Title	企業破産予測のためのテキストベースのアンサンブル学習モデル
Author(s)	NGUYEN, BA HUNG
Citation	
Issue Date	2020-09
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/16992
Rights	
Description	Supervisor:Huynh Nam Van, 先端科学技術研究科, 博士

DOCTORAL DISSERTATION

Textual-based Ensemble Learning Models for Corporate Bankruptcy Prediction

BA-HUNG NGUYEN

Supervisor: Professor Huynh Van Nam

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Doctor of Philosophy

September 2020

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

List of Algorithms

ACKNOWLEDGMENTS

I would like to thank many people who helped me along my path to writing this dissertation.

I would especially like to thank my supervisor, Professor Nam Huynh for taking me under his wing, his exceptional supports, and for running a lab as a big family with a fun and intellectual atmosphere. I would also like to thank my second supervisor Professor Takashi Hashimoto for all of his knowledge he has shared with me, Professor Kiyoaki Shirai for all of value comments and advises for my minor research, and Professor Galina Andreeva for her unconditional supports at Edinburgh during my visiting research.

I'd like to thank several people at John von Neumann Institute who were instrumental in getting me interested in machine learning in finance and starting me along this path, including Professor Hi-Duc Pham, Professor Tu-Bao Ho, Dinh Hai, Viet Hung, Quoc Huy, and Tung Nguyen.

I'd like to thank JAIST for awarding me the Doctoral Fellowship and Off-campus grant. The grant has given me the opportunity to spend time on one of the leading credit research center with projects like the survival analysis in credit risk and Professor Galina's Digital Footprint project. I'd especially like to thank Matsumoto for her supports for grant preparation and Professor Dam Hieu Chi for his inspiring discussions to apply the grant.

I'd like to thank the people I worked with at Edinburgh for making my visiting period an enjoyable time and providing a lot of help and mentorship. I'd especially like to thank Scarlet, Victor, Oscar, Antonia, Andrej, Yagmur and Leone for many good lunches and interesting discussions. I'd especially like to thank Professor Raffaella Calabrese for all of her supports.

I'd like to thank JAIST Supercomputer Unit for keeping all of the computing and software infrastructure running smoothly, and Student Welfare Section, Educational Service Section, and especially Career Support Office for unconditional helps.

I'd especially like to thank the support from The US Office of Naval Research Global under Grant No. N62909-19-1-2031: "Multi-source Learning Coupled with Reasoning Capabilities for Large-scale Decision Support Environments" for this dissertation.

Several people were very supportive in my personal life during the past three years. I'd like to thank Xuan Thang for all of his preparation when I first arrived in Nomi. I'd like to thank my friend and karatedo partner Vinh Vo for joining me in dojo center while working hard on my research. Finally, I'd like to thank my parents for raising me to acknowledge difficulties and to become a valuable person, my wife Phuong for all of her support and understanding, and the many sacrifices she has made to let me continue pursuing my research.

ABSTRACT

A credit score is an estimation of the likelihood that a borrower will show some undesirable behaviors in the future and supports decision making in credit risk modelling. Nevertheless, the majority of studies were usually based on a snapshot of financial-related data at a specific time point in the past, excluded the trend in business performance over years, and ignored up-to-date business/social activity information that might suggest an early warning of changing in credit worthiness. In addition, advances in data mining for social media and machine learning in application for text mining can be applied for the identification of key features for credit scoring models in term of timeliness, to improve the trade-off between cost and accuracy. Hence, the research that utilises both time series data and textual data can help not only to address the shortage of data types and sources, but also to introduce a new approach in credit scoring.

My research tackle these crucial issues with (i) examining more recent and time-series based financial data with a trendy approach adapted from epidemiology and (ii) the development of new ensemble learning approaches that combine tradition statistical models and machine learning models in credit risk modelling capable of handling corporate rich-featured data, including both numeric and textual data.

First, this study employs a large longitudinal data for the UK SMEs to examine their time-to-liquidation using survival analysis, a well-known technique from clinical research. Despite of severely lacking financial data, this study shows the significant effects of SME's demographic characteristics and also further stresses on improvement both in causal interpretation and in model discrimination power when utilising the extended hazard models using the time-varying nature of SMEs financial variables. Another crucial finding in the implication of using some traditional statistic models is the bias in decision-making, where we show that excluding the gender feature eventually reduce the acceptance rates of the better credit worthiness class in both traditional statistical and machine learning-based models. Which questions on the current inconsistencies of existing regulations for the automated decision-making tools.

With two recent, imbalanced corporate credit datasets, this study then sheds more light on the comparison of corporate credit risk models with different balancing strategies and performance measurements. This study shows that the AUC is not a sufficient measure for the imbalanced dataset as the classifiers tend to overfitted toward the majority class with extremely low value of precision and recall, and second, sampling methods provide significant improvement toward the correctness of classifiers in problems that minority class play an important role as in credit risk management. As any single model has its drawbacks and advantages in a specific domain, combining several models might result in improvement in classification accuracy. In the light of reducing the risk of overfitting as well as underfitting, my research combine models using three approaches to build meta-algorithm including bagging, boosting, and stacking. This study shows that homogeneous and simple heterogeneous ensemble classifiers show better performance compared with the traditional individual classifiers. These findings based on two recent loan portfolios of Vietnamese and US corporate data provide more insights to the practice of corporate credit risk modelling.

Finally, to the utilisation of textual data in credit risk modelling, this study employs topic model on textual data to (i) explore the aspects that defines creditworthiness, (ii) learn the distributed representation of textual data, and (iii) combine it with traditional industry standard to improve the credit risk prediction. I uncover 30 topics embedded in the financial reports which reflect important business aspects and the evolution of words in many topics are in line with crucial economics events. More importantly, the topical features alone provide comparable performance with industrial standard using z-score. And by concatenating the topical features and z-score features, the classifier demonstrates the state-of-the-art performance in corporate bankruptcy prediction. In addition, I proposed novel models that learn from both numeric and textual data from financial reports to examine the predictability of models built from dictionary-based count vectorisation of financial report and dictionary-based sentiment classifier using a financial dictionary. The approach provides comparable and consistent predictive results, yet with more simple and intuitive features compared with the deep learning model.

Keywords: bankruptcy prediction · ensemble model · textual analysis · topic modelling · sentiment analysis

Chapter 1

Introduction

1.1 Credit Scoring and Credit Risk Modelling

A credit score is an estimation of the likelihood that a borrower will show some undesirable behaviors in the future and supports decision making in credit retail sector. In credit risk assessment, Tsai and Hung (2014), Wu et al. (2014) and a review of Chen et al. (2016) have indicated that most previous studies examined static models from historical financial statements and/or finance related data/surveys which are limited due to the fact that these data are far behind the recent financial crisis. Nevertheless, these studies were usually based on a snapshot of financial-related data at a specific time point in the past, excluded the trend in business performance over years, and ignored up-to-date business/social activity information that might suggest an early warning of changing in credit rating of a loan application. They also pointed out that more recent data especially time series and alternative sources of data are of the urgent consideration in modeling credit risk. In addition, advances in data mining for social media and machine learning in application for text mining can be applied for identification key features for credit scoring models in term of timeliness, to improve the trade-off between cost and accuracy. Hence, the research that utilizes both time series data and social media data can help not only address the shortage of data types and sources, but also introduce a new approach in credit scoring.

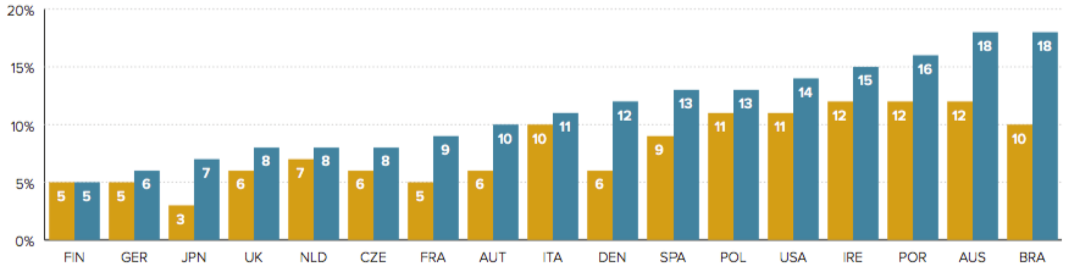
1.2 Modelling Approaches

In credit worthiness modeling, Huang et al. (2004) categorised those in to two main types of models: traditional statistical models and artificial intelligence (AI) models:
- For statistical models, Altman et al. (1977) examined the predictability of seven-variable discriminant analysis and showed the improvement towards his previous model with five variables. Wiginton (1980) showed that logistic regression (LR) surpassed multiple discriminant analysis.

These statistical models, in general, were succinct and were easy to explain. However, the agreement on which variables to use has not been reached yet and the multivariate normality assumption on variables might be violated in financial dataset (Chen et al., 2016).

- As for AI models, the pioneer work of Jo et al. (1997) pointed out that artificial neuron network (ANN) outperformed statistical models. Wu et al. (2014) reaffirmed the performance of support vector machine (SVM) that it is comparable with back-propagation neuron network (BPN) and genetic programming (GP). The study of Tsai and Chen (2010) for four approaches to combine statistical models and machine learning technique indicated that the combined model of LR and ANN gave the best result. Zhao et al. (2015) showed that they could improve multilayer perceptron (MLP) by using their Average Random Choosing method in imbalanced dataset (IDS). In general, AI models are state of the art and can improve the predictive accuracy, however, there application in real world is questionable since it lacks the ease of interpretation, explanation and understanding from the industrial users viewpoint (Sun et al., 2014).

To overcome these drawbacks and exploit the information gain by combining different models, hybrid or ensemble techniques are currently widely used. Huang et al. (2015) showed that the proposed algorithm outperforms popular models including LR, ANN and SVM. Also, in a review of Sun et al. (2014), an early warning system will be crucial for enterprise practice. From the current trend in data-driven decision aids and the richness of online-mediated forms of credit granting and credit-related knowledge, especially those have high potential to affect business activity, they are being produced at a quick growing rate (Fig. 1) through social media including: news, articles, Facebook, Tweeter, LinkedIn, and so forth. And they exist mostly in text form a high dimensional data. Hence apply data mining techniques particularly natural language processing (NLP) on these data in order to improve the discrimination power of both statistical and leaning models will be the future tendency to add more on the accuracy and timeliness of creditworthiness assessment.



Q4. You say you've used these sources of news in the last week, which would you say is your MAIN source of news. *Base:* All in 2015/2016 who used a source of news in the last week: (between around 1500 and 2000 in each country)

Figure 1.1: Growth in social media as main source of news between 2015 and 2016 (Reuters Institute)

1.3 Challenges

There are problems of current practices in credit scoring modeling in both data and methodology, to be specific:

1. Time series dataset and Imbalanced dataset (IDS) treatment:
   More time series-based data is needed to confirm the applicable of model performance over time and it should include not only just quantitative variables but also the qualitative ones. Text data from both financial reporting explanation and social media will also be mined, with sentiment analysis, these results as of financial aspect/sentiment will contribute in overall scorecard of a loan application. On the other hand, the IDS characteristic is a real phenomenon, exists in every dataset relating to loan portfolios of banks since most loan applications naturally are good and those bad ones are rare. Hence, the proportion of major class toward minor class could be over 9:1 and this will distort the models prediction capability (Chen et al., 2016). In this research, sampling techniques and optimization on learning algorithm using cost sensitivity leaning (CSL) will be applied to examine the performance of different classifiers on skewed dataset.

2. Combination of models using hybrid and ensemble techniques: Combining different models results is of current interest since it can benefit from individual classifiers strength. Koh et al. (2015) concluded that in addition with increasing the classification performance, combined models could also reduce the subjectivity and increase objectivity in risk assessment. Xiao et al. (2016) coped with IDS using a dynamic classifier and dynamic ensemble selection of features resulted in better performance compare to ensemble static classifiers. On the conclusions of Lessmann et al. (2015) and Ala'raj and Abbod (2016), ensemble models on which classifiers learnt from different data sources are of the future direction. This research will examine the ensemble form of classifiers based on accounting data with those learnt from social media data.

3. Application of deep learning for financial textual data: Mai et al. (2019) is the first to utilise the deep learning model to learn the textual representation for the task of bankruptcy classification, and they showed potential improvement towards traditional statistical models. However, there are many works to be done including (i) examining the word level predictive power, (ii) identifying the gap between dictionary-based analysis with deep learning based representation learning.

4. In examining the business performance, experts usually rely on the keys financial elements in financial statements. However, there is a crucial problem in this practice - **window dressing** (Nemoto et al., 2018; Gandhi et al., 2019) which is the manipulation of financial statements.

1.4 Research Questions and Contributions

My research tackle these crucial issues with the development of new ensemble learning approaches that combine tradition statistical models and machine learning models in credit risk modelling capable of handling corporate rich-featured data, including both numeric and textual data.

My PhD Dissertation aims to address three main questions:

1. **How traditional statistical and machine learning models perform in bankruptcy classification with complex and skewed dataset?**

   First, this study employs a large longitudinal data for the UK SMEs to examine their time-to-liquidation using survival analysis, a well-known technique from clinical research. Despite severely lacking financial data, this study shows the significant effects of SME's demographic characteristics including location, number of shareholders, trading addresses, directors, contacts, subsidiaries, auditors/accountants, bad debtors, and unsecured creditors. This study also further stresses on improvement both in causal interpretation and in model discrimination power when utilising the extended hazard models using the time-varying nature of SMEs financial variables. Another crucial finding in the implication of using some traditional statistic models is the bias in decision-making, where I show that excluding the gender feature eventually reduce the acceptance rates of the better credit worthiness class in many both traditional statistical and machine learning-based model. Which questions on the current inconsistencies of existing regulations for the automated decision-making tools

   Credit risk modelling especially for corporate segment experiences the high level of imbalanced dataset, this phenomenon can affect model performance as the accuracy is only reflecting the underlying class distribution. With two recent, imbalanced corporate credit datasets, I shed more light on the comparison of corporate credit risk models with different balancing strategies and performance measurements. Specifically, the performance of twelve classifiers belong to linear-based, kernel-based, tree-based, homogeneous ensemble-based, and heterogeneous ensemble-based classes are examined under three sampling strategies with five performance measurements. I show that the AUC is not a sufficient measure for the imbalanced dataset as the classifiers tend to overfitted toward the majority class with extremely low value of precision and recall, and second, sampling methods provide significant improvement toward the correctness of classifiers in problems that minority class play an important role as in credit risk management.

2. **In term of hybrid/ensemble forms of classification method, what forms lead to the improvement in term of classification performance?**

   As any single model has its drawbacks and advantages in a specific domain, combining several models might result in improvement in classification accuracy. In the light of reducing the risk of overfitting as well as underfitting, My research combine models using three approaches to build meta-algorithm including bagging, boosting, and stacking.

This study employs six homogeneous ensemble classifiers including Bagged Logistic Regression, Random Forest (Bagged Tree), Bagged SVM, Bagged MLP, Boosted LR, and Boosted Decision Tree and two heterogeneous classifiers which are Simple Average Ensemble and Hill-climbing Selection Ensemble. The results show that homogeneous and simple heterogeneous ensemble classifiers show better performance compared with the traditional individual classifiers. These findings based on two recent loan portfolios of Vietnamese and US corporate data provide insights for the practice of corporate credit risk modelling.

3. **How should knowledge discovery in databases be applied to corporate textual dataset? What are the most significant factors that affect their creditworthiness?**

   Along with accounting data, corporate business is also very sensitivity with the information in textual data, hence, finding the sentiment with which topics their managers are talking about in financial reports is definitely a crucial task leads to improvement of models predictability. Together, these aspect/sentiment components could be used to examine in which business aspects a specific corporate are doing good or bad. This information then later plays as a comparison/complement role for other machine learning models built on accounting data, which ultimately result in a better models predictability.

   In order to address these questions, I proposed novel models that learn from both numeric and textual data from financial reports to:

   - Examine the predictability of models built from (i) dictionary-based count vectorisation of financial report and (ii) dictionary-based sentiment classifier using a financial dictionary. The textual features are significant, and they improve the predictive power of the classification model.

     My approach provides comparable and consistent predictive results, yet with more simple and intuitive features, compared with the deep learning model (Mai et al., 2019). The largest improvement comes from the SMEs segment with the gain in AUC ranging from 8.4% to 11.5% followed by recall ranging from 3.3% to 9.7%. Besides, by using one-year-head prediction, I provide a practical investigation on improvement of the predictive power using the textual features where the combined features could significantly increase the AUC from 1.1% to 7.6% in the three corporate segments.

   - Employ topic model on textual data to (i) explore the aspects that defines creditworthiness, (ii) learn the distributed representation of textual data, and (iii) combine it with traditional industry standard to improve the credit risk prediction.

     I uncover 30 topics embedded in the financial reports which reflect important business aspects such as energy, partnership, research and development, loan and interest rate, and so forth. In addition, the evolution of words in many topics are in line with crucial economics events such as the financial crisis, the big reduction in coal production in 2015, or the cycles of competition of electronic devices producers. More importantly, the topical features alone provide comparable performance with industrial standard using z-score. And by concatenating the topical features and z-score features, model demonstrates the state-of-the-art performance in corporate bankruptcy prediction.

1.5 Dissertation Organisation

This dissertation begins with the introduction to traditional statistical models in Chapter 2 where current industrial standards are discussed along with adaptations of survival analysis, a well-known technique from epidemiology to credit risk modelling. Chapter 3 discusses potential bias resulting from the inconsistencies of the current law and automated decision-making tools in credit scoring. Chapter 4 investigates in a great detail the comparison of performance of several machine learning models with logistic regression under three main balancing strategies. Chapter 5 and 6 experiment with the textual data and emphasise on how this new sources of data could improve both understanding of the nature of risk, and classifier predictability. Specifically, Chapter 5 explores the topics in financial filings, discusses on how their evolution are in line with crucial economics events. Chapter 6 focuses on the sentiment on word level based on a domain-specific wordlist, in this chapter I propose a financial dictionary-based sentiment classifier to construct new sentiment-based features for bankruptcy prediction task. Chapter 7 concludes the dissertation and discuss my current limitations and further research directions.

Chapter 2

Statistical Models

2.1 Discriminant Analysis

Discriminant analysis is a set of methods to distinguish groups in data and to assign new observations into one of the existing groups. Linear discriminant analysis and multiple discriminant analysis refer to methods when number of groups is two and more than two, respectively.

In credit quality assessment, the objective of multiple discriminant analysis is to distinguish default from non-default firms[1] as accurately as possible by a function of several independent creditworthiness factors (financial ratios, indicators from financial statements). We classify a firm into one of several groups base on their individual characteristics.

In multiple discriminant analysis, a weighted linear combination of factors is formed in order to classify default or non-default customers as much discriminatory power as possible on the basis of the discriminant score D:

$$D = a_0 + a_1 K_1 + ... + a_n K_n, \tag{2.1}$$

where n is the number of financial ratios, K_i is the specific ratio value, and a_i is the ratio's coefficient.

2.2 z-score Model and Its Extensions

With multiple discriminant analysis models, the selection of ratios is the most important and credit experts mostly use their experience and their risk appetite in their choice. The most

[1] In the simplest case, we regard a firm is either default or non-default. In practice, we can allocate firms to several groups based on their discriminant score.

popular and well-known factors are **z-score** of Altman (1968), specifically:

$$D = a_0 + a_1 Z_1 + a_2 Z_2 + a_3 Z_3 + a_4 Z_4 + a_5 Z_5, \tag{2.2}$$

with:
Z_1: Working Capital/Total Assets
Z_2: Retained Earning/Total Assets
Z_3: Earning Before Interest and Taxes/Total Assets
Z_4: Market Value Equity/Book value of total Debts
Z_5: Sales/Total Assets

Discriminant analysis is the first tool to be used in developing credit rating models Altman (1968); Harrell (2001). Nonetheless, the implementation of multiple discriminant analysis has been criticised because of its normal distribution assumptions on those financial ratios. In addition, multiple discriminant analysis is limited in its assumption that the data should have homogeneous variance-covariance matrices (Harrell, 2001). These are strong statistical assumptions that are rarely met in practice let alone the sample size and outlier restrictions. There are many extensions for multiple discriminant analysis and z-score factors, and the most notable is those for small and medium enterprises (SMEs) in Altman and Sabato (2007) where the authors propose the alternative financial ratios to apply for SMEs credit assessment using logistic regression:

Table 2.1: Altman's SMEs factors

Z_1	Cash Flow from Operating Activities / Current Liabilities
Z_2	Short Term Debt / Equity Book Value
Z_3	Cash / Total Assets
Z_4	EBIT / Interest Expenses
Z_5	Account Receivable / Liabilities

2.3 Logistic Regression

Regression models show the relationship of a dependent variable with other independent variables. In practical credit assessment procedures, certain creditworthiness factors (independent variables) will help classification model to decide whether a loan could be classified as default or not (dependent binary variable). Using regression models also enable us to calculate membership probabilities and thereby to determine default probabilities directly from the model function.

We present in this chapter logistic regression. Denotes Φ is the cumulative standard normal distribution function, and $\sum$ represents a linear combination of the financial factors:

$$\sum := b_0 + b_1 K_1 + ... + b_n K_n, \tag{2.3}$$

where:
n: the number of financial factors
K_i: the specific value of creditworthiness criteria i
b_i: factor's coefficient within the rating function (for $i = 1, ..., n$)

In binary classification using logistic model or logit, the default probability p of a given loan is calculated as follow:

$$p_{Logit} = \frac{1}{1 + exp[-(\sum)]}. \tag{2.4}$$

Logistic regression has a number of strong points compare with multiple discriminant analysis. It not only does not require normal distribution in input variables which enable logistic regression to undertake qualitative creditworthiness factors directly but also its result can be interpreted as the probability of group membership (Harrell, 2001). And, logistic regression provides more robust and accurate results than those generated by multiple discriminant analysis if its assumptions are hold and there is a large number of observations in training data. Logit model is easy to implement and links to other crucial elements of Basel II and III[2] such as probability of default (PD), loss given default (LGD) and expected loss (EL).

There are other traditional statistical models such as naïve bayes, decision tree, support vector machine, or even genetic algorithm could be use in credit risk assessment, however, this is out of scope of this thesis, reader could refer to the book of Thomas et al. (2017) for more detail. We especially mention logistic regression because (i) it is a well-known industrial standard, (ii) logit model is embarrassingly simple and explainable, and (iii) it could present state of the art performance in many performance metrics compare with advanced statistical and machine learning models as shown in below chapters. Next, I present new approach to credit risk using the survival analysis - a popular method in epidemiology which is specially designed for censored data.

2.4 Survival Analysis in Credit Risk

As there are increasingly attention being paid on the application of lifetime analysis on credit risk modellings as stated in BASEL accords and IFRS 9 documentations[3], the time until an event (often regarded as default, in liquidation, or dissolved) happens is of the objective of the survival analysis.

Survival analysis is a branch of statistics for analysing the expected duration of **time** until one or more **events** happen. Survival analysis attempts to answer questions such as:

[2] Basel Committee on Banking Supervision - Revisions to the Standardised Approach for credit risk
[3] Macro Econometric IFRS9 and Stress Test models using Survival Analysis, Ribeiro, 2016

- What is the proportion of a population which will survive past a certain time?
- Of those that survive, at what rate will they default/failure?
- Can multiple causes of default/failure be taken into account?
- How do particular circumstances or characteristics increase or decrease the probability of survival?

Why not Linear Regression?

- First, survival times are typically positive numbers; ordinary linear regression may not be the best choice unless these times are first transformed in a way that could remove this restriction.
- Second, and more importantly, ordinary linear regression cannot effectively handle the censoring of observations.

2.4.1 Censoring

Observations are called censored when the information about their survival time is incomplete, for example, a corporate purchased a loan and paid back during study time, hence, their default time is not observable (corporate E in Figure 2.1). The most commonly encountered form is right censoring.

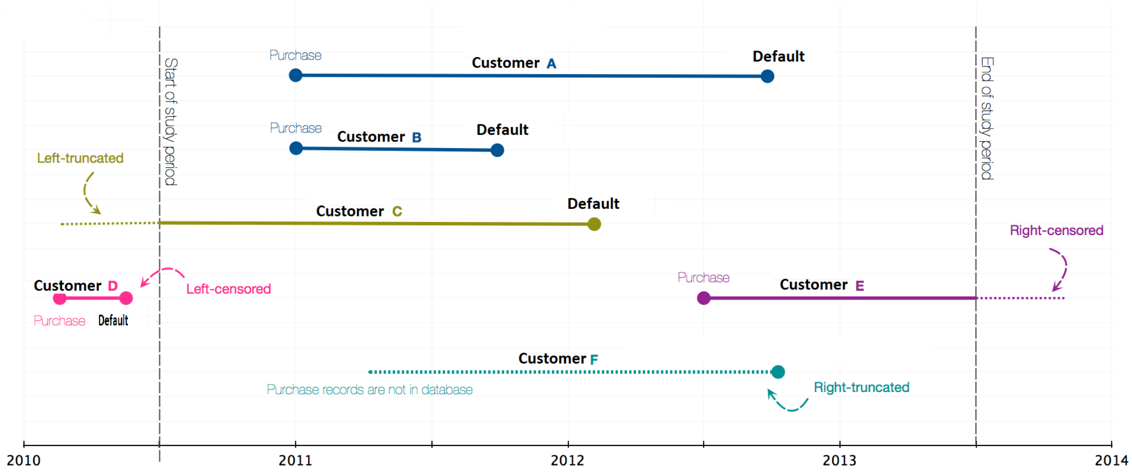


Figure 2.1: Censoring in credit data.

2.4.2 Modelling time to event

The basic of this section is mostly adapted from Cox (1972), Allison (1982), Banasik et al. (1999), and Harrell (2015). Suppose we have a sample of n independent corporate ($i = 1, .., n$) and we start monitor each corporate i from time $t = 0$ up to time t_i. At t_i corporate i either be censored or an event occurs (liquidation in this research). Censoring here means the corporate can not be monitored after time t_i because it is lost to follow-up for reasons like our study stop, being dormant, owners retire, and so forth. Introducing δ_i, the dummy variable for this censoring, i.e $\delta_i = 1$ if corporate i is censored or 0 otherwise. And for each corporate i, we have a vector of independent variables or predictors $\mathbf{X}_i$.

Denote:

- $S(t)$: **survival function** (non-increasing) which is the probability that the time of liquidation T is later than some specified time t: $S(t) = p(T > t)$
- $F(t) = p(T \leq t) = 1 - S(t)$: lifetime distribution function (the cumulative distribution function for T)
- $f(t) = F'(t)$: event density - the rate of liquidations per unit time (the probability density function of T evaluated at t)
- $\lambda(t)$: **hazard function**, the probability that the event will occur in a small interval around t, given that the event has not occurred before time t,

$$\lambda(t) = \frac{\lim_{\Delta \to 0} p(t \leq T < t + \Delta | T \geq t)}{\Delta} = \frac{f(t)}{S(t)}, \tag{2.5}$$

  the hazard function can also be expressed as the negative of the slope of log of $S(t)$[4]:

$$\lambda(t) = -\frac{\partial log(S(t)}{\partial t}. \tag{2.6}$$

- $\Lambda(t)$: cumulative hazard function, the area under $\lambda(t)$. We have

$$\Lambda(t) = \int_0^t \lambda(v)dv = -logS(t). \tag{2.7}$$

  A useful property of the cumulative hazard function is (Harrell, 2015)

$$E[\Lambda(min(T, z))] = 1 - S(z) = F(z). \tag{2.8}$$

  $\sum_1^n \Lambda(min(T_i, z)$ estimates the expected number of events happen before time z among n subjects.

[4] this enable easier determination of the phases of increased risk than looking for sudden drops in $S(t)$

There are several methods could be used to model the time to event T, which could be divided to three main branches:

1. **Nonparametric models** such as Kaplan-Meier estimator or Altschuler-Nelson estimator which useful for descriptive analysis of survival time.
2. **Parametric models** using exponential distribution, Weibull distribution and so forth as a functional form of $S(t)$, to model data in more detail to
   - easily compute selected quantiles of the survival distribution, and
   - estimate (usually by extrapolation) the expected failure time.
3. **Semiparametric models** such as Cox proportional hazard model (CPH, Cox (1972)) which makes a parametric assumption for the effect of the predictors on the hazard function (the regressors are linearly related to log hazard), but no assumption to the nature of the hazard function $\lambda(t)$. As the form of the true hazard function is unknown or complex, the Cox model has definite advantages, especially when we are more interested in *the effects of the predictors* than in the shape of $\lambda(t)$.

Cox Proportional Hazard (PH)

Cox represents the hazard function as a function of both time and covariates using a proportional hazards model (Cox, 1972),

$$\lambda(t|\mathbf{X}) = \lambda_0(t) exp(\beta \mathbf{X}). \tag{2.9}$$

where $\lambda_0(t)$, the baseline hazard, is a function of time which could imply the distribution of T:

- $\lambda_0(t) = \lambda$: exponential distribution.
- $\lambda_0(t) = \lambda_0 + \lambda_1 log\ t$: Weibull distribution.
- $\lambda_0(t) = \lambda_0 + \lambda_1 t$: Gompertz distribution.

Let the set of firms at risk of liquidation before failure time t_i be the risk set at time t_i, denoted as R_i - the set of firms had not been bankrupt or censored by time t_i, we have $R_i = \{j, Y_j \geq t_i\}$ where Y_i is the failure/censoring time of firm j.

Cox used partial likelihood to estimated β:

$$\mathcal{L}(\beta) = \prod_{Y_i \text{uncensored}} \frac{exp(X_i \beta)}{\sum_{Y_j \geq Y_i} exp(X_j \beta)}. \tag{2.10}$$

[4] Cox argued that when the proportional hazard model holds, information about $\lambda(t)$ is not useful in estimating the parameters of primary interest, β.

Two popular methods to approximate the true likelihood in case of ties (multiple event at the same time t_i) are Breslow and Efron (Harrell, 2015).

Survival Analysis with time-varying covariates

The hazard function in Eq. 2.9 could be extended to account for time-varying covariates whose value change with time:

$$\lambda(t|\mathbf{X(t)}) = \lambda_0(t)exp(\beta\mathbf{X(t)}). \tag{2.11}$$

This is particularly true for internal predictor including financial elements as in business data. The value of Total Assets, for example, change every financial year. The same likelihood function 2.10 is used to estimate the extended Cox PH to account for time-varying estimate. The difference is the values of X now changes at each risk set.

2.4.3 Performance measurements

To measure the predictability of an survival model, as in regression modeling, we could use R^2. Other measurements such as Kendalls τ and Somers' D_{xy} rank correlation or C-index. C-index is a generalization of the area under the ROC curve. C-index could be applied for a continuous response variable which can be censored such as time-to-liquidation as in this study. The C-index is the percentage of all pairs of SMEs whose survival time can be ordered in a way that the SME with the higher predicted survival is the one who survived longer. Formally, with s_i and T_i are the survival predicted and time to liquidation of SME$_i$, we have:

$$\text{C-index} = P(s_j > s_k|T_j > T_k). \tag{2.12}$$

$$D_{xy} = 2\text{C-index} - 1. \tag{2.13}$$

2.4.4 Experiment with UK SMEs

SMEs[5] in UK do not necessarily have to report their detail financial statements[6], their filings to the Company House could be a briefed statements in which the very general financial elements such as profit and loss accounts are reported[7]. This explains for lacking of financial data for SMEs to examine their lifetime, hence, the alternative data are usually employed as complement sources. To provide updated results on the determinants of SMEs time-to-liquidations, and to examine to what extend, the time-varying covariates could help predicting SMEs lifetime, this

[5] Reader could refer to Andreeva et al. (2016) for the definition of SMEs in the UK and the liquidation state of the companies.

[6] https://www.gov.uk/government/publications/life-of-a-company-annual-requirements/life-of-a-company-part-1-accounts

[7] following Sections 475 and 477 of the companies act 2006.

work examines the survival of more than 67,000 SMEs in the UK with follow-up time from 2004 to 2016. I then compare the baseline hazard model and its extended model with time-varying predictors on predicting time-to-liquidation. The features of input sample including of company fixed demographics and time-varying financial elements. This work especially stresses on the effect of number of directors on the SMEs survival.

The "Company status" indicator is presented on Table 2.2, where the majority of the SMEs are active:

Table 2.2: Company process

Process	Status	Count
0	Active (dormant), petition to wind-up	1
1	Active, meeting of creditors	6
2	Active, app. of liquidator	7
3	Inactive (no precision)	9
4	Active, petition to wind-up	12
5	Active, with vol. arrangement	27
6	Active, in administration	67
7	Active (dormant), in default	101
8	Active (receivership)	271
9	Active, in default	288
10	In liquidation	2484
11	**Active (dormant)**	10174
12	Dissolved	25481
13	**Active**	62207

Excluding the inactive company with no precision reasons, we regard a company as default if it is in the liquidation processes 0-10 according to the UK government guideline[8]. The dissolved category includes those that do not necessarily experience default or liquidation, they might stop operating because the owner retires, dies, or other reasons. We could consider this category as another level instead of only two level Liquidation/Active in further work. With the requirements of SMEs above, the number of SMEs from FAME[9] that does not satisfy and is excluded is 5,617, among them, 299 SMEs are liquidated. Moreover, by setting an observation period of 12 years follow-up, we also exclude SMEs without exact date of incorporation, and we end up with the following number of default and active SMEs:

Table 2.3: Company status

Code	Status	Count	Percentage
1	In Liquidation	1,598	2.38%
0	Active	65,661	97.62%

The final longitudinal data consists of 800,076 SME-period observations with six fixed and six time-varying covariates. Table 2.4 shows the binning of fixed, categorical covariates, and type of the time-varying covariates as follows:

[8] https://www.gov.uk/liquidate-your-company

[9] FAME, Bureau van Dijk, licensed per subscription of Business School, The University of Edinburgh

Table 2.4: SMEs survival predictors

	Fixed covariate	Categories and Binning
1	#Directors	[$\leq$ 3, > 3] (Group 1, Group 2)
2	Region	12 area postcodes in the UK
3	#Shareholders	[0, 1, 2, >2]
4	#Contacts	[0, 1, 2, > 2]
5	#Trading addresses	[0, 1, > 1]
6	#Subsidiaries	[0, 1]
7	#Auditor/Accountants	[0, 1]
9	#Bad debtors	[0, 1]
9	#Unsecured creditors	[0, 1]
10	SIC*	[C, F, O]**
	Time-varying covariate***	**Type**
1	Total Assets (TA)	Numeric
2	Current Assets (CA)	Numeric
3	Net Tangible Assets (NT)	Numeric
4	Current Liabilities (CL)	Numeric
5	Shareholder Funds (SHF)	Numeric
6	Liquidity Ratio (LR)	Numeric

*SIC: Standard Industry Code
**C: Construction; F: Food&Postal Activities; O: Others
***TA and NT are subsequently removed from modeling process because of collinearity.

2.4.5 Experimental results

Denote Group 1 and 2 as SMEs with number of directors not larger than 3 and larger than 3, respectively. If we use Weibull distribution to model the survival time, with maximum likelihood method, we have

$$p(T \geq t) = exp[-exp(\frac{log(t) - X\beta}{0.5596})] \text{ where } X\hat{\beta} = 4.652 - 0.1792[\text{Group 2}], \tag{2.14}$$

the effect of going from Group 1 to Group 2 is to decrease log failure time by 0.18 for using this parametric estimation, giving a Group 2:1 liquidation time ratio of 0.84.

The conventional Cox PH model is first fitted to this data with baseline value for all continuous predictors (measured after the first financial year). I used 3-knot restricted cubic spline (rcs) transformation to account for the non-linear effects of these predictors on hazard ratio[10], specifically, with X as a continuous covariate defined in Table 2.4, its rcs transformation is as follows:

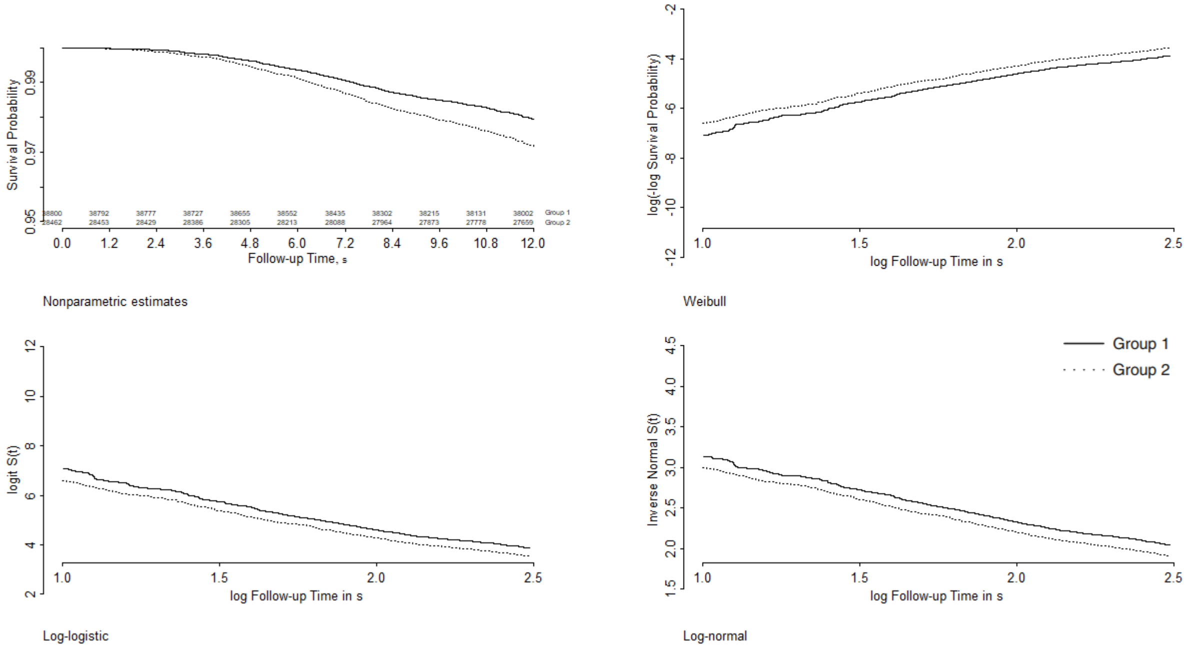
$$rcs(X) = \beta_0 + \beta_1 X + \beta_2 X_2, \tag{2.15}$$

where:

$$X_2 = (X - t_1)^3_+ - (X - t_2)^3_+(t_3 - t_1)/(t_3 - t_2) + (X - t_3)^3_+(t_2 - t_1)/(t_3 - t_2),$$

and $t_1 = 0.1, t_2 = 0.5, t_3 = 0.9$ quantiles of X.

[10] The plain conventional Cox PH model is also fitted without non-linear modification for continuous predictors, however, likelihood ratio test and the test for significant of different in deviance strongly prefer the model with non-linear effects of continuous predictors including SHF, LR, CL, and CA.



Nonparametric estimates

Weibull

Log-logistic

Log-normal

Figure 2.2: Altschuler-Nelson-Fleming-Harrington nonparametric survival estimate along with various parametric estimates. Group 1 and 2 are SMEs with number of directors not larger than 3 and larger than 3, respectively.

The base levels of region and SEC are West.Midlands and SIC of Foods&Postal activities. Those for Number.of.shareholders, Number.of.contacts, Number.of.trading.addresses are 0. Table 2.5 present the summary of fitted Cox PH model, with 95% confident interval for hazard ratio. Some regions have significant higher hazard compare with West.Midlands including North.East, Northern.Ireland, and Scotland. Having more trading addresses or contacts showing the better hazard compare with no trading address or contact. SMEs with subsidiaries or bad.debtor(s) or operate in industries other than food and postal activities have better chance to survive. It is reasonable that SMEs with unsecured creditors will have much lower survival rate. Regarding the continuous predictors, excluding CL, the other financial predictors including SHF, LR, and CA all show significant non-linear effects on hazard ratio.

One of the most important assumption for Cox PH is the proportional hazard, which assumes that the effect on hazard ratio does not change overtime. Figure 2.3 below presents the Schoenfeld residual plot for potential violation of predictors, which shows that the region (missing value category), number of shareholders and number of contacts effects might not satisfy the PH assumption. The bootstrap bias-corrected estimates (Harrell, 2015) of D_{xy} and R^2 are presented as following Table:

A D_{xy} of 0.8853 translates to C-index of 0.94 which means that this model could correctly ranks survival of 9 out of 10 cases. Despite showing good discrimination power in C-index, this model is limited as it has some predictors violates the PH assumption and it does not make use of time-varying predictors. Next, I present the extended Cox PH model which utilises the

Table 2.5: Summary - Conventional Cox PH: “**mtt**” and “**mto**” represent more than two and more than one. “’” denotes the corresponding X_2 of the predictor as defined in Eq. 2.15

	β	Hazard Ratio	2.5%	97.5%	SE	z	p-value
Directors=Group 2	0.0264	1.0267	0.9265	1.1377	0.0524	0.5030	0.6149
region_East.England	-0.0189	0.9813	0.5257	1.8317	0.3184	-0.0594	0.9527
region_East.Midlands	0.0491	1.0503	0.7823	1.4100	0.1503	0.3264	0.7441
region_East.of.England	-0.1284	0.8795	0.6435	1.2021	0.1594	-0.8054	0.4206
region_Greater.London	0.1215	1.1292	0.9005	1.4160	0.1155	1.0524	0.2926
region_NA	-0.1336	0.8749	0.6888	1.1114	0.1220	-1.0948	0.2736
region_North.East	0.3250	1.3840	1.0352	1.8505	0.1482	2.1931	0.0283
region_North.West	0.2420	1.2738	0.9989	1.6244	0.1240	1.9510	0.0511
region_Northern.Ireland	0.5181	1.6788	1.2496	2.2553	0.1506	3.4393	0.0006
region_Scotland	0.3964	1.4865	1.0841	2.0382	0.1611	2.4613	0.0138
region_South.East	0.0138	1.0138	0.7832	1.3124	0.1317	0.1044	0.9168
region_South.West	-0.0961	0.9084	0.6695	1.2325	0.1557	-0.6173	0.5370
region_Wales	0.2211	1.2474	0.8713	1.7860	0.1831	1.2075	0.2273
Number.of.shareholders_mtt	-3.9018	0.0202	0.0167	0.0244	0.0960	-40.6317	0.0000
Number.of.shareholders_two	-4.0659	0.0171	0.0149	0.0198	0.0728	-55.8631	0.0000
Number.of.shareholders_one	-4.3716	0.0126	0.0109	0.0146	0.0732	-59.7072	0.0000
Number.of.contacts_mtt	-2.5282	0.0798	0.0566	0.1125	0.1754	-14.4158	0.0000
Number.of.contacts_two	-2.0042	0.1348	0.1112	0.1634	0.0982	-20.4054	0.0000
Number.of.contacts_one	-0.9372	0.3917	0.3368	0.4556	0.0771	-12.1574	0.0000
Number.of.trading.addresses_mto	-0.0408	0.9600	0.8081	1.1405	0.0879	-0.4643	0.6424
Number.of.trading.addresses_one	-0.1171	0.8895	0.7881	1.0040	0.0617	-1.8957	0.0580
Subsidiaries	-1.3302	0.2644	0.1788	0.3911	0.1997	-6.6596	0.0000
Auditors.Accountants	0.1045	1.1102	1.0007	1.2317	0.0530	1.9732	0.0485
Bad.debtors	-0.4150	0.6603	0.5200	0.8385	0.1219	-3.4045	0.0007
Unsecured.creditors	1.1258	3.0827	2.5790	3.6847	0.0910	12.3690	0.0000
SIC_C	0.0725	1.0752	0.8237	1.4034	0.1359	0.5332	0.5939
SIC_O	-0.4570	0.6332	0.4990	0.8035	0.1215	-3.7608	0.0002
SHF	-0.0001	0.9999	0.9998	1.0000	0.0001	-1.5827	0.1135
SHF’	0.0007	1.0007	1.0001	1.0013	0.0003	2.1566	0.0310
LR	-0.2716	0.7622	0.6999	0.8299	0.0435	-6.2495	0.0000
LR’	0.3227	1.3809	1.2310	1.5489	0.0586	5.5063	0.0000
CL	0.0017	1.0017	0.9999	1.0035	0.0009	1.8090	0.0704
CL’	-0.0031	0.9969	0.9937	1.0002	0.0017	-1.8446	0.0651
CA	0.0039	1.0039	1.0022	1.0056	0.0009	4.5269	0.0000
CA’	-0.0106	0.9895	0.9849	0.9941	0.0024	-4.4546	0.0000

Table 2.6: Bootstrap validation - Conventional Cox PH

Index	Original	Training	Test	Optimism	Corrected	n
	Sample	Sample	Sample		Index	
D_{xy}	0.8877	0.8867	0.8844	0.0023	0.8853	100
R^2	0.2579	0.2549	0.2550	-0.0001	0.2580	100

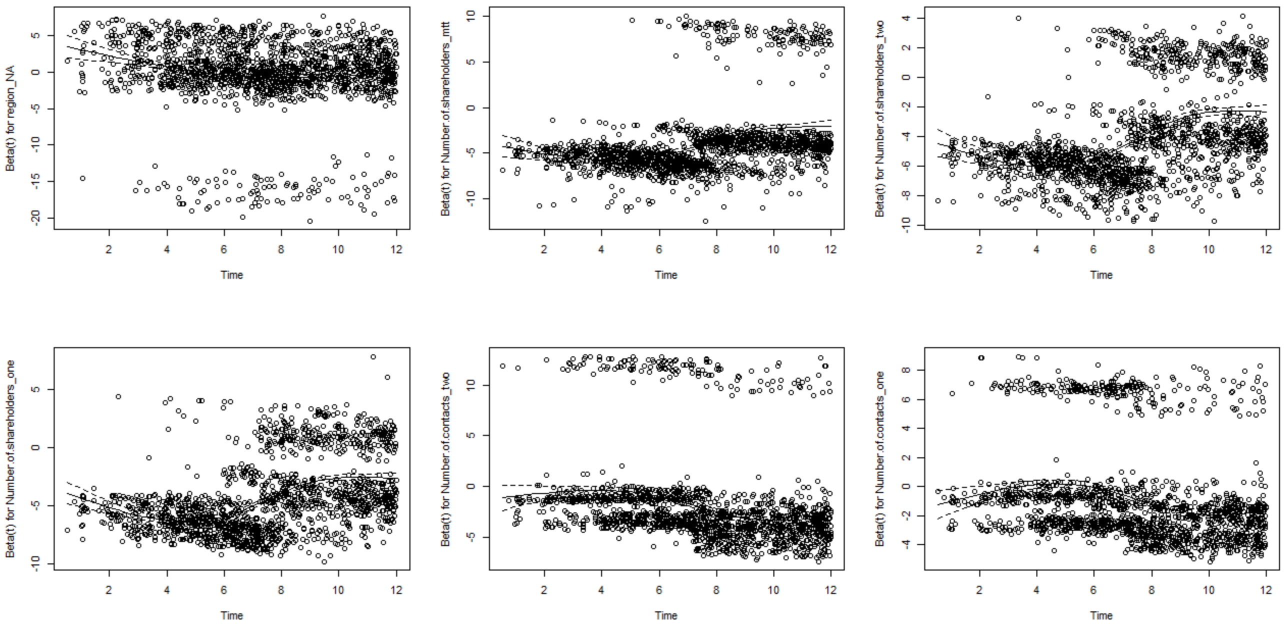


Figure 2.3: Schoenfeld residual plot

time-varying characteristics of financial elements available from SMEs filings to the Company House.

Table 2.7 presents in the same manner the estimated coefficients for several predictor as in Table 2.5 but with financial predictors including SHF, LR, CL, and CA being time-varying. The effect of having more than three directors is now significant and reduce the hazard rate compare with SMEs having less than three directors. SMEs that do not have region, or operate in North East or North West or Scotland have higher rate of liquidation than those from West Midlands. The effects of number of trading addresses, number of contacts, number of shareholders, or having subsidiaries, auditor/accountant, bad debtors, unsecured creditor are similar with that from the conventional Cox PH model. Noticeably, while the effect of CA remains (increase the hazard rate) and of CL turns significant, the effect of LR now increase the hazard of SMEs.

The bootstrap-corrected index for D_{xy} and R^2 are presented in Table 2.8 demonstrates that using updated-value for financial predictors could help improve the discrimination index to 5.5 and 2.9 basic points for D_{xy} and R^2, respectively.

Figures 2.4 and 2.5 below present the summary of effects of 14 predictors with categories for each categorical predictors of the conventional Cox PH and extended Cox PH, respectively. The summary of extended Cox PH not only show better causal interpretation with more significant effects of categorical predictors, and number of directors as this study hypothesise, but also the continuous-financial predictors, when correctly modeled, could improve the model fitness and discrimination power as shown in Table 2.8.

Table 2.7: Summary - Extended Cox PH: “**mtt**” and “**mto**” represent more than two and more than one. “’” denotes the corresponding X_2 of the predictor as defined in Eq. 2.15. Financial predictors are time-varying.

	β	Hazard Ratio	2.5%	97.5%	SE	z	p-value
Directors=Group 2	-0.1105	0.8954	0.8081	0.9922	0.0524	-2.1103	0.0348
region_East.England	-0.1124	0.8937	0.4786	1.6687	0.3186	-0.3528	0.7242
region_East.Midlands	0.0291	1.0295	0.7676	1.3806	0.1497	0.1940	0.8461
region_East.of.England	-0.0853	0.9182	0.6720	1.2546	0.1593	-0.5357	0.5921
region_Greater.London	-0.0078	0.9922	0.7914	1.2440	0.1154	-0.0675	0.9462
region_NA	0.6569	1.9289	1.5189	2.4495	0.1219	5.3882	0.0000
region_North.East	0.3024	1.3530	1.0123	1.8084	0.1480	2.0427	0.0411
region_North.West	0.2544	1.2897	1.0122	1.6433	0.1236	2.0584	0.0396
region_Northern.Ireland	0.1482	1.1598	0.8679	1.5498	0.1479	1.0021	0.3163
region_Scotland	0.5772	1.7811	1.2993	2.4415	0.1609	3.5869	0.0003
region_South.East	0.1531	1.1655	0.8999	1.5095	0.1320	1.1604	0.2459
region_South.West	0.1806	1.1979	0.8833	1.6246	0.1555	1.1615	0.2455
region_Wales	0.3244	1.3832	0.9660	1.9806	0.1832	1.7713	0.0765
Number.of.shareholders_mtt	-2.7473	0.0641	0.0528	0.0778	0.0990	-27.7388	0.0000
Number.of.shareholders_two	-2.7117	0.0664	0.0571	0.0772	0.0769	-35.2536	0.0000
Number.of.shareholders_one	-3.0847	0.0457	0.0394	0.0530	0.0756	-40.8120	0.0000
Number.of.contacts_mtt	-2.3027	0.1000	0.0706	0.1415	0.1772	-12.9931	0.0000
Number.of.contacts_two	-1.5464	0.2130	0.1753	0.2589	0.0996	-15.5302	0.0000
Number.of.contacts_one	-0.7079	0.4927	0.4237	0.5729	0.0770	-9.1955	0.0000
Number.of.trading.addresses_mto	-0.0911	0.9130	0.7678	1.0855	0.0883	-1.0308	0.3026
Number.of.trading.addresses_one	-0.2080	0.8122	0.7189	0.9175	0.0622	-3.3429	0.0008
Subsidiaries=1	-1.5388	0.2146	0.1450	0.3177	0.2002	-7.6880	0.0000
Auditors.Accountants=1	0.1166	1.1237	1.0121	1.2475	0.0533	2.1859	0.0288
Bad.debtors=1	-0.4450	0.6408	0.5047	0.8136	0.1218	-3.6526	0.0003
Unsecured.creditors=1	1.3925	4.0249	3.3512	4.8342	0.0935	14.8980	0.0000
SIC_C	-0.1160	0.8904	0.6829	1.1610	0.1354	-0.8571	0.3914
SIC_O	-0.5294	0.5890	0.4644	0.7469	0.1212	-4.3672	0.0000
SHF	0.0001	1.0001	1.0000	1.0001	0.0000	1.3413	0.1798
SHF’	-0.0011	0.9989	0.9983	0.9996	0.0003	-3.3607	0.0008
LR	1.5018	4.4896	4.0407	4.9883	0.0538	27.9396	0.0000
LR’	-2.6167	0.0730	0.0602	0.0886	0.0985	-26.5530	0.0000
CL	0.0295	1.0299	1.0272	1.0327	0.0014	21.4168	0.0000
CL’	-0.0551	0.9464	0.9417	0.9512	0.0026	-21.3802	0.0000
CA	0.0177	1.0178	1.0164	1.0193	0.0007	24.5641	0.0000
CA’	-0.0551	0.9464	0.9422	0.9505	0.0022	-24.5746	0.0000

Table 2.8: Bootstrap validation - Extended Cox PH

Index	Original	Training	Test	Optimism	Corrected	n
	Sample	Sample	Sample		Index	
D_{xy}	0.9460	0.9492	0.9455	0.0037	0.9423	100
R^2	0.3132	0.3195	0.3121	0.0074	0.3058	100

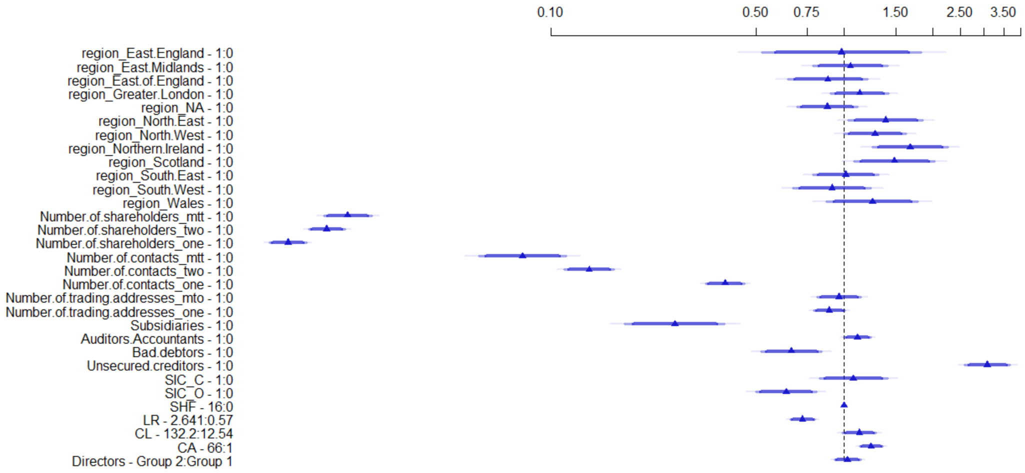


Figure 2.4: Hazard ratios and confidence bars for effects of predictors - Conventional Cox PH model

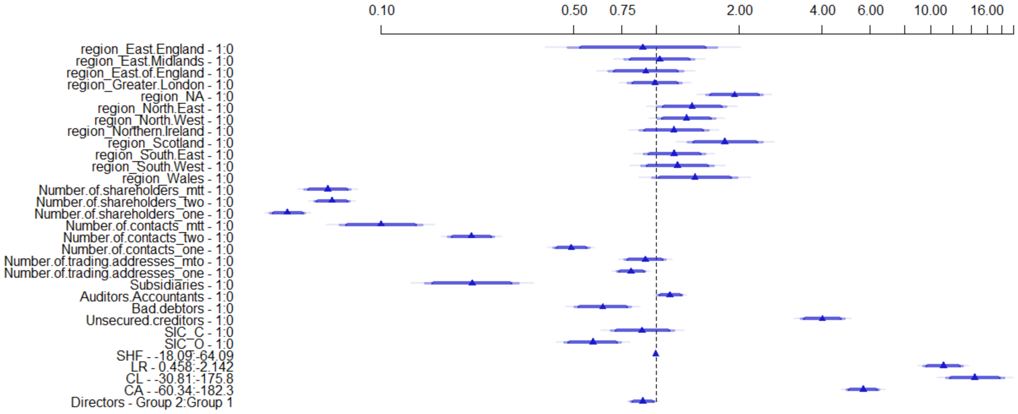


Figure 2.5: Hazard ratios and confidence bars for effects of predictors - Extended Cox PH model

2.4.6 Conclusion

This study first shows the significant effects of SME's demographic characteristics including location, number of shareholders, trading addresses, directors, contacts, subsidiaries, auditors/accountants, bad debtors, and unsecured creditors. This study also further stresses on improvement both in causal interpretation and in model discrimination power when utilising the extended hazard models using the time-varying nature of SMEs financial variables.

Our study is without limitations, first, it focuses on the UK SMEs, which might behave significantly different with other peers since the Brexit referendum in 2016. Hence, more up-to-date data are needed to present this difference. In addition, missing values for SMEs data are prevalence, a comparative analysis on the effectiveness of imputation methods on the performance of survival models is also expected.

Chapter 3

Bias in Automated Decision-making for Credit Scoring

3.1 Problem Statement

Discrimination (defined as unequal outcome for protected group) happens in several classification problems including crime recidivism, social welfare, credit assessment, and so forth. In the previous work, Andreeva and Matuszyk (2019) has demonstrated how existing regulations can disadvantage women. In many countries there are special laws that promote the equality by prohibiting the use of certain characteristics in various decisions, e.g. when selecting job applicants or when deciding who should be given or not given the credit. The list of prohibited characteristics vary from country to country, but most often it includes gender or sex. The law requires that sexes are treated equally, i.e. no distinction is made between men and women. At the same time there is an expectation that the outcome should also be equal between sexes. However, Andreeva and Matuszyk (2019) show that in algorithmic decision-making (i.e. when decisions are based on models trained on real-life data), removal of gender does not lead to equality of outcome (rejection rates) for men and women. This happens because of other characteristics that remain in the model and that are correlated with gender. Furthermore, for women the chances of being rejected for credit would be lower if gender is included into the model, because their historic probability of default is lower.

One of the main approach proposed by machine-learning community towards eliminating such discrimination is to balance the dataset. This research investigates two approaches at data-level preparation: balancing towards the target outcome, and balancing towards the group of interest, and how effective they are in eliminating the discrimination. As sophisticated and powerful machine learning classifiers (e.g. multilayer perceptron (MLP), extreme gradient boosting (XGB, Sun et al. (2014)) are complex and hard to explain, this study also performs the experiments with standard classifiers such as logistic regression. This is of particular interest especially in the area that requires transparency in decision-making process[1]. Specifically, this research examines the performance of several machine learning classifiers under several data balancing methods with a two-fold objective. First, it aims to compare the predictive performance of machine learning classifiers with logistic regression and, second, it examines the discrimination outcomes for men and women by simulating and comparing the rejection rates.

First, this analysis shows how gender bias and discrimination can arise from real-life data under many automated decision-making tools. Second, we advocate that the protected characteristics should be allowed to use in order to correct for the bias in the data, yet with separately estimated models. Finally, we show that balancing towards the protected class or the target outcome in the training data does not necessarily produce equal rejection rates for men and women.

3.2 Data and Methods

3.2.1 Dataset overview

The dataset is granted from Andreeva and Matuszyk (2019), which comprises a portfolio of car loans from a European bank. The summary of train and test sets is given in Table 1. Bad refers to customers who missed two consecutive monthly payments. The features (and their types) of this data include Number of children (categorical), Car price (numeric), Down payment (numeric), Car age (numeric), Loan duration (numeric), Time in employment (numeric), Net income (numeric), Marital status (categorical), Car engine (categorical), Phone (categorical), and Occupation (categorical).

On the distribution of classes, as shown in Fig 3.1 and Table 3.1, women constitute approximately 26% of the samples. While men have more applications, they also have higher bad rate than women (1.34% compared with 0.34%). In addition, the data is highly imbalanced where almost all observations belong to the majority class - good loans (98.32%), and the minority class - bad loans is only 1.68%.

Table 3.1: Target outcome and gender

	Train Set			Test Set		
	Good	Bad	Total	Good	Bad	Total
Count						
Female	16828	216	17044	4104	59	4163
Male	45613	851	46464	11507	208	11715
Total	62441	1067	63508	15611	267	15878
Percentage						
Female	26.50	0.34	26.84	25.85	0.37	26.22
Male	71.82	1.34	73.16	72.47	1.31	73.78
Total	98.32	1.68	100.00	98.32	1.68	100.00

[1] The Inquiry on "Algorithms in Decision-Making", The Alan Turing Institute submission to the House of Commons' Science and Technology Committee, 28 Feb. 2017

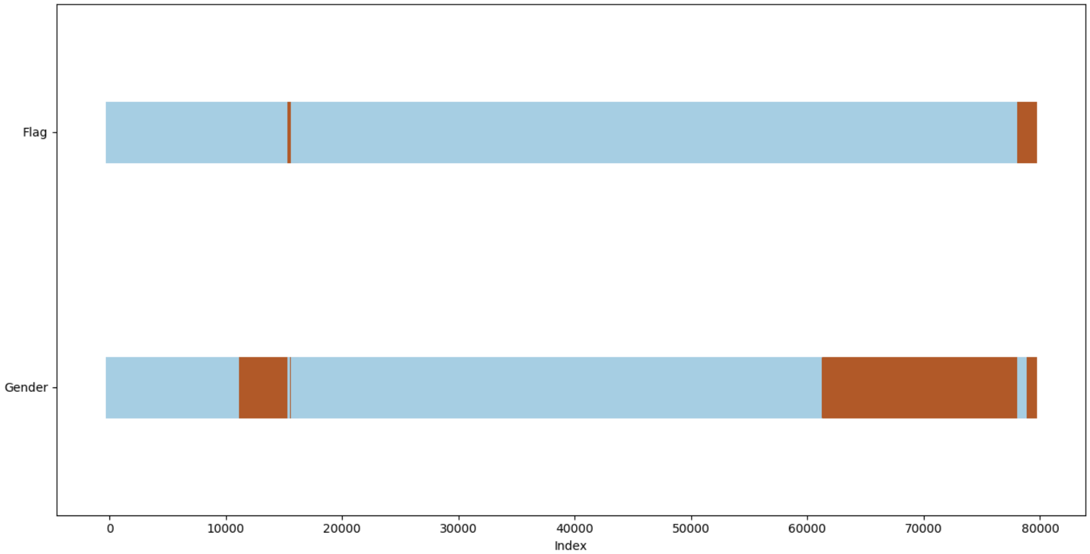


Figure 3.1: Default flag and gender

3.2.2 Imbalanced treatment

To handle imbalanced dataset, random oversampling the minority class and undersampling the majority class are two common sampling methods. Despite literature shows that there are sophisticated methods of creating additional data for the minority class, including SMOTE (Chawla et al., 2002) or ROSE (Menardi and Torelli, 2014), however, these methods can not effectively deal with categorical features and might produce unrealistic samples (e.g. customer with -1 child).

3.2.3 Classifier

Five classifiers belong to Bayesian-based (Naïve Bayes - NB), linear-based (logistic regression - LR), tree-based (decision tree - DTC), network-based (multilayer perceptron - MLP), and homogeneous ensemble-based classes (extreme gradient boosting with tree-based classifier - XGB) are examined under three sampling strategies with area under the ROC curve (AUC Goadrich et al. (2006)). The classifiers and their tuning parameters are as follows:

Table 3.2: Classifier hyperparameters

Model	**Parameter(s)**	**Tuning Range**
NB	Priors Probability	[[0.9, 0.1], [0.8, 0.2], [0.7, 0.3]]
LR	Regularization strength: C	$linspace(0.5, 1.5)$
DTC	Max depth	range (3, 5)
	Max feature	range (5, 20)
MLP	Hidden Layers Sizes	[(10,10,10), (10,10), (10,)]
	α	[0.0001, 0.05]
XGB	Learning rate	$logscale(-4, 1)$
	Number of estimators	range (20, 50)

To ensure the same imbalanced characteristics for each of training dataset before balancing and modelling, we utilyse stratified k-fold to have the same proportion of both target outcome and gender across all folds. Our stratified folds are as follows:

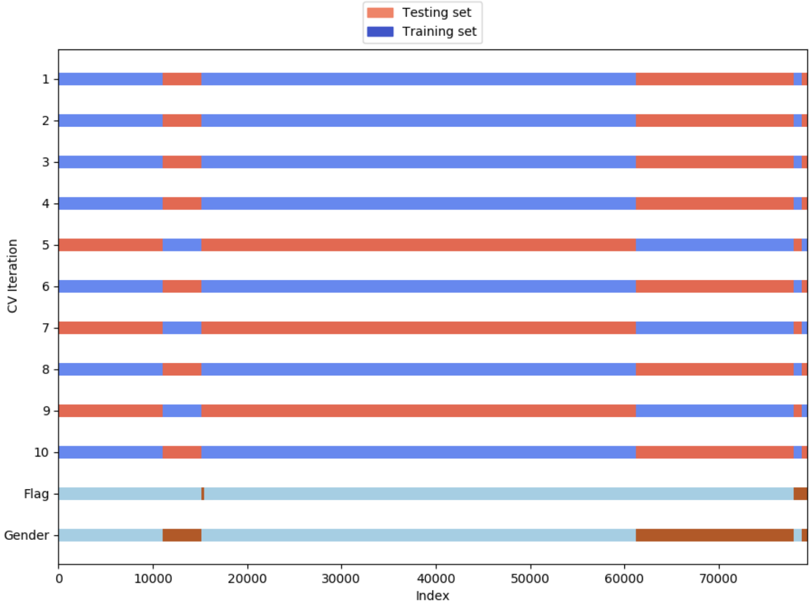


Figure 3.2: Stratified split

Tuning for parameter optimization is repeated within each loop of the cross-validation. And we further set an outer cross-validation (nested cross-validation) to report the final performance metrics to compensate for potential random train-test sample separation bias.

As for the preprocessing, we employ multiple imputation (Buuren and Groothuis-Oudshoorn, 2011) for missing numerical values and include additional missing category for categorical values. Then, for feature transformation, robust scaler and dummy coding are used for numerical and categorical features, respectively. Finally, to compare several setting of models, We fit four models for each classifier in this study as follows: Model with GENDER feature (training sample comprising both men and women, M1); Model without GENDER feature (M2), Model for male segment only (training sample comprising men only, M3), and Model for female segment only (training sample comprising women only, M4). We build segmented Models 3 and 4 as opposed to including interactions, because this approach is preferred in practice to account for different segments in loan portfolios (Banasik et al., 1999; Bijak and Thomas, 2012; Thomas et al., 2017). This approach also makes it possible to accept or reject the same proportion of men and women, thus ensuring the outcome is equal between them.

3.3 Results

In order to compare how access to credit changes for men and women when Gender is used (or not used), we calculate proportions of men and women rejected by Models 1 and 2 (with and without Gender). We experiment with different cutoff levels that would lead to a range of rejection/acceptance rates from 0.1 to 0.9. Since in families it is usually men who apply for credit, it is difficult to compare the rejection rates for married men and women. Therefore, we in this section we restrict out comparison to unmarried customers, which consist of single, divorced and widowed. The results are reported in the following order. First, we reproduce the regression results from Andreeva and Matuszyk (2019) in Table III, in order to show the features used and their effects. They are followed by the performance of classifiers on different segments of data and balancing methods in Table IV. Finally, the rejection rates are illustrated in Figure 3.3.

3.3.1 Regression

We first fit a Logistic Regression to examine the effects of variables. All variables are categorized and dummy coded to several categories as shown in Table 3.3. All variables are categorized and dummy coded to several categories as shown in Table 3.3. The majority of variables are highly significant with small standard error (SE), including Gender that indicates that being a woman has a negative effect on the Probability of default/'bad'. These results are inline with the general literature on credit scoring (Crook and Banasik, 2004; Thomas et al., 2017).

3.3.2 Predictive power on segmentation

We proceed to examine the effect of Gender removal on classifiers performance in two balancing strategies: target and gender balancing. Five classifiers performance on test set under 10-fold nested cross validations are reported in Table 3.4 below. In terms of predictive power, multilayer perceptron presents the strongest performance followed by Logistic Regression in target balancing strategy, whereas in gender balancing, LR shows the best performance across train/test sets and balancing strategies, and is followed closely by MLP. This provides further evidence on the comparable performance of Logistic Regression, an industry standard for credit scoring, with more advanced classifiers. In terms of classifier performance on balanced data, apart from Naive Bayes , all the sampling methods for target balancing strategy modestly improve the full model performance on test set. However, MLP shows superior results with undersampling strategy.

Interestingly, Gender balancing strategy while does not enhance individual classifier under three sampling methods, it does improve the performance for classifiers built on full and male segment data compare with Target balancing strategy, this does not apply for female segment data however. In addition, we observe no significant different in AUC for M1 and M2 in

Table 3.3: Logistic Regression, reproduced from Andreeva and Matuszyk (2019)

	β	SE	z	p-value	[0.025	0.975]
Intercept	-7.286	0.151	-48.127	0.000	-7.583	-6.990
Children 1 kid	0.219	0.088	2.480	0.013	0.046	0.391
Children 2 kids	0.126	0.118	1.067	0.286	-0.105	0.357
Children 3+ kids	0.280	0.212	1.322	0.186	-0.135	0.696
Children missing	-0.663	0.113	-5.887	0.000	-0.883	-0.442
CP Cheap	-0.993	0.116	-8.563	0.000	-1.220	-0.765
CP Expensive	1.099	0.100	11.010	0.000	0.904	1.295
CP Mid Price 2	0.438	0.097	4.518	0.000	0.248	0.628
DP 25-35%	0.818	0.110	7.419	0.000	0.602	1.034
DP 50%+	-1.113	0.167	-6.667	0.000	-1.440	-0.786
DP < 25%	1.285	0.098	13.158	0.000	1.093	1.476
CA 2yrs	1.378	0.130	10.622	0.000	1.123	1.632
CA 3-4yrs	1.860	0.106	17.585	0.000	1.653	2.068
CA 4+	2.505	0.120	20.962	0.000	2.271	2.740
Duration 30-60	0.641	0.103	6.242	0.000	0.440	0.843
Duration 60+	1.528	0.101	15.138	0.000	1.330	1.725
ToE 1-4	1.202	0.091	13.193	0.000	1.023	1.380
ToE 4-7	0.695	0.099	7.040	0.000	0.501	0.888
ToE < 1	0.713	0.111	6.413	0.000	0.495	0.931
NI High inc	-0.490	0.086	-5.736	0.000	-0.658	-0.323
NI Low inc	-0.409	0.088	-4.635	0.000	-0.583	-0.236
NI Mid inc1	-0.134	0.094	-1.421	0.155	-0.319	0.051
Marital D	1.983	0.113	17.537	0.000	1.761	2.205
Marital S	1.446	0.078	18.430	0.000	1.292	1.599
Marital W	1.257	0.186	6.741	0.000	0.891	1.622
Engine 1.4-1.6	-0.056	0.134	-0.415	0.678	-0.319	0.208
Engine 1.6+	0.111	0.109	1.019	0.308	-0.102	0.325
Engine missing	0.616	0.090	6.875	0.000	0.440	0.791
Phone C	0.052	0.095	0.549	0.583	-0.135	0.240
Phone H	0.416	0.075	5.520	0.000	0.268	0.564
Phone N	0.308	0.099	3.107	0.002	0.114	0.503
Occup. Female	-0.423	0.166	-2.538	0.011	-0.749	-0.096
Occup. Male	-0.292	0.102	-2.857	0.004	-0.493	-0.092
Gender	-0.458	0.078	-5.912	0.000	-0.610	-0.306

Model's fit statistics:

Pseudo R^2	AIC	BIC	Log-Likelihood
0.354	8817.9166	9133.5072	-4375.0

CP is Car Price; DP is Down Payment; CA is Car Age; ToE is Duration (in years) of Employment; NI is Net Income; Marital D and S and W are Divorced, Single, Widow respectively; Phone C and H and N are Company phone number, house phone number, no phone number provided respectively; Occup. Female and Male are Professional Occupation as Female and Male, respectively.

three sampling settings among each balancing strategy. This might indicate that removing discrimination feature does not affect the model performance (or lender in general) at all, later, we will illustrate that this initial misconception leads to discrimination in the final stage of loan application below.

With segmented models for male and female applications (M3 and M4), the M3 or M4 taken separately are not directly comparable with M1 or M2 since they are estimated on different samples. To make a comparison possible the predicted values of M3 and M4 are combined together in the third column of Table 3.4. Across all models and balancing strategies, the combination of segmented models does not improve and, even in some cases, weaken (under target balancing strategy) classifier performance as compared with full models (M1, M2). When measuring the model performance separately on Male and Female segments, on Male segment, the models estimated on full data perform better than the model estimated in men only. In contrast, on Female segment, segmented model provides higher AUC compared with models built on full data. We attribute these results to the imbalanced nature of our data, as we have much lower number of female applications, this leads to (i) classifiers built on female segment fit better with female test set, and (ii) gender balancing produces almost identical performance on segmented data.

3.3.3 Rejection rate for unmarried customers

In what follows, we present several graphs in Figure 3.3 on the rejection rates of unmarried customers to robustly examine the impact of the removal of Gender in the final stage of credit application. Due to space limitation, we do not present that for Gender balancing strategy, although the results are similar with target balancing. The rejection rates are computed by varying the overall acceptance threshold from 0 (do not accept any applications) to 1 (accept all applications) and they are presented on the x-axis. The y-axis shows the proportions of men and women rejected under model M1 and M2, for each threshold. The left, middle, and right subfigures are the graphs for the corresponding classifier under no sampling, undersampling and oversampling methods, respectively. In general, the rejection rates for female applicants for M2 model (without gender) are significantly lower than that of M1 (with gender) for: logistic regression using all three samplings strategies, na ive bayes and decision tree using no sampling and oversampling, MLP using undersampling. No significant difference is observed for XGB. This indicates that (i) gender removal, especially in applying for traditional classifiers is actually disadvantaging female applications rather than protect them, and (ii) advanced classifiers including MLP or XGB are less prone to this discrimination, however, they are limited in their interpretability.

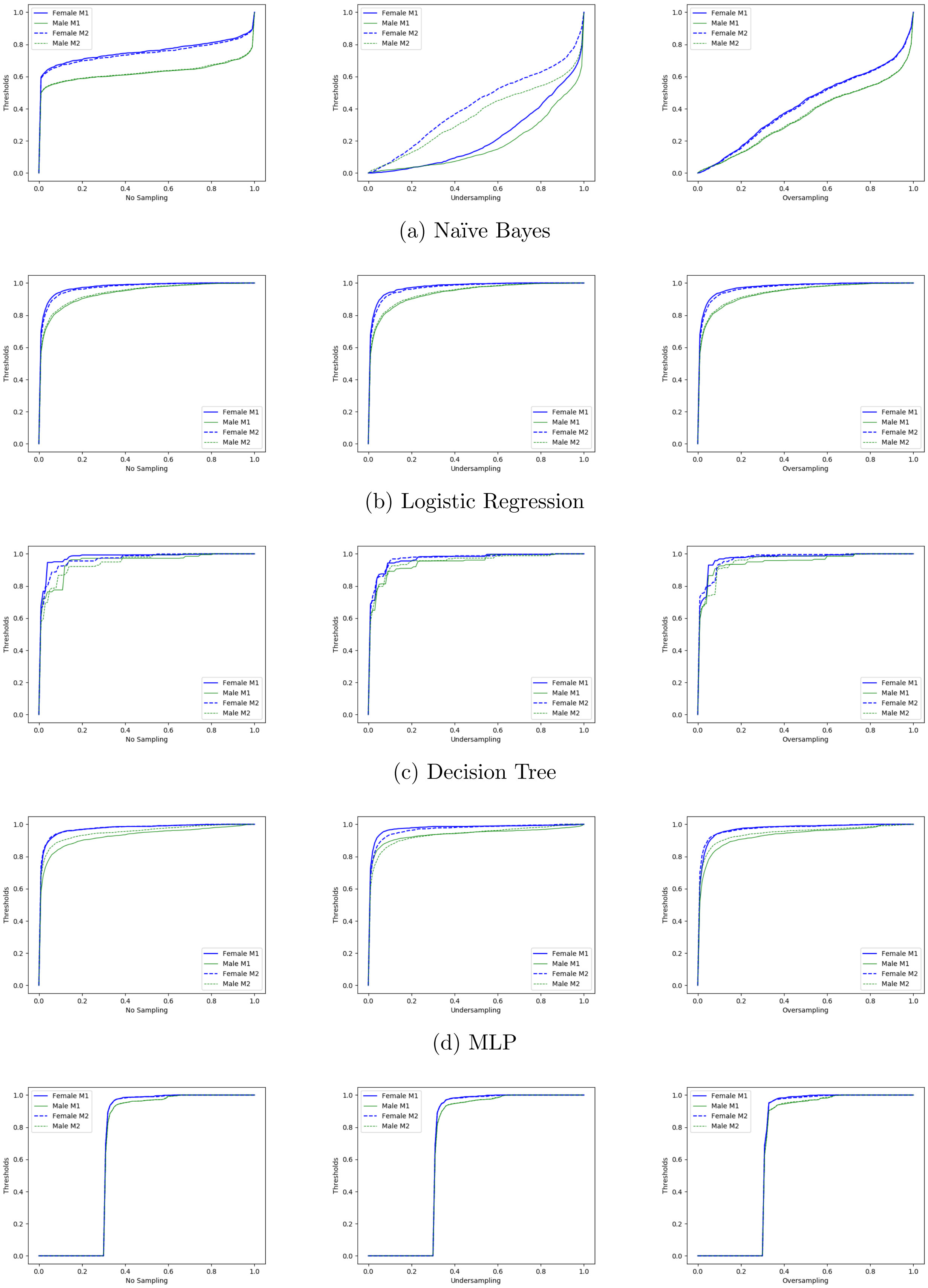
Overall, we document the following findings:

- Multilayer perceptron (MLP) shows the strongest performance across train/test sets and balancing settings, and is followed closely by logistic regression, although this result is reversed for the gender balancing strategy.

Table 3.4: AUC (Bold faces indicate highest values in the respective columns)

	Full Data			Male Segment			Female Segment		
	M1_Full	M2_Full	M3M4	M1_Male	M2_Male	M3	M1_Female	M2_Female	M4
PANEL 1: TARGET BALANCING									
No Sampling									
NB	0.895	0.895	0.838	0.914	0.914	0.863	0.824	0.824	0.902
LR	**0.906**	**0.907**	0.896	**0.929**	**0.929**	0.858	0.825	0.825	**0.936**
DTC	0.880	0.891	0.856	0.903	0.913	0.830	0.795	0.813	0.880
MLP	0.905	0.897	**0.913**	0.922	0.913	**0.901**	**0.842**	**0.836**	0.926
XGB	0.897	0.896	0.886	0.919	0.918	0.893	0.819	0.818	0.845
Undersampling									
NB	0.881	0.882	0.820	0.899	0.899	0.863	0.816	0.816	0.901
LR	0.907	0.908	0.895	0.931	0.930	0.858	0.827	0.827	**0.932**
DTC	0.907	0.884	0.873	0.924	0.905	0.870	0.846	0.810	0.879
MLP	**0.930**	**0.932**	**0.916**	**0.943**	**0.940**	**0.910**	**0.879**	**0.903**	0.922
XGB	0.903	0.904	0.870	0.923	0.924	0.875	0.832	0.832	0.896
Oversampling									
NB	0.881	0.881	0.842	0.898	0.898	0.872	0.816	0.816	0.902
LR	0.907	0.908	0.896	0.931	0.930	0.859	0.827	0.826	0.934
DTC	0.913	0.869	0.868	0.927	0.888	0.843	**0.861**	0.804	0.885
MLP	**0.916**	**0.917**	**0.913**	**0.934**	**0.934**	**0.890**	0.849	**0.850**	**0.936**
XGB	0.902	0.901	0.898	0.920	0.919	0.870	0.836	0.836	0.922
PANEL 2: GENDER BALANCING									
No Sampling									
NB	0.898	0.898	0.865	0.907	0.907	0.907	0.862	0.862	0.860
LR	**0.921**	0.922	**0.923**	**0.934**	0.934	**0.934**	**0.876**	**0.876**	**0.879**
DTC	0.891	0.892	0.890	0.906	0.907	0.905	0.836	0.839	0.803
MLP	0.916	**0.923**	0.899	0.929	**0.936**	0.912	0.869	0.874	0.854
XGB	0.909	0.909	0.897	0.926	0.926	0.924	0.848	0.848	0.823
Undersampling									
NB	0.897	0.897	0.865	0.907	0.907	0.907	0.861	0.861	0.856
LR	**0.922**	**0.923**	**0.923**	**0.935**	**0.935**	**0.934**	**0.877**	**0.877**	**0.879**
DTC	0.880	0.873	0.888	0.896	0.888	0.907	0.822	0.819	0.800
MLP	0.916	0.916	0.916	0.930	0.928	0.929	0.866	0.868	0.865
XGB	0.903	0.901	0.897	0.920	0.919	0.924	0.843	0.839	0.823
Oversampling									
NB	0.897	0.897	0.865	0.907	0.907	0.907	0.863	0.863	0.860
LR	**0.923**	**0.923**	**0.923**	**0.936**	**0.936**	**0.934**	**0.877**	**0.877**	**0.879**
DTC	0.886	0.888	0.886	0.903	0.904	0.905	0.826	0.831	0.798
MLP	0.912	0.916	0.916	0.926	0.930	0.933	0.862	0.865	0.8507
XGB	0.907	0.906	0.897	0.924	0.923	0.924	0.845	0.844	0.823

M1 and M2 are models estimated on full data with Gender removed on M2. M3 and M4 are models estimated on Male and Female segments, respectively. M1_Male and M2_Male are the M1 and M2 models performance on Male segment. M1_Female and M2_Female are the M1 and M2 models performance on Female segment. M3M4 is the combination of M3 and M4 models to make it comparable with M1 and M2 which are estimated using full data.



(a) Naïve Bayes

(b) Logistic Regression

(c) Decision Tree

(d) MLP

(e) XGB

Figure 3.3: Rejection rate

- Balancing settings modestly improve full model performance, but not for the models built on segmented data.

- Balancing towards Gender does not improve performance for models built on full data neither for ones built on segmented data. However, it produces better performance than balancing towards the target.

- Both in no sampling and oversampling, DTC, MLP, and XGB do not reduce rejection rate for females if we retain gender while the results for undersampling are mixed.

- The removal of Gender does not make the reject rates equal for both sexes for three out of five classifiers (except for Decision Tree and XGB, although there are still some areas where the small difference remains).

- For the same threshold, removing gender increases the rejection rate for women while decreases it for men.

3.3.4 Conclusions

Using the data on car loans with imbalanced classes, this study examined the utilisation of different automated, machine learning-based decision-making tools and balancing strategies on achieving equal outcomes for men and women.

We considered four model specifications: Model 1 with gender and Model 2 without gender, and two segmented models, Model 3 and 4, for men and women, respectively. For each specification we experimented with five different classifiers (Multilayer Perceptron, Extreme Gradient Boosting, Decision Tree, Naïve Bayes, Logistic Regression). We also balance the data with two standard balancing strategies (undersampling and oversampling) towards the target - bad loan indicator and also towards the protected characteristic of interest gender.

This study shows that the target balancing improves the predictive performance of most classifiers, and especially undersampling for multilayer perceptron. However, gender balancing has little effect on predictive accuracy. Neither target nor gender balancing has a pronounced effect on achieving the equal outcome. Nevertheless, application of Extreme Gradient Boosting based on decision tree showed the promising results for removing the gap in rejection rates between men and women, and further research could explore this in more detail.

Chapter 4

Imbalanced Dataset and Performance Evaluation in Credit Risk

4.1 Nature of Problem

A crucial problem of current practices in credit risk modelling is the imbalanced dataset (IDS) where classifiers tend to perform poorly on minority class despite presenting high overall accuracy measures. The IDS exists in many industrial data especially in credit portfolios data since the majority of corporate borrowers are good. To handle this problem, at algorithm level, we could use cost sensitive learning, at data level, researchers could perform undersampling or oversampling either as duplicating the minority class or creating synthetic samples.

In addition, combining many model advantages using ensemble methods is of the current trend since this can not only increase the classification performance but ensemble models could also reduce the subjectivity and increase the objectivity in risk assessment. In this line of literature, Lessmann et al. (2015) showed the advantages of simple ensemble models with individual classifiers and other ensemble families. Xiao et al. (2016) used a dynamic dynamic ensemble selection of features to result in better performance compares with the static ensemble selection. However, many works on corporate credit scoring consist of small number of features as in Angelini et al. (2008), Tsai and Chen (2010), or small number of samples Huang et al. (2004), Yu et al. (2008), Wang et al. (2011), and the data is not new and this might not be appropriate in data-driven practice as in credit scoring.

This chapter presents a comparative study on recent, skewed, and rich-feature corporate datasets to determine the performance of both individual and ensemble models on predicting the bad loan applicants. And to further investigate the applicability of imbalanced treatment methods at the data level, I also perform the comparison using several performance measurements under three sampling strategies.

4.2 Sampling Techniques

The credit application data are one of the sources that produce high-degrees of IDS where almost all observations belong to the majority class - good applications, and the remaining belong to the minority class - bad applications. Sampling approaches help by balancing the two classes. The popular methods are random oversampling the minority class, undersampling the majority class, or synthesizes artificial data in the minority class (He and Garcia, 2009; Chawla et al., 2002; Menardi and Torelli, 2014).

As creating artificial samples might lead to unrealistic samples, this study focuses on undersampling and oversampling and then relatively compare them under several performance metrics described below.

4.3 Experimental Setup

A long with logistic regression, an industrial standard, this study employs the best performing classifier from the work of Baesens et al. (2003) and its extended version of Lessmann et al. (2015), their works performed a general comparison of recent credit scoring models where individual, homogeneous ensemble, and heterogeneous ensemble classifiers are examined. The best individual classifiers are logistic regression (LR), decision tree (DTC), support vector machine (SVM), and multilayer perceptron (MLP). For the homogeneous ensemble models, we employ bagged and boosted version of the weak individual classifiers. For the heterogeneous ensemble classifiers, we employ the simple average ensemble (SAE) and hill-climbing ensemble selection (HES). Generally, twelve classifiers belonging to linear-based, kernel-based, tree-based, homogeneous, and heterogeneous ensemble-based classes are examined in this research, Table 4.1 presents more detail about these classifiers.

Table 4.1: Classifier summary

	Abbrev.	#Models
Individual Classifier		
Logistic Regression	LR	5
Decision Tree	DT	64
Support Vector Machine	SVM	16
Multilayer Perceptron	MLP	64
Homogeneous Classifier		
Bagged Logistic Regression	BaLR	5
Bagged Tree	RF	64
Bagged SVM	BaSVM	16
Bagged MLP	BaMLP	64
Boosted LR	BoLR	5
Boosted DT	BoDT	64
Heterogeneous Classifier		
Simple Average Ensemble	SAE	1
Hill-climbing Selection Ensemble	HSE	149*
Total		
12		617

* HSE includes all the estimations of base classifiers before fitting to the validation set for final classifier selection

Classifier	Parameter	Value
LR	$l2$ regularization	$1e{-}2$, $1e{-}1$, 1.0, 5, 10
DT	max depth max features	3,...,10 3,...,10
SVM	C γ	$1e{-}1$, 1 10, 5, 2, 1, $1e{-}1$,$1e{-}2$,$1e{-}3$,$1e{-}4$
MLP	#node in hidden layer 1 #node in hidden layer 2	2,...,8 2,...,8

Table 4.2: Hyper-parameters of individual classifiers

4.3.1 Assessment metrics

In financial industry, the loss resulted from bad loans might wipe out all the interest profit of the entire loan portfolio. Hence, it is crucial that the classifier for credit scoring need to maintaining minimum error on potential bad loans while not being too conservative on others (He and Garcia, 2009). Further, by simply predict all observations as majority class, the accuracy or categorical error might become misleading in the present of IDS without other proper measurements. Therefore, more informative metrics including the area under receiver operating characteristics curves (AUC, independent with class distribution), area under Precision-Recall curves (AUPRC, pay more attention to characteristics of Precision-Recall curve) (Davis and Goadrich, 2006), or Recall (focus on misclassification of minority class) are necessary for the consistently and concisely evaluations of classifier performance in the presence of IDS on corporate credit scoring.

In this study, we employ five metrics to:

- assess the correctness of the model categorical predictions, including:
  - Precision:
  $$P = Pr(Y = 1|\hat{Y} = 1).$$
  - Recall:
  $$R = Pr(\hat{Y} = 1|Y = 1).$$
  where 1 denotes a default class, 0 denotes a non default class, and $\hat{Y}$ is the estimate of the true class label Y.

- assess the discriminatory ability of the classifier:
  - Area Under ROC curve (AUC). This metric shows the probability of the classifier rank a randomly chosen bad loanx^+ higher than a randomly chosen normal loan (x^-), i.e.:
  $$\text{AUC} = p\big(score(x^+) > score(x^-)\big).$$
  - Area Under Precision-Recall Curve (AUPRC). How meaningful is a default loan application predicted by the classifier given the baseline probabilities of loan assessment problem:
  $$\text{AUPRC} = \sum_n (R_n - R_{n-1}) P_n.$$

where P_n and R_n are the precision and recall at the n^{th} threshold.

- assess the accuracy of the classifier probability predictions: Brier Score (BS), it can be computed using the following formula:

$$BS = \frac{1}{N}\sum_{t=1}^{N}(predict_t - o_t)^2.$$

where $predict_t$ is the predicted probability and o_t is the actual label of the observation t, respectively.

4.3.2 Test of significant in ranking classifiers

To compare the classifiers performance, the Friedman test is usually employed for comparing the ranks of many classifiers, and after these tests showing significant, a post-hoc test such as Neymenyi is then performed to compare all classifiers. Finally, when comparing other classifiers with a control classifier, this study follows Bonferroni correction or Hommel procedure (Garca et al., 2010) to account for the family-wise error.

To compare performance of different classifiers, we use the traditional Friedman test (Demar, 2006) to test H_0 : no difference between the classifier ranks. The test statistic is as follows:

$$\chi_F^2 = \frac{12D}{K(K+1)}\Big[\sum_{j=1}^{K} AR_j^2 - \frac{K(K+1)^2}{4}\Big], \tag{4.1}$$

with:

- D is the number of datasets from IDS treatment,
- K is the number of classifiers,
- $AR_j = \frac{1}{D}\sum_{i=1}^{D} r_i^j$, r_i^j is the rank of classifier j on sampling method i, and
- χ_F^2 has Chi-square distribution with $K-1$ degree of freedom.

We then employ the post-hoc Nemeyi test to perform pair-wise comparison of the individual classifiers (Garca et al., 2010). Two classifiers performance are significantly different if their average ranks differ by at least the following critical difference:

$$CD = q_\alpha\sqrt{\frac{K(K+1)}{6D}}. \tag{4.2}$$

Where critical values q_α are based on the Studentised range statistic, let R_i and R_j be the average ranks of classifier i and j, the test statistic is:

$$Z = (R_i - R_j) \Big/ \sqrt{\frac{K(K+1)}{6D}}. \tag{4.3}$$

Finally, when compare several individual classifiers with the best classifier to control for the family-wise error this study follows the Bonferroni-Dunn (Demar, 2006) procedure.

4.3.3 Datasets

Two practical datasets are employed in this study. The first is Vietnamese dataset, consists of 7,316 applications of corporate borrowers of a local bank. From financial statements, I calculate and divide financial ratios into 7 categories: Cashflow, Day Sale Coverage Ratio (DSCR), Efficiency, Leverage, Liquidity, Profitability, and Return. Table 4.3 shows a list of 189 financial ratios under each category. Along with the quantitative-financial variables, Table 4.4 presents other nine qualitative variables relating to the status of the creditworthiness from credit information center (CIC), historical credit profile of the borrower, and their demographic characteristics.

Table 4.3: Quantitative variables

Category	**#Financial Ratios**
Cashflow	4
DSCR (Day Sale Coverage Ratios)	31
Efficiency	30
Leverage	28
Liquidity	18
Profitability	24
Return	42
Total	177

Table 4.4: Qualitative variables

Name	**Description**
Q1	Duration of oldest bank contract (month)
Q2	Duration of youngest bank contract (month)
Q3	# Inquiries to credit center during last 12 months
Q4	# Inquiries to credit center during last 6 months
Q5	# Inquiries to credit center during last 3 months
Q6	# Years of establishment
Q7	# Banks in contract
Q8	Outstanding debt at all banks
Q9	Geographical area

The other dataset is from the U.S. Small Business Administration (SBA), this provides historical data from 1987 through 2014. This US SMEs dataset has 899,164 observations with 27 features. Each observation is a loan that was partially guaranteed by the SBA. The variable MIS_Status indicates if the loan was paid-in-full or defaulted/charged-off. This study randomly chooses 1%

sample of this dataset (preserving the good:bad ratio) which results in 1,549 default loans and 7,332 paid-in-full loans. Table 4.5 shows the features of this dataset.

Table 4.5: US dataset variables

Name	**Type**	**Description**
NewExist	Boolean	New or existing Business
FranchiseCode	Boolean	Franchise Code
LowDoc	Boolean	LowDoc Loan Program
UrbanRural	Categorical	Urban, Rural, Undefined
DisbursementGross	Numeric	Amount Disbursed
BalanceGross	Numeric	Gross amount outstanding
GrAppv	Numeric	Gross Amount of Loan
SBA_Appv	Numeric	SBA Approved Loan
Term	Numeric	Loan term in months
NoEmp	Numeric	Number of Employees
CreateJob	Numeric	Number of jobs created
RetainedJob	Numeric	Number of jobs retained
State	Text	Borrower State
Bank	Text	Bank Name
BankState	Text	Bank State
NAICS	Text	Industry code
RevLineCr	Text	Revolving Line of Credit
MIS_Status	Text	Loan Status

Table 4.6: Imbalance degree

	#Sample	**#Default**	**Imbalanced Ratio**
VN	7,316	375	5:95
US	8,881	1,549	17:83

The good:bad ratios in Table 4.6 show that while the US data have a low imbalanced ratio, the Vietnamese dataset is extremely imbalanced with the fraction of bad applications is 4.7%. Figure 4.1 and Figure 4.2 present the scatter plot of the second component against the first component for both datasets. While the scatter for VN dataset does not shows any particular cluster pattern, it does reveal three clusters for US dataset, one of them on the bottom has significantly less number of bad loans than the others, this could suggest the classification task for US dataset might be less tricky than that for the VN dataset.

4.4 Results

4.4.1 Classifier ranking

Table 4.7 presents the performance of classifiers for the original datasets. Precision and recall for the Vietnamese dataset are extremely low than the US dataset which translates to almost no bad loan applications were detected on the test set, which also means that the problem of imbalanced class is more severe for the Vietnamese dataset compares with that of the US dataset. Ensemble classifiers perform poorly both in homogeneous and heterogeneous settings. First, there is no significant improvement using bagging ensemble in term of recall and precision. Nevertheless, boosting methods including BoostedLR and BoostedDT show slight improvement

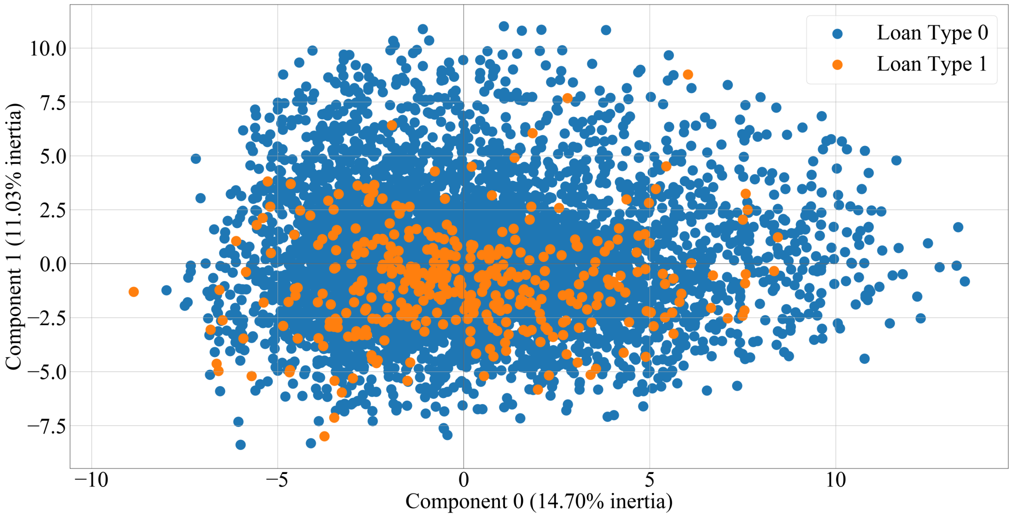


Figure 4.1: PCA VN dataset: Loan Type 0 and 1 are the Non-Default and Default customers

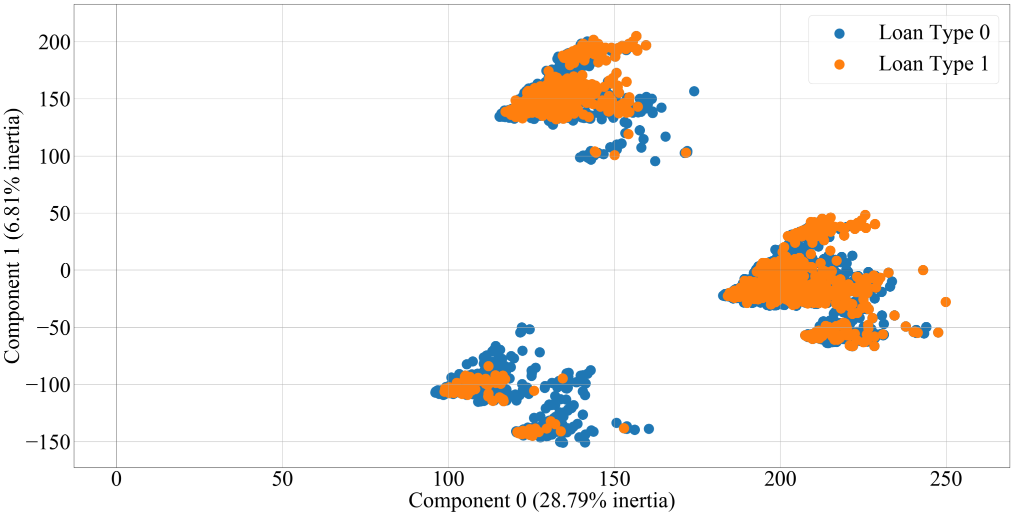


Figure 4.2: PCA US dataset: Loan Type 0 and 1 are the Non-Default and Default customers

on Precision and Recall but lower on the other performance metrics compared with their base classifiers. Second, heterogeneous ensembles, despite presenting modest improvements in AUC, AUPRC, and BS, they are struggle with extremely low Recall and Precision. This provides supporting evidence for the low performance of corporate credit scoring models on the present of high IDS.

Table 4.7: Performance of classifiers on the original dataset

	AUC	AUPRC	BS	P	R
Panel 1: VN Dataset					
LR	0.585	0.070	0.049	0.000	0.000
DT	0.589	0.070	0.053	0.015	0.053
SVM	0.555	0.061	0.049	0.000	0.000
MLP	0.547	0.073	0.049	0.000	0.000
BaLR	0.584	0.071	0.049	0.000	0.000
RF	0.590	0.074	0.049	0.000	0.000
BaSVM	0.495	0.054	0.049	0.000	0.000
BaMLP	0.582	0.075	0.048	0.000	0.000
BoLR	0.572	0.069	0.249	**0.117**	**0.068**
BoDT	0.580	0.070	0.135	0.014	0.046
SAE	**0.610**	**0.083**	0.053	0.000	0.000
HSE	0.587	0.072	0.049	0.000	0.000
Panel 2: US Dataset					
LR	0.834	0.550	0.107	0.419	0.688
DT	0.841	0.566	0.101	0.578	0.654
SVM	0.875	0.688	0.087	0.568	0.769
MLP	0.873	0.661	0.090	0.624	0.702
BaLR	0.813	0.530	0.119	0.051	0.605
RF	**0.920**	**0.765**	**0.075**	**0.682**	0.793
BaSVM	0.855	0.628	0.111	0.163	**0.863**
BaMLP	0.851	0.637	0.103	0.367	0.827
BoLR	0.818	0.503	0.210	0.289	0.643
BoDT	0.905	0.703	0.153	0.666	0.714
SAE	0.903	0.725	0.093	0.565	0.834
HSE	0.891	0.706	0.084	0.624	0.754

Bold face and underline indicate the best and worst classifiers.

By contrast, with the imbalanced ratio of 17:83 for the US dataset, ensemble methods do provide an uplift in performance especially in precision and recall. Heterogeneous ensembles also present significant improvement in all metrics compared with the base classifiers. However, individual classifiers have competitive performance, they have higher precision but lower recall compare with bagging and boosting alternatives, with an exception from Random Forest. Random Forest presents the best performance in 4 over 5 metrics compare with other classifiers.

Table 4.8 shows the ranking of classifiers in three sampling methods and five performance measurements.

The non-parametric Friedman test statistics in Table 4.8 reject the null-hypothesis, and there is significant difference between the classifier ranks for all metrics, I then proceed with the Nemenyi *post hoc* test for the pair-wise comparison of all classifiers. Figure 4.3 presents the critical differences:

The performance of the best classifier, RF, is significantly better than LR family, an industry standard. SAE is better than BaggedLR and BoostedLR but not significantly better than LR.

Table 4.8: Classifiers ranks

C	AUC	AUPRC	BS	P	R	AR	FR
RF	1.4	1.5	2.4	2.0	3.5	2.2	1
SAE	1.8	2.1	4.4	4.4	2.5	3.0	2
HSE	5.4	5.1	3.9	7.2	6.8	5.7	3
BaMLP	6.0	4.8	6.6	6.6	6.8	6.2	4.5
BoDT	6.4	6.4	7.8	5.5	4.8	6.2	4.5
DT	7.6	9.4	7.1	4.5	6.1	7.0	6
MLP	7.6	6.2	7.2	6.6	8.6	7.3	7
SVM	9.0	8.6	4.0	8.8	7.1	7.5	8.5
BaSVM	9.2	8.0	5.6	9.7	4.9	7.5	8.5
LR	7.1	8.4	7.8	7.2	8.5	7.8	10
BaLR	7.9	8.2	9.6	8.2	9.1	8.6	11
BoLR	8.6	9.2	11.6	7.1	9.4	9.2	12
χ^2_{12}	45.2	48.2	46.5	30.7	33.7		
p	$4e{-}6$	$1e{-}6$	$2e{-}6$	$1e{-}3$	$4e{-}4$		

AR and FR stand for Average Rank and Final Rank. We sort the classifier performance for the final rank decreasingly and we use the average method in case of ties. Average rank is the mean rank of each classifier rank across performance measurements and sampling methods. p is p-value of the non-parametric Friedman test statistics

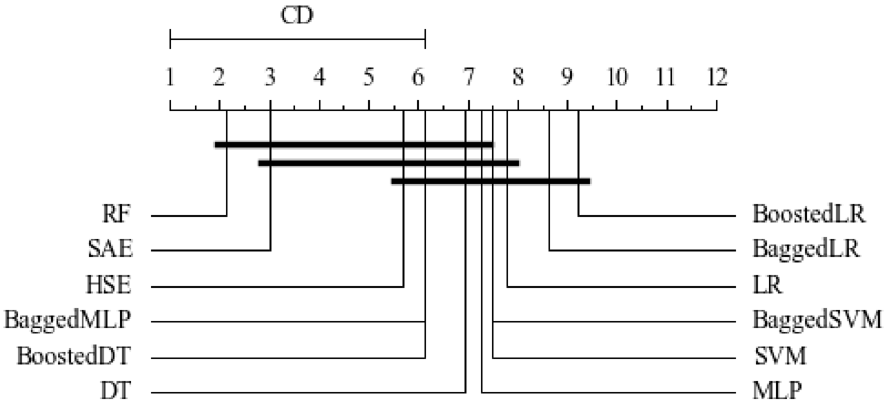


Figure 4.3: Critical difference

We then perform p-value adjustment with Bonferroni-Dunn (Demar, 2006) procedure to compare the best classifier with selected classifiers including a heterogeneous ensemble (SAE), a tree-based (DT), a kernel-based (SVM), a network-based (MLP), and an industry-standard (LR). Table 4.9 presents the adjusted p-values with Bonferroni-Dunn procedure.

Table 4.9: Comparison of RF with other classifiers

p-values (Bonferroni-Dunn (Demar, 2006) Procedure)

SAE	**DT**	**SVM**	**MLP**	**LR**
0.0538	0.0000	0.0000	0.0000	0.0000

Despite RF shows significant difference in the ranks compared with DT, SVM, MLP, and LR. However, no significant difference in the ranks of RF and SAE is found.

4.4.2 Effects of sampling methods

Table 4.10 provides more insights on the performance of sampling methods. There are no significant differences in AUC and AUPRC by using these three sampling methods. This provides evidence to the current debate using several metrics to compare different classifiers as in Hand and Anagnostopoulos (2013).

Table 4.10: Two tail t-stat for performance of sampling compare with no sampling

	AUC	**AUPRC**	**BS**	**P**	**R**
VN Dataset					
UNDER	-0.091	0.491	-9.065	-10.308	-6.736
	0.929	0.628	0.000	0.000	0.000
OVER	0.499	0.259	-2.558	-3.734	-4.041
	0.623	0.798	0.018	0.001	0.001
SMOTE	-0.970	-0.884	-3.535	-7.133	-7.090
	0.343	0.386	0.002	0.000	0.000
US Dataset					
UNDER	0.731	1.611	-4.008	-1.420	9.477
	0.473	0.121	0.001	0.170	0.000
OVER	1.328	1.354	-1.962	-0.310	3.513
	0.198	0.189	0.063	0.760	0.002
SMOTE	0.095	0.480	-1.564	-2.005	4.578
	0.925	0.636	0.132	0.057	0.000

Underline values indicate H_0-no significant difference is rejected. Under each sampling method, we compare the performance with no sampling setting, the first row presents the two-tail test statistics and the second row provides the accompanied p-values.

For the remaining three metrics, there are mixed results for two datasets, as for the VN dataset, using sampling methods improves Brier Score, Precision, and Recall. However, it is not the case for the US dataset when Recall decreases, this could be partially explained by the modest degree of IDS for the US dataset. This suggests further generalised examination on the effects of imbalance ratios to classifier performance.

4.5 Conclusions

This study provides recent evidence on the comparison of scoring models for corporate credit risk modelling on several balancing strategies and performance measurements with two imbalanced corporate loan datasets. Specifically, this study examines under three sampling strategies and five performance metrics the performance of 12 classifiers belong to linear-based, kernel-based, tree-based, homogeneous ensemble-based, and heterogeneous ensemble-based classes. The results first shows many classifiers tend to be overfitted toward the majority class through the extremely low precision and recall. Second, as classifiers perform extremely poor for the minority class - bad applications, it is advisable to have diverse measurements that center effectively on (i) probability predictions and (ii) categorical prediction (i.e. on minority class).

On this merit, sampling methods could be employed to enhance categorical prediction and using simple averaging method for heterogeneous classifiers could improve the performance.

This suggests further works to focus on examining the performance of sampling methods with different degree of imbalanced ratio and multi-source learning models.

Chapter 5

Textual Analysis in Credit Risk

As of June 4^{th}, an extended version of this chapter is accepted to present at ECML-PKDD2020.

5.1 Overview of Textual Data Sources

Using the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system, the United States Securities and Exchanges Commission (SEC) requires listed US enterprises to file their financial reports with the 10-Ks and 10-Qs forms for annual and quarterly reports, respectively. In these forms, along with the detail financial accounting statements, enterprises need also to include their Management Discussion & Analysis[1] (MD&A), which is a forward-looking statement. In that section, the top managers or the boards need to explain their company's performance, address the compliance & risks, and express their views on the company future goals and projects.

Together with examining and monitoring the company performance through traditional financial statements, these textual data provide a rich-feature dataset that could be effectively exploited for understanding the evolution of embedded aspects in manager sentiment and further assist on other predictive tasks. In addition, common predictive tasks such as stock return or firm liquidations experience many limitations if they rely solely in accounting data because of the corporate window dressing to enhance their credit or financial performance quality (Guan et al., 2008; Gandhi et al., 2019), let alone the benefit of building supportive systems based on the alternative data to complement the traditional predicting or scoring systems.

Most current works in the textual analysis in the financial industry were quantifying the textual data to (i) form the predictors for future company financial performance by explaining the manager sentiment and stock returns (Nguyen et al., 2015; Lopez Lira, 2019); (ii) understand

[1] https://www.sec.gov/corpfin/cf-manual/topic-9

the role of investment analyst report (Huang et al., 2018); or (iii) improve the manager sentiment tone understanding (Zhou, 2018). However, little works have been paid to understanding what these predictors actually represent for and how they evolve through economic cycles and crisis. Our study contributes to the current literature of textual analysis for the financial reports by first uncovering the latent topics from the MD&A of the filings, and then further examining the predictive power of those textual topics and investigating how they could improve the traditional bankruptcy prediction methods. Specifically, by using Latent Dirichlet Allocation (Blei et al., 2003) and Dynamic Embedding Topic Models (Dieng et al., 2019b), we find the hidden, yet interpretable topics in the MD&A section and show that they are significant and deliver comparable predictive performance just only using the MD&A textual data. More importantly, the topics show effective complement role to the traditional z-score based bankruptcy prediction model.

By and large, this study objectives are first to explore the hidden topics in the management's reflection on corporate business, then examine how the latent topics from textual data evolve, and qualitatively on how the MD&A contents inline with market movement. By using more than 20 years of textual data in the 10-K filings of US listed firms collected from EDGAR from 1997 to 2016, we found that there are potentially 30 topics embedded in the MD&A and illustrated their evolution overtime. Our study further enhances the industrial accounting-based bankruptcy prediction model with the MD&A data, we demonstrate the significant of the topics by a comparative experiment on the prediction of firm bankruptcy taking into account of several well-known baselines both in financial-accounting and natural language processing literature.

In what follows, we first present the relevant literature on the topic modelling and the textual analysis in liquidation modelling in section 2. Section 3 devotes to the process of mining MD&As, estimating the latent topics, summary of the data, and experimental settings. The results are presented and discussed in section 4, and we conclude our work in section 5.

5.2 Backgrounds

In uncovering the latent topics embedded in textual data, the pioneer work of Blei et al. (2003) introduced latent Dirichlet allocation (LDA) which uses the variational Bayesian inference to infer the latent topics from large corpus. Then, based on an intuitive that the document collections should reflect evolving content, Blei and Lafferty (2006) proposed the dynamic topic models (DTM), which is a dynamic version of LDA to examine the evolution of topics overtime, and showed the superior performance in term of likelihood for the hold-out dataset compare with the traditional LDA. However, as existing topic models fail to learn interpretable topics when working with large and heavy-tailed vocabularies, Embedding Topic Models (Dieng et al., 2019a) (ETM) bridges this gap by utilising the word embeddings (skip-gram, Mikolov et al. (2013a)). ETM incorporates the embeddings into the inference procedure of the traditional LDA. Specifically, it combines traditional topic models with word embeddings and models each word with a categorical distribution whose natural parameter is the inner product between a word embedding and an embedding of its assigned topic. ETM discovers interpretable topics even with large, imbalanced vocabularies that include rare words and stop words.

To effectively model topic evolution in embedding spaces, Dieng et al. (2019b) further introduced Dynamic Embedding Topic Models (DETM) which combines the ETM with DTM by modelling each word with a categorical distribution parameterised by the inner product between the word embedding and a per-time-step embedding representation of its assigned topic. The DETM learns smooth topic trajectories by defining a random walk prior over the embedding representations of the topics. DETM is fitted using structured amortized variational inference with a recurrent neural network.

In applications of topic modelling for financial data, one of the most initial work in summarising the textual financial data is from Bao and Datta (2014), where they employed the modified LDA to fit the sentence level analysis with the assumption of one topic for each sentence. Their empirical analysis focused on the risk-disclosure in the filings, and by unsupervised learning, they found 30 risk types (topics), and among them there are new and significant risk types to predict the risk perceptions of investors, significantly, they discovered five more important topics compare with a large-scaled supervised learning in the work of Huang and Li (2011). In the application of topic modelling in stock market analysis, Nguyen et al. (2015) utilised the sentiment analysis, specifically, they analysed the financial social media data using a combination of topic modelling and sentiment classifier in predicting stock price movements, they showed that their ensemble model achieves better performance in predicting stock price movements compared with the traditional time series and human sentiment methods.

Recently, Zhou (2018) and Jiang et al. (2019) analysed and examined the relationship between homogeneous and heterogeneous sources of financial texts and indicated that higher manager sentiment followed by lower earnings disruption and higher investment growth. To our knowledge, there are two works in examining the liquidation of firms by utilizing the textual data, which are Gandhi et al. (2019) and Mai et al. (2019). The first one used sentiment words to examine the financial distress of the US banks, their findings suggested that more negative words in the reports are related to a higher probability of distressed delisting subsequently, and the latter employed deep learning model, specifically the convolutional neural network model, to examine the gain in predictive power using the word embedding techniques. However, neither the relationships between those sentiment words and the likelihood of liquidation have been investigated for non-banking sector data nor how the topics in textual reports relate to firm liquidation. In addition, since deep learning models is a black-box model which is very tricky to decipher its feature-constructing process as well as its classification output, little works have been devoted to the performance of classification models trained on more intuitive and interpretable textual features, and relatively compare them with the traditional ones. Hence, to better model how the latent topics from textual data evolve, and qualitatively analysing on how the MD&A contents inline with market movement, we employ DETM both as a explorer for hidden topic in MD&A of the financial filings, and as a feature engineering method for enhancing the traditional bankruptcy prediction model by leveraging the textual data.

5.3 Data and Methods

5.3.1 Data

Accounting data

In this study, we collect the accounting data from the Wharton Research Data Services (WRDS[2]) for all listed firms in the US from 1997 to 2016, the detail statistics of their financial report elements are presented in Table A.1 and A.2 of A. We exclude firms in the financial and regulated utilities sectors with SIC from 6000 to 6999 and from 4900 to 4949, respectively. As for the liquidation flags, a firm is marked as liquidation if it filed for liquidation under Chapter 7 or Chapter 11 bankruptcy filings[3]. We further divide the data for the small and medium enterprises (SME), non-SME, and all samples. We regard a corporate to the SME category if its sale less than or equal to $65 million (Altman and Sabato, 2007) which is in line with the Basel Capital Accord[4].

10-K filings

The raw filings of listed firms in the US could be retrieved from EDGAR[5]. After removing Tables, Figures, attached PDFs, and other redundant elements, we extract the MD&A section in each filing. We cover 10-K and 10-KSB (SB is for small business) filings in this study as other types of filing either notice a delay in document filings (10-K405) or a transition of the accounting period (10-KT and 10-K405T). We present the descriptive statistics of MD&A data extracted from SEC filings in Table 5.1.

At the final stage, we merge the financial data and SEC filings by matching the CIK and the fiscal year-end of financial reports. The final data consist of 51,128 firm-year observations, of which 213 firms are liquidation (approximately 0.42%). Using the SME definition in Section 3 above, we have the number of liquidations under each business segment as follows:

5.3.2 Latent Dirichlet Allocation

LDA (Blei et al., 2003) is considered as a generative probabilistic model, it assumes that each document in the corpus is represented as random mixtures over latent topics. Each topic is characterized by a distribution over terms in a vocabulary. The generative process is shown in Algorithm 1 where the meaning of each variable is described in Table 5.3.

[2] Licensed per subscription of Business School, The University of Edinburgh

[3] https://www.sec.gov/reportspubs/investor-publications/investorpubsbankrupthtm.html

[4] Basel Committee on Banking Supervision, June 2004.

[5] Electronic Data Gathering, Analysis, and Retrieval system - https://www.sec.gov/edgar/aboutedgar.htm

Table 5.1: MD&As extracted from the filings

Year	#MD&As	#NF	#unc	#omitted	#Sent.	mean	median	#tokens
1997	8711	422	7	561	740208	85	64	47328
1998	8931	454	4	563	862791	96.6	73	51372
1999	8934	391	2	611	1044325	116.9	89	54665
2000	9508	458	8	580	1065612	112.1	79	57295
2001	9447	424	1	510	1131682	119.8	83	59454
2002	10179	434	2	806	1488084	146.2	89	64979
2003	11878	419	6	1236	2107816	177.5	117	74079
2004	12124	390	4	1496	2274859	187.6	115	76837
2005	12475	635	2	2016	2501571	200.5	120	81210
2006	12251	678	5	1863	2472559	201.8	135	81163
2007	12087	485	5	1840	2526728	209	145	83147
2008	11432	443	6	1470	2519519	220.4	156	83350
2009	9919	366	4	769	2525077	254.6	189	79620
2010	9165	190	3	676	2405854	262.5	199	78967
2011	8840	162	2	659	2290261	259.1	193	78147
2012	8393	175	1	693	2214756	263.9	195	75322
2013	8105	186	1	677	2183919	269.5	203	74772
2014	8084	184	1	751	2193408	271.3	202	76434
2015	7985	182	1	912	2181273	273.2	204	76066
2016	7589	158	2	1081	2077774	273.8	201	74669
2017	7248	184	1	1113	1931412	266.5	192	71713
Total	203,285	7,420	68	20,833	40,739,488			

#MD&As is the total number of MD&As; #NF is the number of filings that do not have MD&A; #unc is the number of uncommon MD&As that we are unable to trace the sections they begin or end with; #omitted is the number of filings that have the MD&A section omitted; #Sent. is the total number of sentences of all MD&As; mean and median are the mean and median of number of sentence in MD&As; #tokens is the total number of unique words in all MD&As.

Table 5.2: Number of liquidations in two corporate segments

	non-SME	SME	All
1997	12	5	17
1998	11	9	20
1999	5	5	10
2000	8	4	12
2001	19	4	23
2002	21	14	35
2003	22	16	38
2004	13	13	26
2005	16	12	28
2006	7	10	17
2007	6	7	13
2008	7	5	12
2009	16	8	24
2010	2	3	5
2011	1	3	4
2012	5	3	8
2013	3	3	6
2014	3	0	3
2015	3	3	6
2016	1	5	6
Total	181	132	313

Terms	Description
M	# documents in the corpus (constant scalar)
K	# topics (constant scalar)
V	# terms in vocabulary (constant scalar)
N_d	length of the document d
α	hyper-parameter for topic proportions (vector in $\mathbb{R}^K$ space or a scalar if symmetric)
β	hyper-parameter for term proportion (vector in $\mathbb{R}^V$ or a scalar if symmetric)
θ	topic mixture proportion for document m. One proportion for each topic in the document
ϕ	term mixture proportion for topic k. One proportion for each term in the vocabulary
$w_{d,n}$	the term indicator for n^{th} word in document d
$z_{d,n}$	topic assignment of n^{th} word in the document d

Table 5.3: Notations and Terminologies for LDA Model

Algorithm 1 Generative Process of LDA

```
for each topic $k \in [1, K]$ do
    sample mixture proportion $\phi \sim Dir(\beta)$
end for
for each document $m \in [1, M]$ do
    Sample mixture proportion $\phi \sim Dir(\alpha)$
    Sample document length $N_m \sim Poiss(.)$
    for each word $n \in [1, N_m]$ do
        Sample topic index $z_{m,n} \sim Mult()$
        Sample a term for word $w_{m,n} \sim Mult(\phi_{z_{m,n}})$
    end for
end for
```

The main objectives of LDA model is to infer (1) the term distribution $p(t|z = k) = \vec{\phi}_k$ for each topic k; and (2) the topic distribution $p(z|doc = d) = \vec{\theta}_d$ for each document d. That is, we are interested in determining the joint distribution of the topic mixtures Θ, the set of topic assignment $\mathbf{Z}$, the terms in corpus $\mathbf{W}$, and the topic Φ, given the parameters α and β determined by:

$$P(\mathbf{Z}, \Theta, \Phi, \mathbf{W}|\alpha, \beta) = \prod_{i=1}^{K} P(\phi_k|\beta) \prod_{d=1}^{M} P(\theta_d|\alpha) \prod_{n=1}^{N_d} P(z_{d,t})P(w_{d,n}|\phi_{z_{d,n}})$$

Equivalently, we need to compute the posterior

$$P(\mathbf{Z}, \Theta, \Phi|\mathbf{W}, \alpha, \beta) = \frac{P(\mathbf{Z}, \Theta, \Phi, \mathbf{W}|\alpha, \beta)}{P(\mathbf{W}|\alpha, \beta)} \tag{5.1}$$

This distribution is intractable because of the calculation in denominator over all $\mathbf{Z}$, Θ, and Φ. Therefore, practitioners turn to approximate inference approaches. Two common possible methods are variational inference (VI) and Markov chain Monte Carlo sampling (MCMC) (e.g. Collapsed Gibb Sampling) (Hoffman et al., 2010). MCMC asymptotically approaches to the true posterior distribution. However, it is computationally expensive. In contrast to MCMC, VI tends to be faster and easier to scale to large data (Blei et al., 2017).

5.3.3 Evolution of topics in textual data

Blei and Lafferty (2006) introduced the dynamic topic model (DTM) which analyses the evolution of topics in large document collections over time. The method is to use state space models on the natural parameters of the multinomial distributions that represent the topics. The author also derived the variational approximation algorithm based on Kalman filters or non-parametric wavelet regression to estimate the posterior distribution over the latent topics. Essentially, DTM is an extension to LDA to adapt with sequential documents. LDA assumes that the word order within a document and the document order within the corpus are processed as in the same priority. In DTM, words are still assumed to be exchangeable, but the document order holds a vital role. Particularly, the documents are divided into groups by time slice (e.g. quarters, half-years, years) and DTM assumed that the documents in each group come from a set of latent topics that evolved from the ones in the previous time slice. In this paper, we employ the online variational inference (Hoffman et al., 2010) for the textual data of bigram tokens in our data to examine the optimal number of topics, we then use the dynamic embedding topic model to capture the evolution of financial topics over time. In the next section, we briefly review the main idea of dynamic topic model in embedding spaces.

5.3.4 Topic modelling in embedding spaces

Existing topic models fail to learn interpretable topics when working with large and heavy-tailed vocabularies. Embedding Topic Models (Dieng et al., 2019a) (ETM) bridge this gap by utilising the word embeddings (CBOW, Mikolov et al. (2013a)):

$$w_{dn} \sim \mathrm{softmax}(\rho^\top \alpha_{dn}). \tag{5.2}$$

The embedding matrix ρ is a $L \times V$ matrix whose columns contain the embedding representations of the vocabulary, $\rho_v \in \mathbb{R}^L$. The vector α_{dn} is the *context embedding*. The context embedding is the sum of the context embedding vectors (α_v for each word v) of the words surrounding w_{dn}.

And to effectively model the dynamic of topics in embedding spaces taken into consideration of the imbalanced word distribution and topic evolution, Dieng et al. (2019b) introduced Dynamic Embedding Topic Models (DETM) which inherits the strengths of the ETM and DTM by modelling each word as a categorical distribution parameterised by the inner product of the word embedding and a per-time-slice topic embedding.

Denote $\alpha_k^{(t)}$ as *topic embedding* (Dieng et al., 2019a) of the k^{th} topic at time slice t. In DETM, the probability of a word belongs to a topic is given by the (normalized) exponentiated inner product between the word and the topic's embedding at the corresponding time slice,

$$p(w_{dn} = v | z_{dn} = k, \alpha_k^{(t_d)}) \propto \exp\{\rho_v^\top \alpha_k^{(t_d)}\}. \tag{5.3}$$

The topic embeddings evolve under Gaussian noise with variance γ^2,

$$p(\alpha_k^{(t)} | \alpha_k^{(t-1)}) = \mathcal{N}(\alpha_k^{(t-1)}, \gamma^2 I). \tag{5.4}$$

The prior over θ_d depends on a latent variable η_{t_d}, where t_d is the time slice of document d:

$$p(\theta_d | \eta_{t_d}) = \mathcal{LN}(\eta_{t_d}, a^2 I) \tag{5.5}$$

where $p(\eta_t | \eta_{t-1}) = \mathcal{N}(\eta_{t-1}, \delta^2 I)$ and $\mathcal{LN}$ denotes a log-normal distribution. And the generative process of DETM is as follows:

Algorithm 2 Generative process of DETM

```
Draw initial topic embedding $\alpha_k^{(0)} \sim \mathcal{N}(0, I)$
Draw initial topic proportion mean $\eta_0 \sim \mathcal{N}(0, I)$
for time step $t = 1, \ldots, T$ do
    Draw topic embeddings $\alpha_k^{(t)} \sim \mathcal{N}(\alpha_k^{(t-1)}, \gamma^2 I)$ for $k = 1, \ldots, K$
    Draw topic proportion means $\eta_t \sim \mathcal{N}(\eta_{t-1}, \delta^2 I)$
end for
for each document $d \in \mathcal{D}$ do
    Draw topic proportions $\theta_d \sim \mathcal{LN}(\eta_{t_d}, a^2 I)$.
    for each word $n$ in the document do
        Draw topic assignment $z_{dn} \sim \mathrm{Cat}(\theta_d)$.
        Draw word $w_{dn} \sim \mathrm{Cat}(\mathrm{softmax}(\rho^\top \alpha_{z_{dn}}^{(t_d)}))$.
    end for
end for
```

The inference procedure in DETM is also made possible by optimising the Kullback-Leibler divergence of the approximation to the true posterior distribution $p(\theta, \eta, \alpha|\mathcal{D})$, in addition, the authors speed up algorithm via amortisation inference, an black box VI with the distribution over the topic proportions $q(\theta_d|\eta_{t_d}, \mathbf{w}_d)$ parameterised by a neural networks (either recurrent neural network or long short-term memory). In this paper, we employ DETM to create the hidden-topic vector representation of the textual data and use them as input features for predicting corporate bankruptcy.

5.3.5 Number of topics assessments

Identifying the optimal number of topic is of the most important task in topic modelling. We compute the perplexity of a held-out test set to evaluate the model with different setting of number of topics (Blei et al., 2003). The perplexity is monotonically decreasing in the likelihood of the test data, and it is algebraically equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity score indicates better generalisation performance. The perplexity of test set D_{test} of M documents is defined as follows (Blei et al., 2003; Řehůřek and Sojka, 2010):

$$perplexity(D_{test}) = 2^{\left\{-\frac{\sum_{d=1}^{M} log[p(\mathbf{w}_d)]}{\sum_{d=1}^{M} N_d}\right\}} \tag{5.6}$$

Since we cannot directly compute $log[p(\mathbf{w}_d)]$, we compute the lowerbound of perplexity (Hoffman et al., 2010):

$$perplexity(n^{test}, \lambda, \alpha) \leq 2^{\{-bound\}} \tag{5.7}$$

where

$$bound = \frac{\sum_i (\mathbb{E}_q[log_p(n_i^{test}, \theta_i, z_i|\alpha, \beta)] - \mathbb{E}_q[log_q(\theta_i, z_i)])}{\sum_{i,w} n_{iw}^{test}} \tag{5.8}$$

and n_i^{test} denotes the vector of word count of i^{th} document in a test set of M documents. The per-word perplexity in Eq. 5.8 is obtained by computing the probability of each word in the second half of a test document, conditioned on the first half.

5.3.6 Bankruptcy prediction feature sets

One of our experiment in this study is to examine to which extend, the hidden topics in MD&A, once inferred, could help on leveraging the real-world practice in bankruptcy prediction. Thus, along with our topic distributed representations, we compare the distributed representation of MD&A with the industrial standard feature set in predicting corporate bankruptcy which is z-score, we further relatively compare it with other baselines including (i) dictionary-based count vectorisation based on a financial dictionary, and (ii) traditional word embedding. Particularly, we construct the following feature sets:

- S1: Altman z-score with 5 factors (Altman, 1968) for the general corporate and 5 factors for the SMEs (Altman et al., 2010) bankruptcy prediction
- S2: Dictionary-based count (relative to length of MD&A) vectorisation of the sentiment wordlists in Loughran and McDonald, 2011 (Loughran and Mcdonald, 2011)
- S3: Distributed representation of MD&A using *doc2vec* (Le and Mikolov, 2014)
- S4: Topic representation for the MD&A using LDA model
- S5: Topic representation for the MD&A using DETM model

Specifically, the feature sets in S1 and S2 are as follows:

Table 5.4: Altman's 5-factor

Notation	**Formula**
Panel I	**z-score 5-factor (Altman et al., 1977)**
Z1	Working Capital / Total Assets
Z2	Retained Earnings / Total Assets
Z3	EBIT / Total Assets
Z4	Market Value of Equity / Total Liabilities
Z5	Sales / Total Assets
Panel II	**SME 5-factor (Altman et al., 2010)**
A1	Cash Flow from Operating Activities / Current Liabilities
A2	Short Term Debt / Equity Book Value
A3	Cash / Total Assets
A4	EBIT / Interest Expenses
A5	Account Receivable / Liabilities

Table 5.5: Loughran and Mcdonald (2011) wordlists

Wordlist	Sample words
Positive	enthusiastically, assures, improve, empower, complimented
Negative	investigating, complaining, confusion, severe, exculpations
Uncertainty	risking, anticipated, hidden, unobservable, imprecision
Litigious	prosecute, referenda, presumptively, licensable, sequestrator
Modal Strong	highest, strongly, will, unequivocally, lowest

5.3.7 Experimental setup

We filter words with document frequency above 70%, as well as standard stop words from a list. Additionally, we remove low-frequency words that appear in less than 10 documents. To determine the number of topics, first we run 10 epochs of the traditional LDA model with online learning (Hoffman et al., 2010) for all MD&As, the number of topics ranging from 10 to 100, and we compute the perplexity on a hold-out dataset to determine the optimal number of topics. We then run the LDA and DETM using full dataset of 21 years from 1997 to 2017 to examine the topics and their evolution. Finally, we set the predictive scenario as out-of-sample and out-of-time prediction commonly used in credit risk. Specifically, a moving temporal window of 10 years of data is used to train the logistic regression with l_2 regularisation, and the next year as the test data. For example, the data from 1997 to 2006 will be used to train and then predict the bankruptcy in 2007 data. After that, the training and testing window will shift ahead one year and we repeat the whole process. The final performance is reported for each year and averaged for all years as shown below.

For the DETM, we follow Dieng et al. (2019b) to set the following components:

Table 5.6: DETM settings

Component	Setting
Word2Vec (Mikolov et al., 2013b)	Skip-gram, 300 dimensions
Batch size	200
Activation	ReLU
Number of hidden layers	2 (800 nodes for each layer)
Drop-out	0.1
Epochs	100

At the final stage, we employ logistic regression with l_2 regularisation and inverse strength λ ranging from 0.001 to 0.01 (the smaller the λ the stronger the penalty to less influential features). Despite being simple, logistic regression is an industrial standard in bankruptcy prediction and is proved to give comparable performance with other advanced classifiers (Lessmann et al., 2015; Altman et al., 2017), and more importantly it is straightforward to explain its predictors.

5.4 Experimental Results and Discussions

5.4.1 Number of topics

As stated in Section 3, we find the optimal number of topics based on the bound of the perplexity of the second half of the test set conditioned on its first half. Figure 5.1 presents the bounds for different settings of number of topics ranging from 5 to 100, each LDA model (Hoffman et al., 2010) is trained using asymmetric α learned from corpus, 10 epochs, and maximum number of iterations through the corpus when inferring the topic distribution of a corpus is 100.

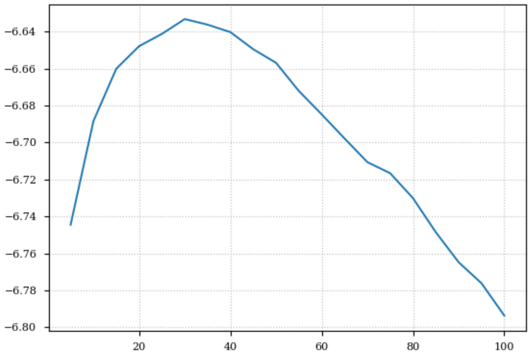


Figure 5.1: Bound of perplexity on the test set

There is a sharp dip in $k = 30$ which shows that the perplexity is deteriorating, and it is encouraging that the possible number of topics could be 30. Based on this number of topics, we estimate the full topic models for topic exploration task and moving-window based topic models for bankruptcy prediction task.

5.4.2 Topics in MD&A and their evolution

MD&A, as discussed in Section 1, is of important source of corporate information which reflects not only its past and current financial strength but emphasises the significant of the new products, services, collaboration projects, strategic partners, potential M&A deals, etc. However, our study, to the best of our knowledge is the first to utilise this large-scale dataset to provide an meaningful representation overtime of hidden topics using unsupervised methods. We present some significant topics inferred through both LDA and DETM models using wordcloud, the size of each word is plotted based on the word probability in a topic as follows:

The wordclouds in Fig. 5.2 shows some related, interesting words to name topics such as partnerships, research&development, energy price, investment loss, sale&store, tax&currency

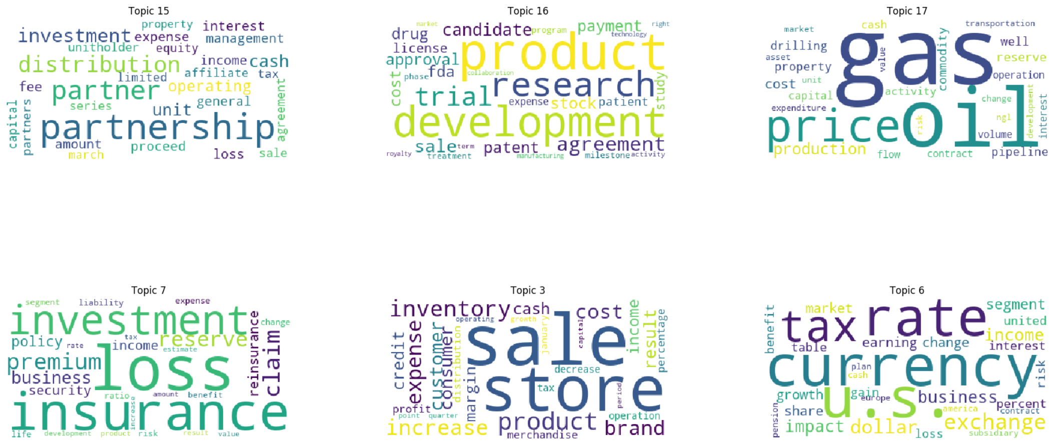


Figure 5.2: Topics discovered by LDA

rate. This shows that our approach using topic modelling is effective to uncover the hidden topics in the management discussion and analysis of the corporate filings. For the full list of topic wordclouds, please refer to Fig. A.1, A.2, and A.2 of Appendix A.

One of the essential follow-up concerns that might attract the policy makers and macroeconomists is how these topics evolve through the market cycles. We provide further explanation to this question by using DETM which can leveraging the word embedding with smooth transition of topics overtime. This is made possible by plotting the tensor $\beta = \text{softmax}(\rho^\top \alpha)$ of words in a topic over all time slice of our data. The following Figure presents the evolution of words in their associated topic. Specifically, in *investment loss* topic, words associated with financial

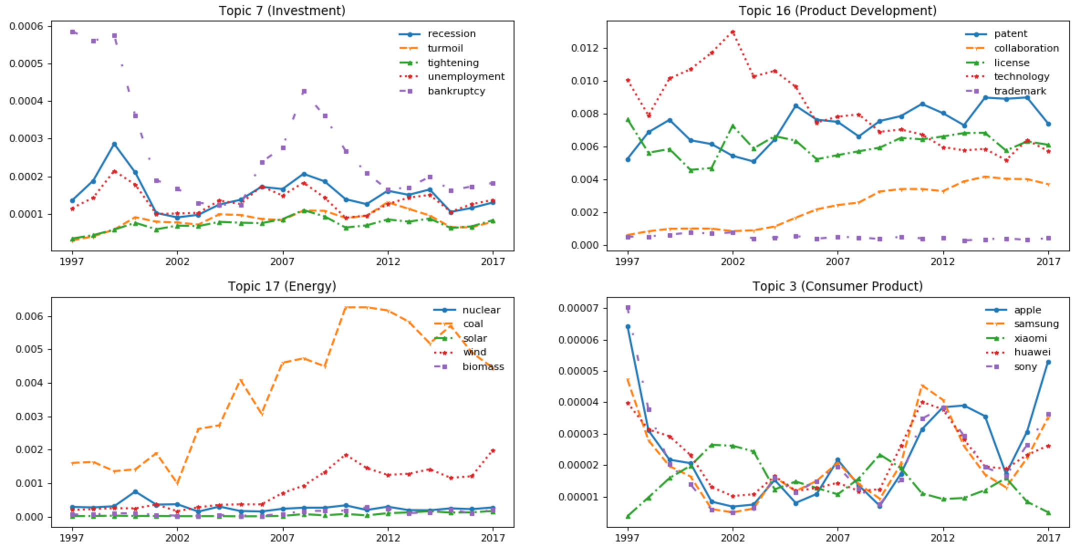


Figure 5.3: Evolution of topics

crisis such as ‘recession’, ‘turmoil’, ‘tightening’, and ‘unemployment’ show strong movement correlation and peak at the crisis period accordingly; the word ‘collaboration’ is progressing in topic *product development* where as ‘patent’ and ‘license’ share their top weights; along with the decreasing of ‘coal’ following the sharp decline in coal production in 2015[6], the weights of renewable sources such as ‘wind’ or ‘solar’ are increasing in *energy* topic; and the *consumer product* topic reveals the competition of several main electronic device producers over the last two decades, partially reflecting their up and down as well as their current positions in the US market.

We then proceed to examine to which extend the topic distributed representation of the MD&A section in 10-K filings could help on bankruptcy prediction task in the section below.

5.4.3 Predictive performance

In this prediction task, we compare the topic representation of MD&A using both LDA and DETM with three baselines: z-score from Altman (Altman et al., 2017), dictionary-based count of sentiment words based on the dictionary of Loughran and McDonald (Loughran and Mcdonald, 2011), and *doc2vec* (Le and Mikolov, 2014). We use three performance metrics which are area under the ROC curve (AUC), area under the precision-recall curve (AUPRC), and Brier score (BS). The first metric is commonly employed in assessing predictive performance, however, in the classification task with extremely imbalanced data, this might not present the general view on classifier performance because of the high prevalence of the majority class, and AUPRC is recommended by several research in credit risk modelling (Davis and Goadrich, 2006; Crone and Finlay, 2012; Garca et al., 2019). Particularly,

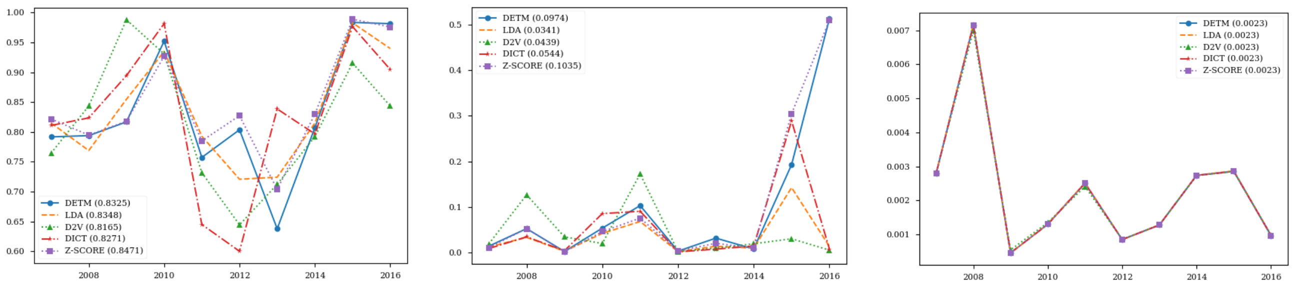
$$\mathrm{AP} = \sum_{n} (R_n - R_{n-1}) P_n$$

where P_n and R_n are the precision and recall at the n^{th} threshold. With naive random predictions, the AP is just the percentage of positive classes. The third metric is a distance between the predicted probability and true label, and is regarded as a cost function, the lower the score is the better the prediction.

We present the performance of the five feature sets from 2007 to 2016 in Fig. 5.4a to Fig. 5.4c. Each point in the graphs is the mean value of 5-fold grid-search for best λ, then the overall performance for the entire period is reported in the parentheses of associated labels. While the Brier score shows no significant difference as expected because of our extremely imbalanced data, the AUC and AUPRC reveal some noticeable results.

First, for both AUC and AUPRC, despite trained on the same text corpus, the topic modelling approaches present superior performance compared with dict-based count vectorisation

[6] “U.S. coal production dropped by more than 10% in 2015 to 897 million short tons, the lowest production level since 1986”, US Energy Information Administration, https://www.eia.gov/todayinenergy/detail.php?id=28732, Retrieved 3rd July, 2020



(a) AUC

(b) AUPRC

(c) BS

Figure 5.4: Performance of classifier

or distributed representation using skip-gram, one possible explanation for this is these two approaches do not have the topical structure for the input text data, hence the lower predictive performance. Second, in terms of AUC, the topic representations (LDA and DETM) provide comparable performance with z-score although they do not use any numerical, demographic, or financial accounting data (83% compared with 84%). In AUPRC, the DETM model produces the features as good as z-score in predicting bankruptcy (0.097 and 0.103).

The seemingly reverse performance trend of LDA topical feature set and z-score suggests that their combination could benefit the classifier. We then examine the improvement in utilising the topical features along with traditional accounting features by a simple concatenation of both feature sets. The predictive power of combined feature set is compared with z-score features using the same performance measures in the following Table 5.7 and 5.8:

Table 5.7: Comparison of z-score and combined feature set using LDA

	z-score				LDA + z-score		
	AUC	AUPRC	BS		AUC	AUPRC	BS
2007	0.8217	0.0110	0.0028	2007	0.8231	0.0250	0.1501
2008	0.7946	0.0525	0.0072	2008	0.8744	0.0936	0.1644
2009	0.8179	0.0023	0.0005	2009	0.8412	0.0027	0.1666
2010	0.9268	0.0459	0.0013	2010	0.9710	0.1298	0.1363
2011	0.7850	0.0756	0.0025	2011	0.8165	0.0947	0.1315
2012	0.8271	0.0036	0.0009	2012	0.9242	0.0145	0.1544
2013	0.7039	0.0203	0.0013	2013	0.8075	0.0061	0.1510
2014	0.8299	0.0098	0.0027	2014	0.8609	0.0283	0.1544
2015	0.9887	0.3044	0.0029	2015	0.9599	0.1429	0.1630
2016	0.9756	0.5098	0.0010	2016	0.9717	0.1088	0.1464
Mean	0.8471	0.1035	0.0023	Mean	0.8850	0.0646	0.1518

Despite a small average increment in AUC of 1.55% for using DETM, we observe better enhancement in AUC of 3.89% by using LDA to build MD&A representation, however, there are declines in BS for LDA and in AUPRC for both combinations. The proposed combinations

Table 5.8: Comparison of z-score and combined feature set using DETM

	z-score				DETM + z-score		
	AUC	AUPRC	BS		AUC	AUPRC	BS
2007	0.8217	0.0110	0.0028	2007	0.8765	0.0274	0.0028
2008	0.7946	0.0525	0.0072	2008	0.8651	0.0785	0.0071
2009	0.8179	0.0023	0.0005	2009	0.7241	0.0016	0.0005
2010	0.9268	0.0459	0.0013	2010	0.9715	0.1485	0.0013
2011	0.7850	0.0756	0.0025	2011	0.7835	0.0650	0.0025
2012	0.8271	0.0036	0.0009	2012	0.9483	0.0459	0.0008
2013	0.7039	0.0203	0.0013	2013	0.7033	0.0047	0.0013
2014	0.8299	0.0098	0.0027	2014	0.8322	0.0102	0.0027
2015	0.9887	0.3044	0.0029	2015	0.9586	0.2456	0.0028
2016	0.9756	0.5098	0.0010	2016	0.9730	0.0382	0.0010
Mean	0.8471	0.1035	0.0023	Mean	0.8636	0.0666	0.0023

are actually worse than the traditional z-score in 2015 and 2016 performance wise when z-score achieved almost perfect AUC of 99% and 98%, respectively.

One of limitations here is that, despite our corpus is all 10-K filings from EDGAR system, the coverage of our corpus is just merely 20 years, which apparently does not include many important economics cycles hence hardly makes the evolution of topics distinguishable when we query the similarity of MD&As in our corpus. We believe sub-sequence research in this avenue will bring more interesting results as more data gathered.

5.5 Conclusions

Understanding management discussion and analysis will help the investors and policy markers to response better to corporate business changes and future prospects. And by utilising the topic modelling approaches on more than 198k filings with 38M words, we found 30 topics embedded in the MD&A which reflect important business aspects such as ‘energy’, ‘partnership’, ‘research and development’, ‘loan and interest rate’, and so forth. In addition, the evolution of words in topics are inline with crucial economics events such as the financial crisis, the big reduction in coal production in 2015, or the competition of electronic devices producers. Crucially, when the problem of window dressing to make the credit quality better become popular not only in financial sector but also in other sectors (Guan et al., 2008; Gandhi et al., 2019), there are increasing needs for employing alternative data to complement with traditional scoring methods, especially the data that reliably represent the forward-looking or future prospect of the businesses. We made this possible in out-of-time and out-of-sample prediction scenario by first examining the predictivity of textual feature built from MD&A, and we showed that they provide comparable performance with industrial standard using z-score. Second, by simply concatenating the textual features and z-score features, we could improve the bankruptcy prediction methods.

There are potential extensions based on our current limitations, first is to employ more textual data to reflect better the economic cycles, then, to employ balancing treatments such as oversampling and undersampling to tackle the severe imbalance classes. Other research could devote to investigate better combination of textual features with traditional features or to examine different training coverage to optimise the prediction performance.

Chapter 6

Sentiment Analysis of Textual Data

As of July 3^{rd} 2020, an extended version of this chapter is accepted for publication at Journal of Operational Research Society.

6.1 Introduction

This study contributes to the current literature of understanding tone in the financial filings. This chapter presents a textual analysis for the default prediction by proposing new features from word-level textual data of the filings, and then demonstrates the competitive predictive power of those textual features. Specifically, we search for sentiment words in a sentence of MD&A, radiate to sentence level based on a dictionary-based financial sentiment classifier and averaging the overall sentiment of an MD&A. This study then compare the predictive performance of traditional models built on accounting data with the ones trained on textual data. The results not only present the significance and the comparable predictive power of textual features but also demonstrate that the combined model built on both types of data could significantly improve both model fitness and model predictivity power. The experiments are carried under three data segments, with three treatments for the imbalanced dataset (IDS), and with six predictive power metrics measured using both stratified k-fold cross validation and one-year-ahead prediction. In what follows, we first present the relevant literature on the default classification models and the textual analysis in credit risk modelling in section 6.2. Section 6.3 devotes to the process of mining MD&As, forming textual features, and comparing frameworks. The data and experiment results are presented in section 6.4 and 6.5, and we conclude this work in section 6.6.

6.2 Related Works

Regarding the employed features for the classifiers, Altman (1968) z-scores are well-known features in building corporate default prediction models. Further extension are also made for the small and medium enterprises (SMEs) such as in Andreeva et al. (2004); Altman and Sabato (2007); Altman et al. (2010) by using other financial and macroeconomics features, or Andreeva et al. (2016) by using the generalised extreme value models. Accordingly, different financial institutes might use different feature set for their credit risk models, depend on their risk apatite. At the current stage, the traditional models are pretty mature and most of them are built on numerical and historical data, and there are in need of works on forming new features, especially those based on text data with their sentiments, to assess credit risk. Hence, it is desirable to have a comparison for the predictive power of other alternative features built from textual analysis with the traditional features.

At the same time, there are several works on quantifying the textual data of the filings into meaningful predictors for predicting future company financial performance (Healy et al., 1999; Lawrence, 2013), or stock returns (Kraus and Feuerriegel, 2017; Zhou, 2018). In the pioneering work of Loughran and Mcdonald (2011), they built a dictionary with six wordlists that offers more accurate tone for financial text compare with the traditional Harvard Dictionary. Based on this dictionary, Engelberg et al. (2012) showed that public news provides valuable trading chances for competent information processing short sellers; Bonsall et al. (2017) created a new financial reporting readability measure and showed that firms with improved readability of their filings have better ratings, lower bond rating disagreement, and lower cost of debt.

More recently, Zhou (2018); Jiang et al. (2019) analyzed and examined the relationship between homogeneous and heterogeneous sources of financial texts and indicated that higher manager sentiment followed by lower earnings disruption and higher investment growth. To our knowledge, the are two closest works in predicting the default companies utilizing the textual data, which are Gandhi et al. (2018) and Mai et al. (2019). The first one used sentiment words to examine the financial distress of the US banks, their findings suggested that more negative words in the reports are related to a higher probability of distressed delisting subsequently, and the latter employed deep learning model, specifically the convolutional neural network model, to examine the gain in predictive power using the word embedding techniques. However, the relationships between those sentiment words and the likelihood of default have not been investigated for non-banking sector data, and further augmented with the classification performance for out-of-time and out-of-sample data. In addition, since deep learning models, in general, is a black-box model which is very difficult to interpret its feature-constructing process as well as its classification output, little works have been devoted to the performance of classification models built on the simple, yet intuitive textual features, and relatively compare them with the traditional ones.

By and large, this study uses textual data in the 10-K filings of US listed firms collected from EDGAR from 1997 to 2017 to examine how could textual data help to predict the default-/liquidation firms. We separate our analysis to different segments of the corporate data to investigate the relationships of the textual features with the corporate probability of default and we further examine their predictivity powers on both cross validation and one-year-ahead prediction.

6.3 Experimental Settings

As logistic regression is the most well-known model for credit scoring modelling thanks to its simplicity and interpretability (Lin et al., 2012), we employ the logistic regression for estimation and prediction of the baseline, text, and combined model. And as z-score (Altman, 1968) is the most commonly used model for bankruptcy prediction on both normal corporate and small businesses (Altman and Sabato, 2007; Altman et al., 2010), we use the Altman factors for constructing the baseline models. We then detail, for the text and combined model, how we form the textual features with counting mechanism and dictionary-based classifier using a financial sentiment dictionary. And we also present six performance metrics we employ for comparing the baseline models with the combined models. All of the experiment results are presented for the small and medium enterprises (SME), non-SME, and all samples. We regard a corporate to the SME category if its sale less than or equal to $65 million (Altman and Sabato, 2007) which is in line with the Basel Capital Accord[1].

6.3.1 Financial ratios

I follow Altman (1968) and Altman et al. (2010) to calculate the z-score factors and the other five factors designed specifically for the small businesses, respectively. Detail formulas of these factors could be found on Table 5.4 of Chapter 5.

6.3.2 Financial wordlist

Loughran and Mcdonald (2011) showed that some words which are negative in Harvard IV dictionary[2] are actually neutral or even positive in the financial context, such as cancer, depreciation, liability, and so forth. To mitigate the proxy for industry or other unintended effects in using the general sentiment dictionary, they proposed six wordlists (LM): positive; negative; uncertainty; litigious; strong modal; and weak modal words[3]. It should be noted that at the time we assess these wordlists, the uncertainty wordlist includes all words in the weak modal wordlist. The randomly 5 words for each wordlist are as follows: (similar with Table 5.5 of Chapter 5)

Using this wordlist, Loughran and Mcdonald (2011) found that firms using more positive, negative, or modal strong words (i.e. stronger language) are more likely to reveal a material

[1] Basel Committee on Banking Supervision, June 2004.

[2] Available at http://www.wjh.harvard.edu/ inquirer/homecat.htm, accessed September 25th, 2019

[3] Available at http://www.nd.edu/mcdonald/Word_Lists.html, accessed September 25th, 2019

Table 6.1: The Loughran and Mcdonald (2011) wordlists

Wordlist	Sample words
Positive	enthusiastically, assures, improve, empower, complimented
Negative	investigating, complaining, confusion, severe, exculpations
Uncertainty	risking, anticipated, hidden, unobservable, imprecision
Litigious	prosecute, referenda, presumptively, licensable, sequestrator
Modal strong	highest, strongly, will, unequivocally, lowest
Modal weak	occasionally, perhaps, somewhat, suggest, nearly

weakness in their internal controls. These wordlists are also showed to have relations with the subsequent market reactions such as trading volume, unexpected earnings, and stock return volatility (Elshandidy et al., 2018). Motivated on these wordlists, we present in section 6.3.4 how we form the textual features using counting sentiment words and counting sentiment sentences with dictionary-based classifier.

6.3.3 Dictionary-based sentiment classifier

The main role of a sentiment classifier is to determine the polarity of a document as positive, negative, or neutral. And the document in this research is a sentence of the MDA. More fine-grained categories could be use such as strongly positive, weak negative, and more sophisticated sentiment classifiers could be employed such as deep learning models including recurrent neural networks, long short-term bi-directional neural networks, or convolutional neural network (Peji Bach et al., 2019). However, they require a large amount of annotated data to work effectively, which is costly to obtain especially in the financial domain. Hence, having the sentiment wordlists for the financial domain, we resource to dictionary-based sentiment classifier to examine how the textual features built on it could help on predicting corporate default.

Dictionary-based (or rule-based) sentiment classifier uses several rules to calculate the sentiment score using sentiment words from a lexicon or wordlist. Each word in a sentence will be assigned a sentiment score based on its corresponding sentiment wordlist. When a negation word or constrastive conjunction word appears preceding that word within a pre-defined distance, e.g three words, its polarity is flipped or switched. Then, using majority voting, a sentiment class is determined for each sentence. Specifically, positive words are assigned +1 sentiment scores, while negative and uncertainty words are assigned -1 scores, the sentence sentiment will be calculated based on the aggregate score adjusted for a set of rules. In this study, we employ three simple rules inspired by Hutto and Gilbert (2014), based on Loughran and Mcdonald (2011) financial wordlists, as follows:

1. assess the negative and positive lexicons in a sentence, and determine the sentence polarity by majority voting,
2. consider the shifting of sentiment with contrastive conjunctions such as ‘however’, ‘but’, ‘despite’, ‘neither’ and so forth, and
3. examine the tri-gram preceding the lexical feature for modifying the valence of a word.

Formally, lets $dict\_lex = \{key : value\}$ be a dictionary of lexical features with key is a negative or positive word in the LM wordlist and $value$ is -1 or 1 for negative or positive words, respectively; $dict\_da = \{key : value\}$ denotes a dictionary of degree adverbs where key is a uncertainty or strong-modal word in the LM wordlist and $value$ is -0.5 or 0.5 for uncertainty or strong-modal words, respectively; $list\_cc$ denotes a list of contrastive conjunction words. The dictionary-based financial sentiment analysis algorithm 3 is as follows:

Algorithm 3 Dictionary-based financial sentiment analysis

```
Input: $s = \{w_1, w_2, .., w_m\}$, a financial sentence of $m$ words.
Parameter: $n$, n-gram preceding a lexical feature.
Output: $score$, a weighted composite sentiment score.
initiate sentiment list $sentiment \leftarrow [\,]$
for each word $w_m \in s$ do
    initiate valence of $w_m, val(w_m) \leftarrow 0$
    if $w_m \in dict\_lex$ then
        $val(w_m) \leftarrow dict\_lex[w_m]$
    end if
    for $i \in [1, n]$ do
        if $w_{m-i} \in dict\_da$ then
            modify $val(w_m)$
        end if
    end for
    add $val(w_m)$ to $sentiment$ list.
end for
if a word $w_m \in list\_cc$ then
    update $sentiment$ list
end if
normalise the total score as composite $score$ from $sentiment$ list.
return $score$
```

We validate the performance of this dictionary-based financial sentiment classifier with threshold set at 0.05 by the gold standard of 2264 annotated financial phrasebanks[4] from Malo et al. (2014). Setting a threshold at 0.05 means that a sentence that has $score \geq 0.05$ is classified as positive, $score \leq -0.05$ is classified as negative, and $-0.05 < score < 0.05$ is classified as neutral. Table 6.2 presents the performance summary:

Table 6.2: Performance of dictionary-based financial classifier

Accuracy	0.7380			
	Negative Sent.	Positive Sent.	Neutral Sent.	Total
precision	0.5779	0.7913	0.7289	
recall	0.8161	0.7952	0.5546	
f1-score	0.6759	0.7930	0.6284	
support	303	570	1391	2264

Performances are averaged from 10 random splits of the financial phrasebanks.

[4] English news on all listed companies in OMX Helsinki

The performance is comparable with linearized phrase-structure sentiment classifier proposed by Malo et al. (2014) and better than other non-financial dictionary based classifiers, and support vector machine classifier presented in that same research. Despite showing high recall for negative sentences, our classifier is somewhat conservative because it presents a low precision of 58% which translates to almost 42 sentences predicted negative are not truly negative out of 100 negative-predicted sentences from our classifier. Table 6.3 reveals some errors among wrongly predicted classes for 18% true negative sentences:

Table 6.3: Error in negative sentence prediction

ID	Negative Sentence	Predict
29	Profitability (EBIT%) was 13.6 % , compared to 14.3% in Q2 2009.	1
31	Shares in Royal and Sun Alliance continued to slide back from a 12-month high of 172p last month...	1
38	Return on investment ROI was 4.1% compared to 43.8% in the first half of 2008	0
45	Diluted earnings per share (EPS) stood at EUR 0.25 versus EUR 0.42	0
69	Marimekko Oyj posted a net profit of 7.99m EUR for 2006, compared to 8.4m EUR for 2005	1
75	Operating profit was EUR 1.6 mn in 2005 compared to EUR 5.9 mn in 2004.	0
141	Uponor OYJ cut its full-year sales growth forecast to 6 pct from 10 pct , blaming tough conditions...	1
277	In the building and home improvement trade, sales decreased by 22.5% to EUR 201.4 mn	1

The majority of errors originate from (i) the inability to detect changes in numbers and time expressions in the sentence as in sentences 29, 38, 69, and 75, (ii) unable to detect context as in sentences 45 and 141, and (iii) mismatched lexicon items as in sentences 31 and 277. This suggests further developments to fix the shortcomings of the current models, a possible approach might focus on the contextual analysis and is beyond the scope of this current study.

6.3.4 Textual features

We then form the following textual features:

1. **In the entire filing**: Percentage of negative, positive, uncertainty, litigious, modal strong, and modal weak words, namely: *pneg*, *ppos*, *punc*, *plit*, *pmods*, and *pmodw*. An sentiment feature (*senti*) is also computed by taking the difference between *ppos* and *pneg*.

2. **In the MDA section**: Count of negative, positive, uncertainty, litigious, modal strong, and modal weak words, namely *NEG*, *POS*, *UNC*, *LIT*, *M1*, and *M3*.

3. **In the MDA section**: Percentage of negative, positive, and neutral sentences namely *PCNEG*, *PCPOS*, and *PCNEU* using our dictionary-based classifier defined above.

These textual features will be used as row labels as illustrated in following Table 6.4 that could be automatically generated by our textual analysis framework:

Table 6.4: Textual features

CIK	Filing ID	pneg	ppos	...	PCNEG	PCPOS	...
151629	6	0.011	0.025	...	16	23	...
708788	17	0.002	0.067	...	54	71	...

CIK - Central Index Key is the corporate ID given by SEC.

6.3.5 Predictive measurement

Regarding the model performance for predicting default/liquidation firms, along with the well know area under the ROC curve (AUC), it is advisable to have measurements that capture different aspects of the classifiers, especially with the present of IDS. Together with recall and precision, Davis and Goadrich (2006) proposed to use Area Under Precision-recall Curve (AUPRC) when we perform classification with an IDS and want to concentrate on the positive examples, the default corporates. Specifically, sensitivity is directly influenced by class imbalance, whereas True Positive Rate only depends on positives. In addition, some practitioners might pay more attention to recall value as in the total number of predicted bad corporate loan applications, how many of them are actually bad as bad loans could easily wipe out all the profit of entire loan portfolio. We also include the Brier score to assess the accuracy of the scorecards probability predictions.

6.4 Data

6.4.1 Financial data

In this study, we also collect the accounting data from the Wharton Research Data Services (WRDS) for all listed firms in the US from 1997 to 2017, similar as in Chapter 5 but with updated company info up to 2017 financial year. Based on these elements, we impute for the missing observations using multiple imputation using chain equations - MICE (Buuren and Groothuis-Oudshoorn, 2011) and calculate the 5-factor z-score (Z1 to Z5) as Altman (1968) and 5-factor as Altman et al. (2010) (A1 to A5) to proxy for the financial accounting features for the non-SME and SME samples, respectively. We also exclude firms in the financial and regulated utilities sectors with SIC from 6000 to 6999 and from 4900 to 4949, respectively. As for the default flags, a firm is marked as default if it filed for liquidation under Chapter 7 or Chapter 11 bankruptcy filings[5] in the 8-K form.

6.4.2 10-K filing

This study also use the same textual data presented in Table 5.1 of Chapter 5.

[5] https://www.sec.gov/reportspubs/investor-publications/investorpubsbankrupthtm.html

The number of defaults under each SIC category is as follows:

Table 6.5: SIC categories

Range	Division	Category	Count	Number of defaults
0100-0999	Agriculture, Forestry and Fishing	SIC_0	404	2
1000-1499	Mining	SIC_1	5032	23
1500-1799	Construction	SIC_2	1091	3
2000-3999	Manufacturing	SIC_3	39627	107
4000-4999	Transportation, Communications	SIC_4	4798	41

At the final stage, we merge the financial data and SEC filings by matching the CIK and the fiscal year-end of financial reports. The final data consist of 51,128 firm-year observations, of which 176 firms are default (approximately 0.34%). Using the SME definition in Section 3 above, we have the number of defaults under each business segment as follows:

Table 6.6: Number of defaults in two data segments

Default	Business Type	Count	Percentage(%)
False	non-SME	28980	99.612
	SME	21972	99.714
True	non-SME	113	0.388
	SME	63	0.286

And the number of defaults for each year is illustrated in Figure 6.1 where we plot three trend lines for the SME, non-SME, and all samples. By matching with the SEC filings data, we do not have any default observations in 2010 for the entire sample, in 2010 and 2013 for the non-SME samples, and in 2010, 2011, and 2016 for the SME sample.

6.4.3 Missing data and correlation

As for the missing data (c.f Appendix), for the financial elements from financial statements, *ipodate* has the highest percentage of missing values of more than 50%. The total market value *mkvalt* has 11% missing values and the remaining elements have less than 10% missing data. For Altman's factors, we find A4 for SMEs has approximately 16.5% missing data while for the remaining 9 factors, missing data consist of around 1%. These observations show that the missing data problem is not severe in our case, we proceed with the multiple imputation using chain equations (Buuren and Groothuis-Oudshoorn, 2011) to impute for the missing values.

In terms of the correlation of both Altman's and textual factors, we present the pair-wise correlation plots in Appendix B.1 while the Altman's factors showing low correlation, we notice the significant correlation coefficients between the three groups of textual features which built on the MDA section. In the seven features belong to the counting of sentiment words for the entire filing, the percentage of negative words (*PCNEG*) highly and negatively correlates with *senti*. All the counting for sentiment words and sentiment sentences are highly correlated in the

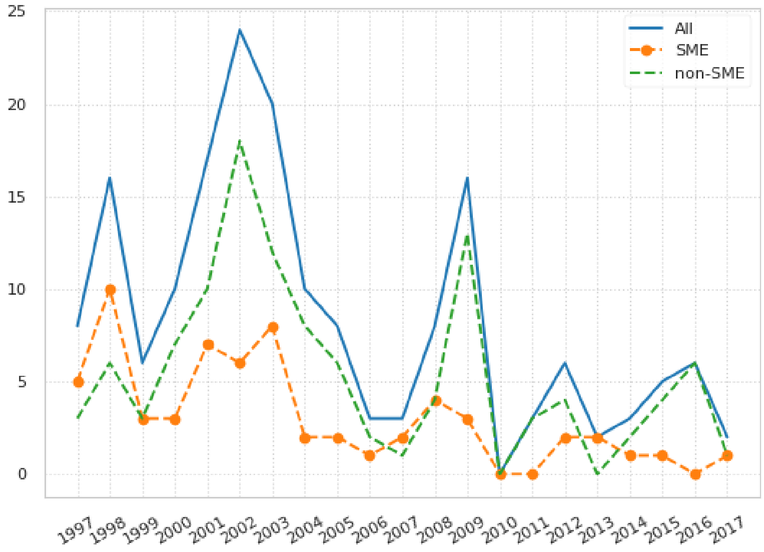


Figure 6.1: Number of defaults for each segment

remaining three groups of features dedicated to the MDA section only. One possible explanation for this is, since they are all intrinsic variables which are basically different manifestations of the same underlying, immeasurable latent variable of the filing sentiment, we would expect a high correlation between them. To account for multicollinearity, we conservatively choose the following textual features to enter to our final models: features relating to the percentage of LM wordlists in the entire filings, consist of *ppos*, *plit*, *pmods*, *pmodw*, *punc*, *senti*, we exclude *pneg* and *pmodw* because they strongly correlate with *senti* and *punc*; difference of percentage of positive and negative words in MDA relative to the total number of words of the filing (*SENTI_MDA*); percentage of negative and percentage of positive sentences in MDA (*PCNEG* and *PCPOS*). We notate the textual features measured based on the entire filing using normal characters while for the ones measured based on the MDA section only we use the capital letters for ease of differentiation.

6.5 Empirical Analysis

In this section, we present the comparison of the regularised logistic regression models fitted for (i) the financial features only, (ii) the textual features only, and (iii) the combine features in Table 6.7 and Table 6.8. The performance measured in terms of the predictive power are presented in Table 6.9 where we show the performance of the traditional financial factors and the combined models using six performance metrics under both 10-fold stratified cross validation and one-year-ahead prediction.

6.5.1 Regression

The regression results in Table 6.7 and 6.8 below show the regression results for SME vs non-SME samples and SME vs all samples, respectively. Each table presents three models fitted using the logistic regression with regularisation, we control for the number of employees (EMP), size measured in terms of logarithm of the squared total assets (LNATSQ), return on assets (ROA), return on equity (ROE), and industry (SIC), and taking 1-year lag in the regression models. The lasso hyperparameter of regularisation strength is defined by cross-validation on finding the highest R^2 and the lasso regularisation path plots for SME, non-SME, and all samples are presented in Figure B.3, B.4, and B.5 of Appendix B.2, respectively. **Altman** is the model fitted using the financial features computed as Altman (1968) for the non-SME samples and as Altman et al. (2010) for the SME samples; **Text** is the model fitted using the textual features which are described in subsection 6.3.4; and the final model, **Combined**, is fitted using both Altman's factors and textual features.

Regarding the significance of financial features, our results are in line with the current literature of corporate default/liquidation modelling (Altman et al., 2010). Besides, the textual features are significant and modestly improve the model fitness. As for the SME sample, the percentage of uncertainty words and **senti** are both significant and negative, the higher the percentage of uncertainty words, and the higher the difference between the number of positive and negative words, the less likely the firm being defaulted. Focusing on the MDA section only, we find that the more positive sentences in a MDA a firm have, the lower the probability of it being defaulted. The signs and significance are consistent for both textual and combined models. For the non-SME sample, the higher the percentage of negative sentences relative to the total number of sentences in the MDA, the lower the default probability. As **senti** is the difference between the percentage of positive and negative words, this shows that, relative to the total number of words in the filing, a firm having more positive words than negative words has a lower probability of default.

In terms of the combined model, for the entire samples, all the financial and textual features remain significant with the same signs. Interestingly, the percentage of the number of negative sentences (PCNEG) relative to the total sentences in the MDA is significant and positive, which suggests that the higher the negative sentences in the MDA, the higher the likelihood of the firm being defaulted. And the higher the number of positive sentences in the MDA, the less likely the firm being defaulted. Our findings for **senti** provide significant evidence to Gandhi et al. (2018), where the authors used the sentiment measures from banks annual reports to show that the percentage of negative words in the annual report has a positive relationship with the likelihood of distress delisting subsequently for the US banks. We further find that the percentage of uncertainty words, the percentage of positive sentences in the MDA are negatively related to the likelihood of the firm being defaulted which might suggest that the more awareness about the uncertainty factors a firm has, the less likely the firm being default.

All in all, we demonstrate that the textual data on the financial reports help not only forming the new and significant textual features on explaining default for the SME and non-SME segment

Table 6.7: Regression models

	SME			non-SME		
	Altman	Text	Combined	Altman	Text	Combined
Intercept	-3.703***	-3.122**	-2.411	-3.133**	-3.465***	-2.285
	(1.034)	(1.359)	(1.467)	(1.527)	(1.334)	(1.403)
A1	1.410*		1.280*			
	(0.729)		(0.754)			
A2	-0.927		-0.894			
	(0.616)		(0.605)			
A3	-1.187**		-0.695			
	(0.477)		(0.486)			
A4	-0.558		-0.605			
	(0.505)		(0.524)			
A5	-0.122		-0.175			
	(0.581)		(0.597)			
Z1				-3.810***		-3.407***
				(0.564)		(0.572)
Z2				-0.183		-0.003
				(0.561)		(0.563)
Z3				-1.353**		-1.226*
				(0.644)		(0.642)
Z4				-0.345		-0.567
				(0.362)		(0.369)
Z5				0.664		0.537
				(0.426)		(0.435)
PCNEG		0.658	0.623		0.935**	0.827*
		(0.552)	(0.554)		(0.435)	(0.440)
PCPOS		-1.209**	-1.158**		-0.408	-0.283
		(0.532)	(0.535)		(0.423)	(0.431)
plit		-0.637	-0.471		-0.024	0.179
		(0.603)	(0.610)		(0.544)	(0.554)
pmods		0.315	0.341		0.556	0.473
		(0.502)	(0.506)		(0.375)	(0.380)
ppos		-0.307	-0.308		-0.052	0.192
		(0.643)	(0.644)		(0.439)	(0.448)
punc		-2.412***	-2.271***		-2.502***	-2.059***
		(0.634)	(0.640)		(0.569)	(0.580)
senti		-2.152***	-2.006***		-0.584	-0.612
		(0.663)	(0.662)		(0.495)	(0.506)
SENTI_MDA		-0.408	-0.418		0.048	0.111
		(0.589)	(0.585)		(0.488)	(0.492)
EMP	-0.437	-0.605	-0.559	0.775	0.571	0.263
	(0.575)	(0.560)	(0.575)	(0.608)	(0.564)	(0.643)
LNATSQ	1.381**	1.818***	1.860***	1.785***	2.078***	2.187***
	(0.611)	(0.676)	(0.674)	(0.669)	(0.585)	(0.705)
ROA	-2.080**	-1.233**	-1.593*	-8.056***	-8.713***	-7.351***
	(0.845)	(0.596)	(0.845)	(1.024)	(0.845)	(1.053)
ROE	-0.484	0.376	-0.510	0.770***	1.220***	0.686***
	(0.590)	(0.434)	(0.589)	(0.264)	(0.248)	(0.266)
SIC_1	-0.998	-0.661	-0.628	-0.705	-0.697	-1.252
	(0.877)	(1.040)	(1.054)	(1.518)	(1.102)	(1.140)
SIC_2	0.195	0.281	0.448	-2.217	-2.376*	-2.922*
	(1.091)	(1.226)	(1.223)	(1.863)	(1.436)	(1.497)
SIC_3	-0.976	-0.442	-0.310	-0.461	-1.246	-1.265
	(0.823)	(1.002)	(1.020)	(1.480)	(1.056)	(1.087)
SIC_4	0.148	0.519	0.546	-0.670	-1.002	-1.532
	(0.861)	(1.030)	(1.044)	(1.492)	(1.072)	(1.103)
N	22035	22035	22035	29093	29093	29093
AIC	857.4562	824.7012	826.5098	1092.4631	1118.9101	1073.2403
BIC	969.4616	960.7078	1002.5183	1208.3587	1259.6404	1255.3619
LLR	34.3774	73.1323	81.3238	415.5929	395.1459	450.8157
R-squared	0.0398	0.0847	0.0941	0.2808	0.2670	0.3046

N is the number of observations and LLR is the log-rank test statistics. ***, **, and * indicate 99%, 95% and 90% significant levels, respectively. EMP is the number of employees (thousand). LNATSQ is the natural logarithm of total assets squared. ROA and ROE are the return on assets and return on equity. SIC_i is the i^{th} SIC category with SIC_0 as the reference level.

Table 6.8: Regression models (cont.)

	SME			**All sample**		
	Altman	Text	Combined	Altman	Text	Combined
Intercept	-3.703***	-3.122**	-2.411	-3.004***	-3.577***	-2.125**
	(1.034)	(1.359)	(1.467)	(0.827)	(0.945)	(1.030)
A1	1.410*		1.280*	0.907		1.123
	(0.729)		(0.754)	(0.707)		(0.714)
A2	-0.927		-0.894	0.131		0.182
	(0.616)		(0.605)	(0.344)		(0.342)
A3	-1.187**		-0.695	-1.680***		-1.531***
	(0.477)		(0.486)	(0.342)		(0.348)
A4	-0.558		-0.605	-2.451***		-2.429***
	(0.505)		(0.524)	(0.661)		(0.681)
A5	-0.122		-0.175	-1.080**		-1.247**
	(0.581)		(0.597)	(0.486)		(0.504)
Z1				-1.003**		-0.486
				(0.439)		(0.438)
Z2				-0.261		0.025
				(0.535)		(0.540)
Z3				-0.932		-0.928
				(0.722)		(0.718)
Z4				-0.650**		-0.776***
				(0.293)		(0.300)
Z5				1.148***		0.880***
				(0.329)		(0.339)
PCNEG		0.658	0.623		1.061***	0.920***
		(0.552)	(0.554)		(0.332)	(0.336)
PCPOS		-1.209**	-1.158**		-0.752**	-0.765**
		(0.532)	(0.535)		(0.329)	(0.331)
plit		-0.637	-0.471		-0.263	-0.120
		(0.603)	(0.610)		(0.393)	(0.400)
pmods		0.315	0.341		0.452	0.525*
		(0.502)	(0.506)		(0.314)	(0.318)
ppos		-0.307	-0.308		-0.287	0.071
		(0.643)	(0.644)		(0.362)	(0.368)
punc		-2.412***	-2.271***		-2.689***	-2.440***
		(0.634)	(0.640)		(0.411)	(0.418)
senti		-2.152***	-2.006***		-1.238***	-1.327***
		(0.663)	(0.662)		(0.396)	(0.401)
SENTI_MDA		-0.408	-0.418		-0.186	-0.226
		(0.589)	(0.585)		(0.372)	(0.373)
EMP	-0.437	-0.605	-0.559	0.081	-0.181	-0.302
	(0.575)	(0.560)	(0.575)	(0.492)	(0.499)	(0.520)
LNATSQ	1.381**	1.818***	1.860***	3.835***	3.983***	3.914***
	(0.611)	(0.676)	(0.674)	(0.555)	(0.545)	(0.599)
ROA	-2.080**	-1.233**	-1.593*	-5.101***	-5.562***	-4.461***
	(0.845)	(0.596)	(0.845)	(0.987)	(0.523)	(0.979)
ROE	-0.484	0.376	-0.510	0.296	0.682***	0.326
	(0.590)	(0.434)	(0.589)	(0.340)	(0.226)	(0.342)
SIC_1	-0.998	-0.661	-0.628	-0.982	-0.735	-0.818
	(0.877)	(1.040)	(1.054)	(0.742)	(0.767)	(0.777)
SIC_2	0.195	0.281	0.448	-0.963	-1.412	-1.216
	(1.091)	(1.226)	(1.223)	(0.923)	(0.946)	(0.946)
SIC_3	-0.976	-0.442	-0.310	-0.752	-0.923	-0.720
	(0.823)	(1.002)	(1.020)	(0.708)	(0.743)	(0.748)
SIC_4	0.148	0.519	0.546	-0.424	-0.448	-0.488
	(0.861)	(1.030)	(1.044)	(0.725)	(0.756)	(0.764)
N	22035	22035	22035	51128	51128	51128
AIC	857.4562	824.7012	826.5098	2008.0285	1990.2131	1930.4706
BIC	969.4616	960.7078	1002.5183	2176.0281	2140.5285	2169.2069
LLR	34.3774	73.1323	81.3238	377.7695	391.5849	471.3273
R-squared	0.0398	0.0847	0.0941	0.1609	0.1668	0.2008

N is the number of observations and LLR is the log-rank test statistics. ***, **, and * indicate 99%, 95% and 90% significant levels, respectively. EMP is the number of employees (thousand). LNATSQ is the natural logarithm of total assets squared. ROA and ROE are return on assets and return on equity. SIC_i is the i^{th} SIC category with SIC_0 as the reference level.

but also complement effectively with the traditional financial-based features. In what follows, we examine the predictive powers of the combined model compared with the traditional ones.

6.5.2 Predictive power comparison

The performance of the combined models trained on the textual features is compared with the models trained on the Altman's factors under no sampling, undersampling and oversampling strategies. We report the 10-fold stratified cross-validation using gridsearch for the regularisation parameter of the logistic regression classifier and separate the comparison for three segments: SME, non-SME, and all samples. In addition, trained on a IDS, classifiers tend to perform extremely poor on minority class despite producing high accuracy measures (Chen et al., 2016). Hence, to reveal different aspects of a credit scoring model, it is desirable to employ undersampling and oversampling at the initial stage of modelling process. In order to complement with the cross-validation comparison in terms of (i) the practical modelling scenario and (ii) compliance with Basel III practice for the default model validation purpose[6], and (iii) replicate the nature of scorecard modelling, we also present the one-year-ahead prediction performance using a rolling window of five year data.

Out-of-sample cross-validation prediction performance

First, the performance of three models using six metrics described in subsection 3.5 are illustrated in Figure 6.2, 6.3, and 6.4 as below:

[6] The IRB Use Test: Background and Implementation, Basel Committee on Banking Supervision, https://www.bis.org/publ/bcbs_nl9.pdf, accessed September 25th, 2019

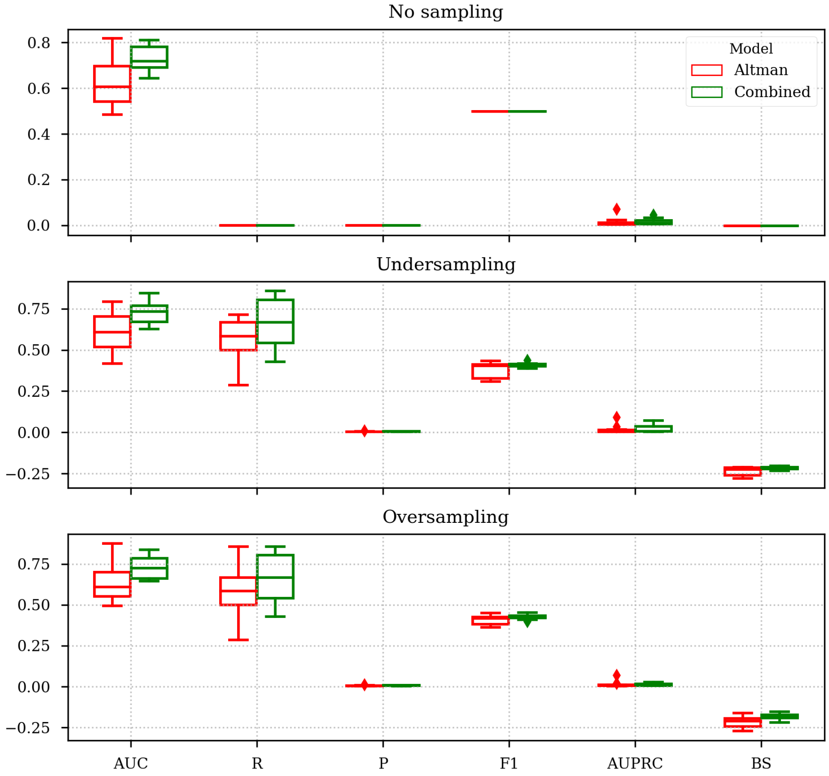


Figure 6.2: Box plots of performance metrics, SME segment

The presentations of six performance metrics in Figure 6.2 to Figure 6.4 are similar in a way that the top subfigures are for the original data, the middle subfigures are for the undersampling data, and the bottom subfigures are for the oversampling data. The box plots of six performance metrics for the SME, non-SME, and all samples are presented in Figure 6.2, 6.3, and Figure 6.4, respectively. And the average performance across 10-fold for 6 metrics is plotted in Table 6.9 below. First, it is apparent that the mean of AUCs for the combined model are all higher than that for the Altman model for the three segments and all three balancing strategies. We observe the very poor performance (especially in recall, Precision, and F1 score accordingly) of both models on the original data, where we do not employ any balancing strategy on the training set.

For the SME sample, under no balancing setting, despite there is no significant difference in recall, Precision, F1, AUPRC, and BS scores, there is an improvement in terms of the AUC. As for the undersampling and oversampling settings, first we notice higher recall compare with the no-sampling setting, and in these two sampling settings, there are improvements in terms of the AUC, recall, and Brier score when using the combined model.

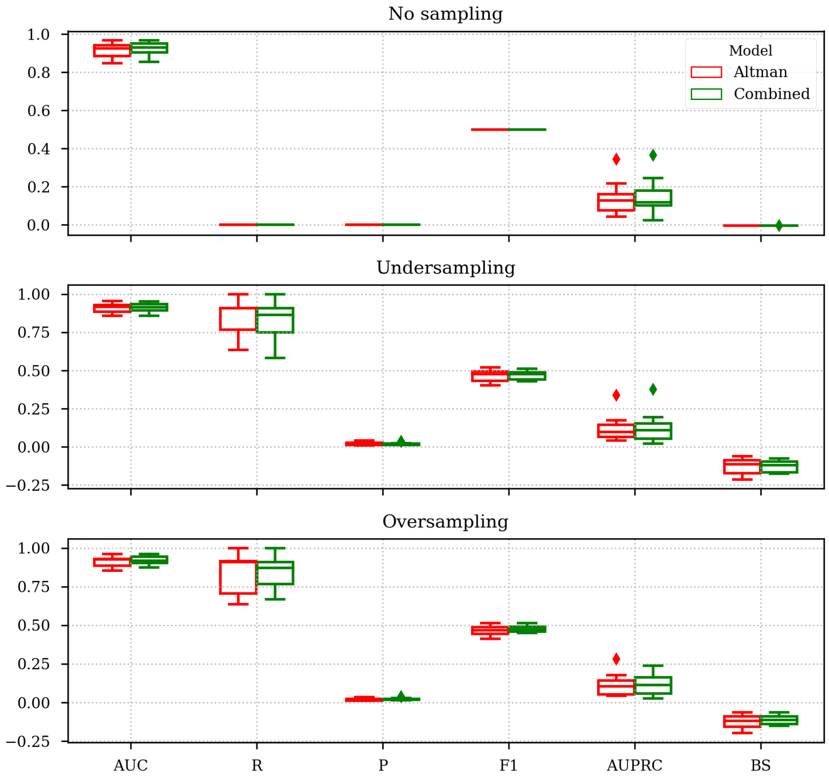


Figure 6.3: Box plots of performance metrics, non-SME segment

For all samples, there are improvements in the AUC, F1, AUPRC, and Brier score by using the textual features.

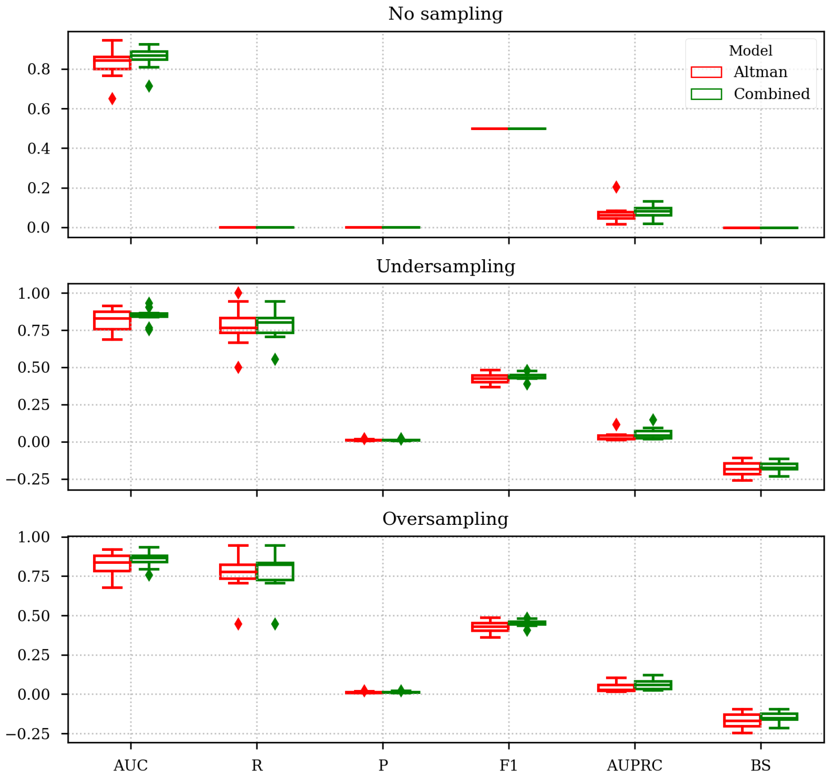


Figure 6.4: Box plots of performance metrics, all samples

We then proceed to summarise the performance of both models, test for the significant difference, and present the results in Table 6.9 as below. First, the textual features improve AUC in all sampling settings for both datasets. Excluding AUPRC for the oversampling, there are increments in all metrics for both undersampling and oversampling of the three segments. Our AUC for all samples using the combined text and accounting features for the original sample is 0.854 which is comparable with that of Mai et al. (2019) of 0.856 where they utilised the three data sources including the text, accounting, and market data for predicting the corporate default. This finding provides further evidence to the debate on to which extent we should employ deep learning techniques taking into consideration its costly computational power and poorly explainable nature.

Table 6.9: Predictive power comparison

	Altman			**Combine**		
	N	U	O	N	U	O
Panel I: SME						
AUC	0.6278	0.6075	0.6397	0.7295***	0.7227***	0.7238***
F1	_	0.3744	0.4085	_	0.4080***	0.4280***
R	0.0000	0.5792	0.5822	0.0000	0.6762***	0.6152*
P	0.0000	0.0044	0.0055	0.0000	0.0059***	0.0066***
AUPRC	0.0149	0.0173	0.0138	0.0172	0.0214***	0.0117***
BS	-0.0029	-0.2394	-0.2151	-0.0028***	-0.2159***	-0.1821***
Panel II: non-SME						
AUC	0.9108	0.9046	0.9086	0.9206***	0.9115***	0.9128***
F1	_	0.4598	0.4684	_	0.4668***	0.4780***
R	0.0000	0.8272	0.8356	0.0000	0.8314	0.8382***
P	0.0000	0.0193	0.0197	0.0000	0.0195***	0.0216***
AUPRC	0.1286	0.1153	0.1184	0.1405***	0.1231***	0.1180***
BS	-0.0037	-0.1466	-0.1212	-0.0037	-0.1383***	-0.1110***
Panel III: All sample						
AUC	0.8253	0.8123	0.8271	0.8545***	0.8457***	0.8542***
F1	_	0.4244	0.4277	_	0.4381***	0.4506***
R	0.0000	0.7750	0.7475	0.0000	0.7761	0.7887***
P	0.0000	0.0105	0.0106	0.0000	0.0113***	0.0130***
AUPRC	0.0640	0.0397	0.0375	0.0754***	0.0529***	0.0550***
BS	-0.0034	-0.1885	-0.1700	-0.0034	-0.1721***	-0.1467***

N is no sampling, U is undersampling, and O is oversampling strategy. If both recall and Precision are zero, the F1 score is ill-defined and hence is reported as '_'. ***, **, and * indicate the corresponding repeated measure t-test is significant at 99%, 95%, and 90% level, respectively.

In Table 6.10, we summarise the performance gains when comparing the Altman model with the combined model, for all metrics under three sampling strategies: Despite having lower recall under oversampling setting for the SME sample, in general, the textual features improve the classification performance. The largest improvement is in the AUC for the SME segment (8.4%-11.5%) followed by recall (3.3%-9.7%) and F1-score (2.0%-3.4%). The increment in terms of the Brier Score is ranging from 0 to 3.3% for all three segments.

One-year ahead prediction

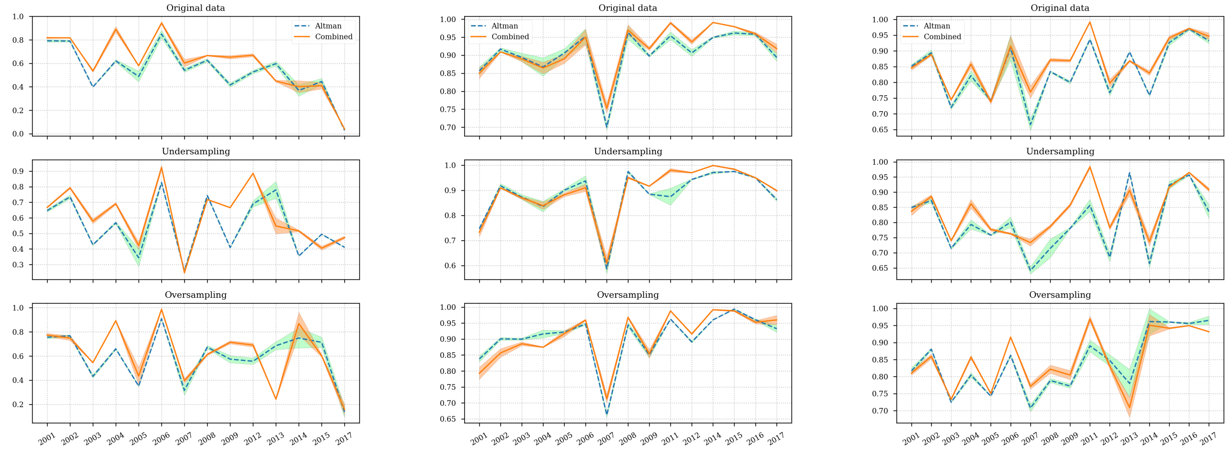
We present the one-year ahead rolling window predictions using four consecutive years to train the prediction models and predict the 5^{th}-year defaults in the test data. Specifically, we train the models in the 1997 to 2000 data and predict the defaults in 2001 data, the modelling window is then shifted one year, e.g using 1998 to 2001 data to train the model and predict the defaults in 2002 data, and so forth. In general, the presentation of predictive power is similar to the previous cross-validation settings in the way we summarise the performance of Altman and combined model for three different segments of data under six performance metrics; test for the significant difference; and aggregate the performance gains. For these detail comparisons, please refer to Table 6.11 in Appendix B.3.

Table 6.10: Performance gain (%)

	No sampling	Undersampling	Oversampling
Panel I: SME sample			
AUC	10.2	11.5	8.4
F1	_	3.4	2.0
R	_	9.7	3.3
P	_	0.1	0.1
AUPRC	0.2	0.4	-0.2
BS	_	2.3	3.3
Panel II: non-SME sample			
AUC	1.0	0.7	0.4
F1	_	0.7	1.0
R	_	0.4	0.3
P	_	_	0.2
AUPRC	1.2	0.8	_
BS	_	0.8	1.0
Panel III: All sample			
AUC	2.9	3.3	2.7
F1	_	1.4	2.3
R	_	0.1	4.1
P	_	0.1	0.2
AUPRC	1.1	1.3	1.8
BS	_	1.6	2.3

Figure 6.5 illustrates the comparison of classification performance between the Altman and combined models. The performance is measured in terms of AUC for the out-of-time test sets from 2001 to 2017. The lines represent the mean of AUC values in the test sets and their shaded areas are the bands of one standard deviation from the mean. First, Figure 6.5a presents the improvement in AUC by using the combined model from 2001 to 2012 for the SME segment, however, the improvements almost vanish from 2013 onward as both models perform poorly. This mainly affected by lacking the default data in the corresponding training windows because our SME sample does not have any default observation in 2010 and 2011, and we have only at most two default observations each year from 2012 to 2017. On the other hand, for the non-SME sample in Figure 6.5b, the AUC varying from around 0.6 to 0.9 on both models, and the combined model clearly performs better than the Altman model from 2007 onwards. Finally, for the experiment of all sample in Figure 6.5c, we observe the clear improvements in AUC from 2003 to 2012. The aggregate results are further tested in Table 6.11, which shows that the combined features help to improve the classification performance.

Our result for the AUC metric is comparable with that from Mai et al. (2019), albeit a modest higher value (0.847 compare with 0.842). However, we would like to emphasise that the AUC is not the alone decisive factor in the credit risk modelling nor we want to show the dominance performance of different ways in constructing the predictors for corporate default. Our results demonstrate that we could effectively construct the simple, significant, and intuitive features for predicting the corporate default and bridge the performance gap between the deep learning and dictionary-based approaches.



(a) SME (b) non-SME (c) All sample

Figure 6.5: One-year-ahead prediction performance (AUC)

Table 6.11: Predictive power comparison, one-year-ahead prediction

	Altman			Combine		
	N	U	O	N	U	O
Panel I: SME						
AUC	0.5458	0.5474	0.5924***	0.6216***	0.6102***	0.6281***
F1	_	0.3423	0.4207	_	0.3589***	0.4398***
R	0.0000	0.5049	0.3717	0.0000	0.5990***	0.3599
P	0.0000	0.0031	0.0037	0.0000	0.0038***	0.0043**
AUPRC	0.0101	0.0081	0.0071	0.0187***	0.0123***	0.0136***
BS	-0.0028	-0.2572	-0.1810	-0.0028	-0.2527***	-0.1461***
Panel II: non-SME						
AUC	0.9045	0.8819	0.9047	0.9176***	0.8933**	0.9086***
F1	_	0.4556	0.4841	_	0.4583***	0.4986***
R	0.0000	0.7858	0.7450	0.0000	0.8310***	0.7857**
P	0.0000	0.0157	0.0229	0.0000	0.0168***	0.0312***
AUPRC	0.0965	0.0832	0.1183	0.1517***	0.1359***	0.1387***
BS	-0.0042	-0.1676	-0.0954	-0.0042	-0.1606***	-0.0809***
Panel III: All sample						
AUC	0.8343	0.7960	0.8357	0.8611***	0.8352***	0.8468***
F1	_	0.4261	0.4481	_	0.4373***	0.4608***
R	0.0000	0.7323	0.7252	0.0000	0.7868***	0.7063**
P	0.0000	0.0086	0.0107	0.0000	0.0099***	0.0121***
AUPRC	0.0552	0.0302	0.0529	0.0685***	0.0390***	0.0559**
BS	-0.0033	-0.1973	-0.1442	-0.0033	-0.1830***	-0.1242***

N is no sampling, U is undersampling, and O is oversampling strategy. If both recall and precision are zero, the F1 score is ill-defined and hence is reported as '_'. ***, **, and * indicate the corresponding repeated measure t-test is significant at 99%, 95%, and 90% level, respectively.

6.6 Conclusions

The textual data are gathering much attention in the financial risk analysis thanks to its complement with other sources of data in explaining the manager sentiment and stock returns (Lopez Lira, 2019), uncovering the role of investment analyst report (Huang et al., 2018), or improving manager sentiment tone understanding (Zhou, 2018). This study further shows that they improve the traditional models built on the accounting data in predicting corporate default/liquidation. By using more than 50 thousand observations of listed firms in the US market from 1997 to 2017, and with simple counting for sentiment words both in the entire filing and in the MDA section of the filing, we demonstrate the high predictivity power of textual features in building the forecasting models. Despite the severe problem of IDS in default/liquidation prediction for all three data segments including SME, non-SME, and all samples, the textual features are significant and they improve the predictive power of the classification model. Our approach provides comparable and consistent predictive results, yet with more simple and intuitive features, compared with the deep learning model in Mai et al. (2019).

We found that the positive sentences in the MDA, the uncertainty words in the entire filings have the positive relationship with the probability of default. In addition, and the more positive words relative to negative words in the entire filing, the lower the probability of default. On the other hand, the higher the negative sentences in the MDA a firm has, the higher the firm being defaulted. The robustness of our findings is further strengthened by the prediction gains among six performance metrics for all three data segments using out-of-sample test sets. Interestingly, the largest improvement comes from the SME segment with the gain in AUC ranging from 8.4% to 11.5% followed by recall ranging from 3.3% to 9.7%. Besides, by using one-year-head prediction, we provide a practical investigation on improvement of the predictive power using the textual features where the combined features could significantly increase the AUC from 1.1% to 7.6% in the three corporate segments.

This study is without its limitation, first, we just examine the US listed firms, which have the benefit of the availability of both textual and financial data. Second, assigning -1 and +1 score for positive and negative words are somewhat harsh, since words might have different degrees of negative or positive, and the dictionary-based sentiment classifier needs to be evaluated in an annotated data, which is, in financial domain, very difficult and costly to obtain. Nevertheless, we believe that, this research could set a starting point on forming the intuitive and explainable textual features that could be utilised on the constructing of the credit risk models, further works could examine to what degree we should assign negative, positive, or neutral to a word or the entire sentence and radiate to the entire MDA or filing.

Chapter 7

Summary and Future Works

This study concentrated on the complement approach to the current well-developed literature in building credit risk models and emphasise on leveraging textual data to improve the financial-accounting data. Chapter 2 and 3 devote to the new survival analysis and how traditional models in credit risk pose the potential problems in bias in decision making in credit scoring in the mist of inconsistencies of laws and the development of automatic decision-making tools. Chapter 4 explores practical implications of of ensemble learning models. Chapter 5 and 6 focus on mining the textual data to improve the current practices in using traditional financial-accounting data.

This dissertation presents some contributions to the knowledge science, specifically in applying textual analysis to enhance the practice of credit risk modelling in financial industry. This is the first to employ dictionary-based and topic modelling on the distributed representation of financial filings for the task of corporate bankruptcy prediction. This dissertation presents a novel model that explore the topics in financial reports and then learn from multi-sources data to provide state-of-the-art classification performance.

This study has improved my understanding of predictive modelling, specifically for this financial textual representation, in the pragmatic and interpretable ways to solve predictive problems in credit risk modelling. As modern research leverage on multi-source and multi-model data analytics to get significant actionable insights, further research could be focusing on mining reliable sources of data to enhance the current combination of financial statements textual and numeric data. In this avenue, deep learning models will play a crucial role to leverage the distributed representation of textual financial data making use of transfer learning from other general knowledge. However, achieving improvement in predictive power while maintaining the model interpretability is still of a great challenge.

Glossary

bankruptcy a legally declared or recognized condition of insolvency of a person or organization.

default or insolvency, the failure to meet the legal obligations (or conditions) of a loan.

filing the SEC filing is a financial statement or other formal documents such as 10K, 10Q submitted to the U.S. Securities and Exchange Commission (SEC).

liquidation a company in liquidation generally is insolvent, i.e. unable to to pay its debts as they fall due. However, if a company files for Members Voluntary Liquidation, its director(s) make a declaration of solvency, confirming that the company is solvent and able to pay all of its debts in full.

z-score the Altman z-score is the output of a credit worthiness test that assess a listed company's likelihood of bankruptcy.

Acronyms

AUC Area Under the ROC Curve.

DETM Dynamic Embedding Topic Model.

LDA Latent Dirichlet Allocation.

LM Loughran and Mcdonald (2011) Financial Wordlists.

MD&A Management Discussion and Analysis.

MLP Multi-layer Perceptron.

SVM Support Vector Machine.

Bibliography

Ala'raj, M. and Abbod, M. F. (2016). Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems*, 104:89–105.

Allison, P. D. (1982). Discrete-Time Methods for the Analysis of Event Histories. *Sociological Methodology*, 13:61–98.

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, 23(4):589–609.

Altman, E. I., Haldeman, R. G., and Narayanan, P. (1977). ZETATM analysis A new model to identify bankruptcy risk of corporations. *Journal of Banking & Finance*, 1(1):29–54.

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., and Suvas, A. (2017). Financial Distress Prediction in an International Context: A Review and Empirical Analysis of Altman's *Z*- Score Model. *Journal of International Financial Management & Accounting*, 28(2):131–171.

Altman, E. I. and Sabato, G. (2007). Modelling Credit Risk for SMEs: Evidence from the U.S. Market. *Abacus*, 43(3):332–357.

Altman, E. I., Sabato, G., and Wilson, N. (2010). The value of non-financial information in SME risk management. *Journal of Credit Risk*, 6(2):95–127.

Andreeva, G., Ansell, J., Crook, J. N., University of Edinburgh, and Credit Research Centre (2004). *Credit scoring in the context of European integration: assessing the performance of the generic models*. Credit Research Centre, University of Edinburgh, Edinburgh. OCLC: 60368040.

Andreeva, G., Calabrese, R., and Osmetti, S. A. (2016). A comparative analysis of the UK and Italian small businesses using Generalised Extreme Value models. *European Journal of Operational Research*, 249(2):506–516.

Andreeva, G. and Matuszyk, A. (2019). The law of equal opportunities or unintended consequences?: The effect of unisex risk assessment in consumer credit. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, page rssa.12494.

Angelini, E., di Tollo, G., and Roli, A. (2008). A neural network approach for credit risk evaluation. *The Quarterly Review of Economics and Finance*, 48(4):733–755.

Baesens, B., Gestel, T. V., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *J Oper Res Soc*, 54(6):627–635.

Banasik, J., Crook, J. N., and Thomas, L. C. (1999). Not if but when will borrowers default. *Journal of the Operational Research Society*, 50(12):1185–1190.

Bao, Y. and Datta, A. (2014). Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures. *Management Science*, 60(6):1371–1391.

Bijak, K. and Thomas, L. C. (2012). Does segmentation always improve model performance in credit scoring? *Expert Systems with Applications*, 39(3):2433–2442.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pages 113–120, Pittsburgh, Pennsylvania. ACM Press.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Bonsall, S. B., Leone, A. J., Miller, B. P., and Rennekamp, K. (2017). A plain English measure of financial reporting readability. *Journal of Accounting and Economics*, 63(2-3):329–357.

Buuren, S. v. and Groothuis-Oudshoorn, K. (2011). **mice**: Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software*, 45(3).

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, N., Ribeiro, B., and Chen, A. (2016). Financial credit risk assessment: a recent review. *Artificial Intelligence Review*, 45(1):1–23.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Crone, S. F. and Finlay, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, 28(1):224–238.

Crook, J. and Banasik, J. (2004). Does reject inference really improve the performance of application scoring models? *Journal of Banking & Finance*, 28(4):857–874.

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM.

Demar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30.

Dieng, A. B., Ruiz, F. J., and Blei, D. M. (2019a). Topic modeling in embedding spaces. *arXiv preprint arXiv:1907.04907*.

Dieng, A. B., Ruiz, F. J. R., and Blei, D. M. (2019b). The Dynamic Embedded Topic Model. *arXiv:1907.05545 [cs, stat]*. arXiv: 1907.05545.

Elshandidy, T., Shrives, P. J., Bamber, M., and Abraham, S. (2018). Risk reporting: A review of the literature and implications for future research. *Journal of Accounting Literature*, 40:54–82.

Engelberg, J. E., Reed, A. V., and Ringgenberg, M. C. (2012). How are shorts informed? *Journal of Financial Economics*, 105(2):260–278.

Gandhi, P., Loughran, T., and McDonald, B. (2018). Using Annual Report Sentiment as a Proxy for Financial Distress in U.S. Banks. SSRN Scholarly Paper ID 2905225, Social Science Research Network, Rochester, NY.

Gandhi, P., Loughran, T., and McDonald, B. (2019). Using Annual Report Sentiment as a Proxy for Financial Distress in U.S. Banks. *Journal of Behavioral Finance*, 20(4):424–436.

Garca, S., Fernndez, A., Luengo, J., and Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, 180(10):2044–2064.

Garca, V., Marqus, A. I., and Snchez, J. S. (2019). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. *Information Fusion*, 47:88–101.

Goadrich, M., Oliphant, L., and Shavlik, J. (2006). Gleaner: Creating ensembles of first-order clauses to improve recall-precision curves. *Machine Learning*, 64(1-3):231–261.

Guan, L., He, S. D., and McEldowney, J. (2008). Window Dressing in Reported Earnings. *Com. Lending Rev.*, 23:26.

Hand, D. J. and Anagnostopoulos, C. (2013). When is the Area Under the Receiver Operating Characteristic Curve an Appropriate Measure of Classifier Performance? *Pattern Recogn. Lett.*, 34(5):492–495.

Harrell, F. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer Series in Statistics. Springer-Verlag, New York.

Harrell, F. E. (2015). *Regression Modeling Strategies*. Springer Series in Statistics. Springer International Publishing, Cham.

He, H. and Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Healy, P. M., Hutton, A. P., and Palepu, K. G. (1999). Stock Performance and Intermediation Changes Surrounding Sustained Increases in Disclosure. *Contemporary Accounting Research*, 16(3):485–520.

Hoffman, M., Bach, F. R., and Blei, D. M. (2010). Online Learning for Latent Dirichlet Allocation. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc.

Huang, A. H., Lehavy, R., Zang, A. Y., and Zheng, R. (2018). Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach. *Management Science*, 64(6):2833–2855.

Huang, G., Huang, G.-B., Song, S., and You, K. (2015). Trends in extreme learning machines: A review. *Neural Networks*, 61:32–48.

Huang, K.-W. and Li, Z. (2011). A multilabel text classification algorithm for labeling risk factors in SEC form 10-K. *ACM Trans. Manage. Inf. Syst.*, 2(3):1–19.

Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., and Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, 37(4):543–558.

Hutto, C. J. and Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Jiang, F., Lee, J., Martin, X., and Zhou, G. (2019). Manager sentiment and stock returns. *Journal of Financial Economics*, 132(1):126–149.

Jo, H., Han, I., and Lee, H. (1997). Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert Systems with Applications*, 13(2):97–108.

Koh, H. C., Tan, W. C., and Goh, C. P. (2015). A two-step method to construct credit scoring models with data mining techniques. *International Journal of Business and Information*, 1(1).

Kraus, M. and Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, 104:38–48.

Lawrence, A. (2013). Individual investors and financial disclosure. *Journal of Accounting and Economics*, 56(1):130–147.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*.

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136.

Lin, S.-M., Ansell, J., and Andreeva, G. (2012). Predicting default of a small business using different definitions of financial distress. *Journal of the Operational Research Society*, 63(4):539–548.

Lopez Lira, A. (2019). Risk Factors That Matter: Textual Analysis of Risk Disclosures for the Cross-Section of Returns. *SSRN Journal*.

Loughran, T. and Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Mai, F., Tian, S., Lee, C., and Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, 274(2):743–758.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., and Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Menardi, G. and Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1):92–122.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *ICLR Workshop Proceedings. arXiv:1301.3781.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Neural Information Processing Systems.*

Nemoto, N., Yoshino, N., Okubo, Y., Inaba, D., and Yanagisawa, K. (2018). Credit risk reduction effect on small and medium-sized enterprise finance through the use of bank account information. ADBI Working Paper 857, Asian Development Bank Institute (ADBI), Tokyo.

Nguyen, T. H., Shirai, K., and Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24):9603–9611.

Peji Bach, M., Krsti, ., Seljan, S., and Turulja, L. (2019). Text Mining for Big Data Analysis in Financial Sector: A Literature Review. *Sustainability*, 11(5):1277.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Sun, J., Li, H., Huang, Q.-H., and He, K.-Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, 57:41–56.

Thomas, L., Crook, J., and Edelman, D. (2017). *Credit Scoring and Its Applications*. SIAM-Society for Industrial & Applied Mathematics, Philadelphia, 2nd revised edition edition edition.

Tsai, C.-F. and Chen, M.-L. (2010). Credit rating by hybrid machine learning techniques. *Applied Soft Computing*, 10(2):374–380.

Tsai, C.-F. and Hung, C. (2014). Modeling credit scoring using neural network ensembles. *Kybernetes*, 43(7):1114–1123.

Wang, G., Hao, J., Ma, J., and Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, 38(1):223–230.

Wiginton, J. C. (1980). A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *The Journal of Financial and Quantitative Analysis*, 15(3):757–770.

Wu, H.-C., Hu, Y.-H., and Huang, Y.-H. (2014). Two-stage credit rating prediction using machine learning techniques. *Kybernetes*, 43(7):1098–1113.

Xiao, H., Xiao, Z., and Wang, Y. (2016). Ensemble classification based on supervised clustering for credit scoring. *Applied Soft Computing*, 43:73–86.

Yu, L., Wang, S., and Lai, K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Systems with Applications*, 34(2):1434–1444.

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., and Wasinger, R. (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scoring. *Expert Systems with Applications*, 42(7):3508–3516.

Zhou, G. (2018). Measuring Investor Sentiment. *Annual Review of Financial Economics*, 10:239–259.

Publication and Working Papers

Nguyen, B.-H., Andreeva, G., Huynh, N., 2019. Time to Liquidation of SMEs: The Predictability of Survival Models, in: Chen, J., Huynh, V.N., Nguyen, G.-N., Tang, X. (Eds.), *Knowledge and Systems Sciences, Communications in Computer and Information Science.* Springer, Singapore, pp. 186200. https://doi.org/10.1007/978-981-15-1209-4_14

Nguyen, H.B., Huynh, V.-N., 2020. On Sampling Techniques for Corporate Credit Scoring. *Journal of Advanced Computational Intelligence and Intelligent Informatics* 24, 4857. https://doi.org/10.20965/jaciii.2020.p0048

Ba-Hung Nguyen and Van-Nam Huynh. Textual Analysis and Corporate Bankruptcy: A Financial Dictionary-based Sentiment Approach. **Forthcoming: *Journal of Operational Research Society***, Taylor and Francis (Accepted Jun. 2020)

Ba-Hung Nguyen, Koyaki Shirai, and Van-Nam Huynh. Topics in Financial Filings and Bankruptcy Prediction with Distributed Representations of Textual Data. **Accepted, ECML-PKDD**, Sep. 2020: *European Conference in Machine Learning and Principles and Practice of Knowledge Discovery in Databases.*

Appendix A

Chapter 5

A.1 Descriptive Statistics of Financial Elements and Demographic Features

Table A.1: Financial Elements and Demographic Features

Name	**Description**
at	Total assets
act	Total current assets
intan	Total intangible assets
invt	Total assets
ch	Cash
dvt	Total dividends
lct	Total current liabilities
lt	Total liabilities
wcap	Working capital
revt	Total revenue
re	Retained earning
ebit	Earning before interest and taxes
mkvalt	Total market value
sale	Sale
seq	Shareholder equity
ni	Net income
dltt	Total long-term debts
dm	Debt mortgages & other secured
emp	Total employees
gdwl	Goodwill
addzip	Zip code
sic	Standard industrial code
ggroup	Global industrial classification (GIC) group
gind	GIC industries
gsubin	GIC sub-industries
idbflag	International, Domestic, Both Indicator
incorp	Date of incorporation
spcsrc	S&P Quality Ranking - Current
au	Auditor
auop	Auditor opinion

Table A.2: Descriptive Statistics for Accounting and Demographic Features

PANEL 1: Financial Statements Elements

	at	act	intan	invt	ch	dvt	lct	lt	wcap	revt
count	86196	84797	81638	85708	85565	84619	84837	86196	84731	85921
mean	4118.302	1217.743	858.391	284.317	272.502	89.814	928.688	2491.773	279.546	2934.885
std	19224.416	5655.499	5462.940	1347.345	1292.234	592.195	4827.252	12177.598	1799.280	14150.218
min	0.001	-0.168	-0.423	0.000	-0.134	-325.377	0.000	0.001	-43132.545	-1964.999
25%	26.927	14.269	0.000	0.381	1.805	0.000	5.922	9.265	1.703	13.857
50%	186.378	82.971	3.870	10.634	13.877	0.000	33.375	72.032	28.528	138.835
75%	1246.340	414.989	116.742	92.331	82.579	3.874	213.332	687.642	155.221	969.236
max	507560.425	192486.646	225278.000	48586.913	53528.000	67643.805	192819.656	460442.000	56120.000	470171.000

	re	ebit	mkvalt	sale	seq	ni	dltt	dm	emp	gdwl
count	84042	85793	61887	85807	86195	85805	86141	80436	78807	80728
mean	821.767	333.038	3064.844	2937.978	1554.264	176.908	921.058	150.610	8.641	498.611
std	7845.878	1800.682	15901.855	14159.247	8001.189	1527.857	4476.683	1075.271	30.442	3138.554
min	-143336.328	-25913.000	0.000	-1964.999	-86154.000	-98696.000	-0.023	0.000	0.000	-0.423
25%	-58.985	-3.571	27.282	13.832	8.472	-8.340	0.000	0.000	0.088	0.000
50%	-1.867	4.762	173.450	138.825	74.588	0.715	8.898	0.180	0.619	0.000
75%	122.538	84.664	989.844	969.980	463.727	37.491	287.779	19.804	4.000	55.402
max	398278.000	71230.000	790050.098	470171.000	284434.000	104821.000	207174.000	59127.799	863.824	146583.307

PANEL 2: Demographic Features

	addzip	city	state	county	sic	ggroup	gind	gsector	gsubind	idbflag	incorp	spcsrc	au	auop
count	86196	86196	86196	86196	86196	86196	86196	86196	86196	86196	86196	86196	86196	86196
unique	4601	2178	62	37	266	25	67	12	148	2	55	10	25	7
top	nan	Houston	nan	nan	2836	3520.0	352010.0	35.0	35201010.0	D	DE	nan	9.0	1.0
freq	1376	3302	13037	86026	5422	11224	6770	17232	6770	72957	45452	37054	17463	57263

A.2 Topic Wordclouds

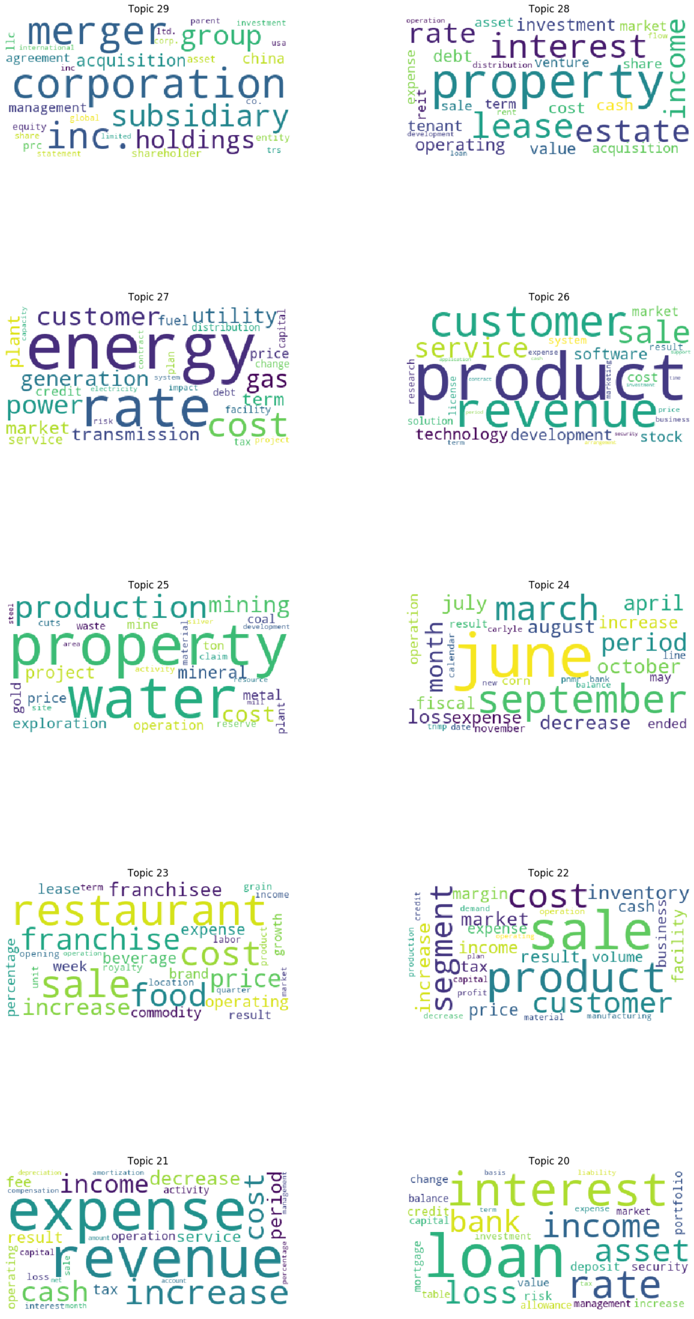
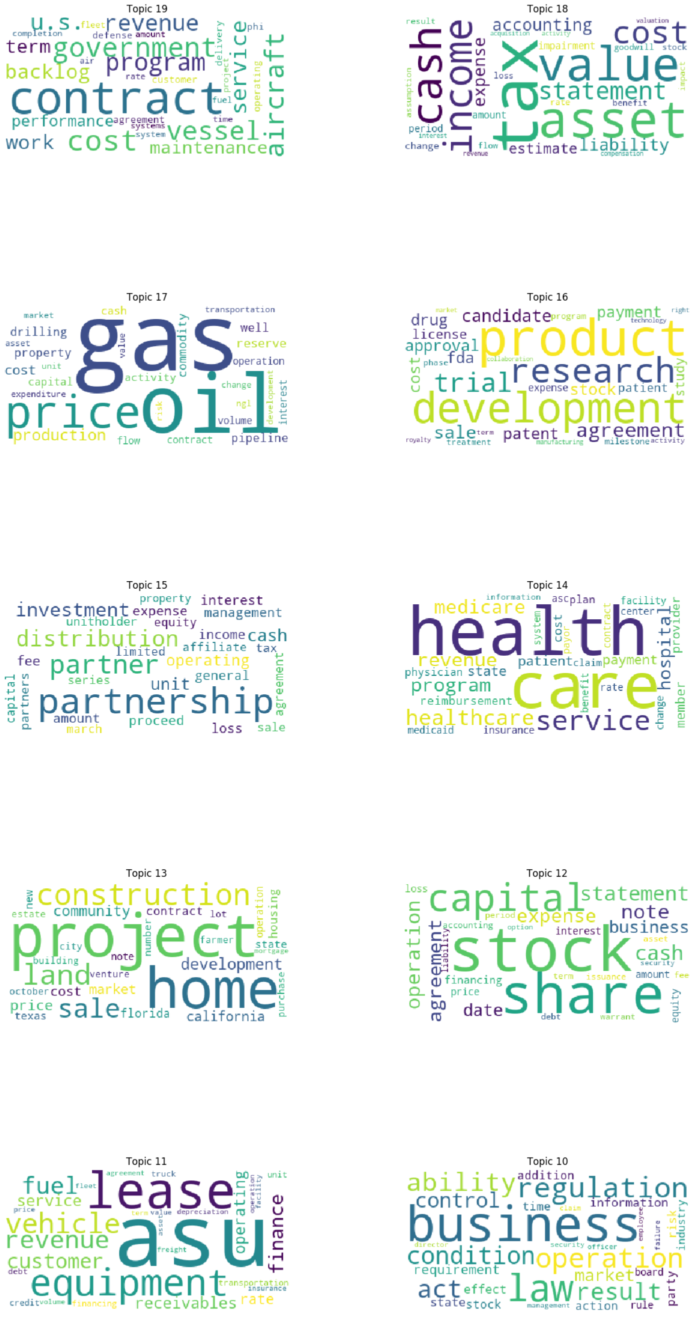


Figure A.1: Topic wordcloud



Topic 19

Topic 18

Topic 17

Topic 16

Topic 15

Topic 14

Topic 13

Topic 12

Topic 11

Topic 10

Figure A.2: Topic wordcloud (cont.)

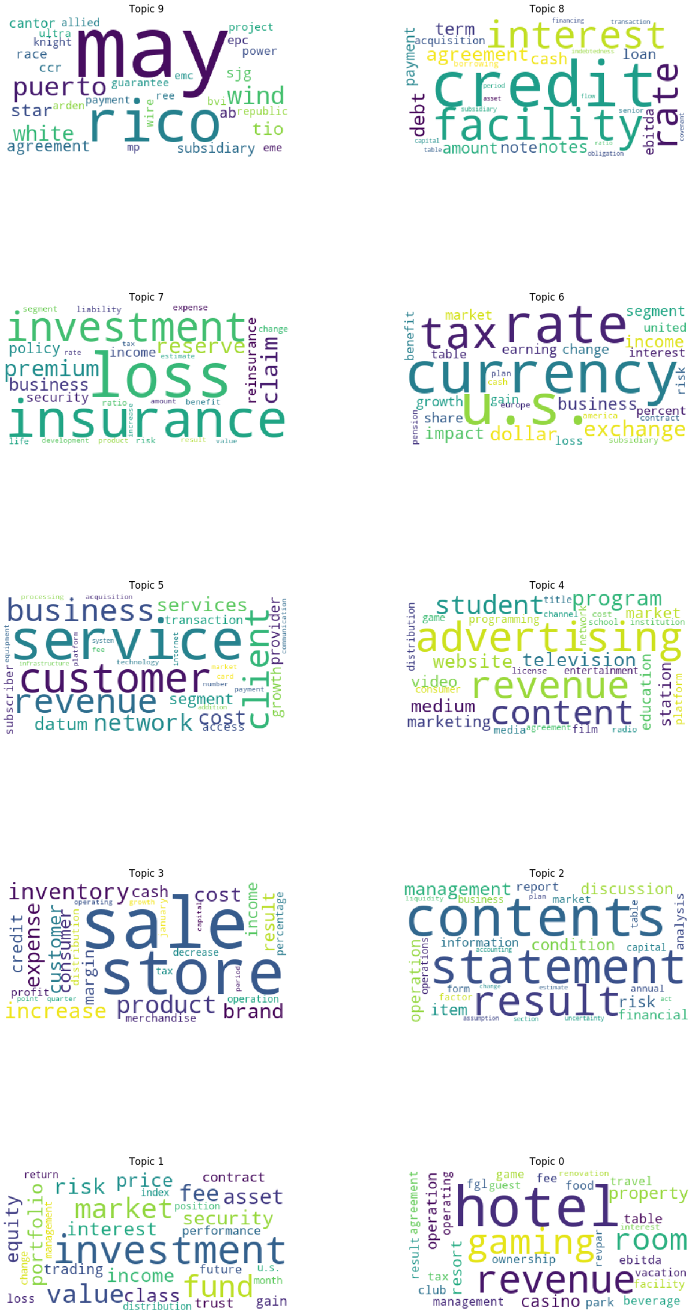


Figure A.3: Topic wordcloud (cont.)

A.3 Performance of Textual Feature Sets

Table A.3: Performance of five textual feature sets

Year	**AUC**	**AUPRC**	**BS**
Panel 1: z-score features			
2007	0.8217	0.0110	0.0028
2008	0.7946	0.0525	0.0072
2009	0.8179	0.0023	0.0005
2010	0.9268	0.0459	0.0013
2011	0.7850	0.0756	0.0025
2012	0.8271	0.0036	0.0009
2013	0.7039	0.0203	0.0013
2014	0.8299	0.0098	0.0027
2015	0.9887	0.3044	0.0029
2016	0.9756	0.5098	0.0010
Mean	0.8471	0.1035	0.0023
Panel 2: dictionary-based features			
2007	0.8116	0.0124	0.0034
2008	0.7980	0.0268	0.0068
2009	0.8408	0.0027	0.0004
2010	0.9712	0.1027	0.0013
2011	0.5321	0.0066	0.0022
2012	0.6200	0.0016	0.0008
2013	0.8384	0.0075	0.0013
2014	0.7992	0.0143	0.0025
2015	0.9712	0.1395	0.0027
2016	0.8771	0.0058	0.0010
Mean	0.8059	0.0320	0.0022
Panel 3: doc2vec features			
2007	0.7643	0.0190	0.0028
2008	0.8434	0.1267	0.0070
2009	0.9878	0.0345	0.0006
2010	0.9306	0.0199	0.0014
2011	0.7318	0.1723	0.0024
2012	0.6445	0.0022	0.0009
2013	0.7123	0.0098	0.0013
2014	0.7922	0.0193	0.0027
2015	0.9153	0.0303	0.0029
2016	0.8433	0.0054	0.0010
Mean	0.8165	0.0439	0.0023
Panel 4: 30-topic features (LDA)			
2007	0.8150	0.0123	0.0028
2008	0.7689	0.0335	0.0072
2009	0.8547	0.0030	0.0005
2010	0.9348	0.0421	0.0013
2011	0.7928	0.0682	0.0025
2012	0.7208	0.0022	0.0009
2013	0.7243	0.0139	0.0013
2014	0.8149	0.0097	0.0027
2015	0.9826	0.1425	0.0029
2016	0.9396	0.0134	0.0010
Mean	0.8348	0.0341	0.0023
Panel 5: 30-topic features (DETM)			
2007	0.7267	0.0125	0.0028
2008	0.7437	0.0335	0.0072
2009	0.7550	0.0018	0.0005
2010	0.9261	0.0549	0.0013
2011	0.8195	0.1932	0.0025
2012	0.6947	0.0020	0.0009
2013	0.6963	0.0355	0.0013
2014	0.8191	0.0089	0.0027
2015	0.9824	0.1533	0.0029
2016	0.9510	0.5049	0.0010
Mean	0.8115	0.1000	0.0023

Appendix B

Chapter 6

B.1 Descriptive Statistics and Pair-wise Correlation Plots

Table B.1: Missing Values of Financial elements, Altman Z-core 5-factor, and Altman et al. (2010) SME 5-factor, All Sample

	Count	Proportion
ipodate	25620	0.5037
mkvalt	5871	0.1154
xint	3290	0.0647
gdwl	2713	0.0533
dm	2363	0.0465
intan	2305	0.0453
emp	1338	0.0263
wcap	809	0.0159
act	805	0.0158
cstk	782	0.0154
re	777	0.0153
lct	758	0.0149
rect	367	0.0072
invt	308	0.0061
ch	281	0.0055
dp	165	0.0032
ebit	151	0.0030
ni	149	0.0029
sale	149	0.0029
np	114	0.0022
dvt	110	0.0022
revt	74	0.0015
dltt	32	0.0006
dlc	21	0.0004
Z1	809	0.0159
Z2	777	0.0153
Z3	151	0.0030
Z4	782	0.0154
Z5	149	0.0029
A1	199	0.0039
A2	28	0.0006
A3	281	0.0055
A4	8395	0.1650
A5	367	0.0072

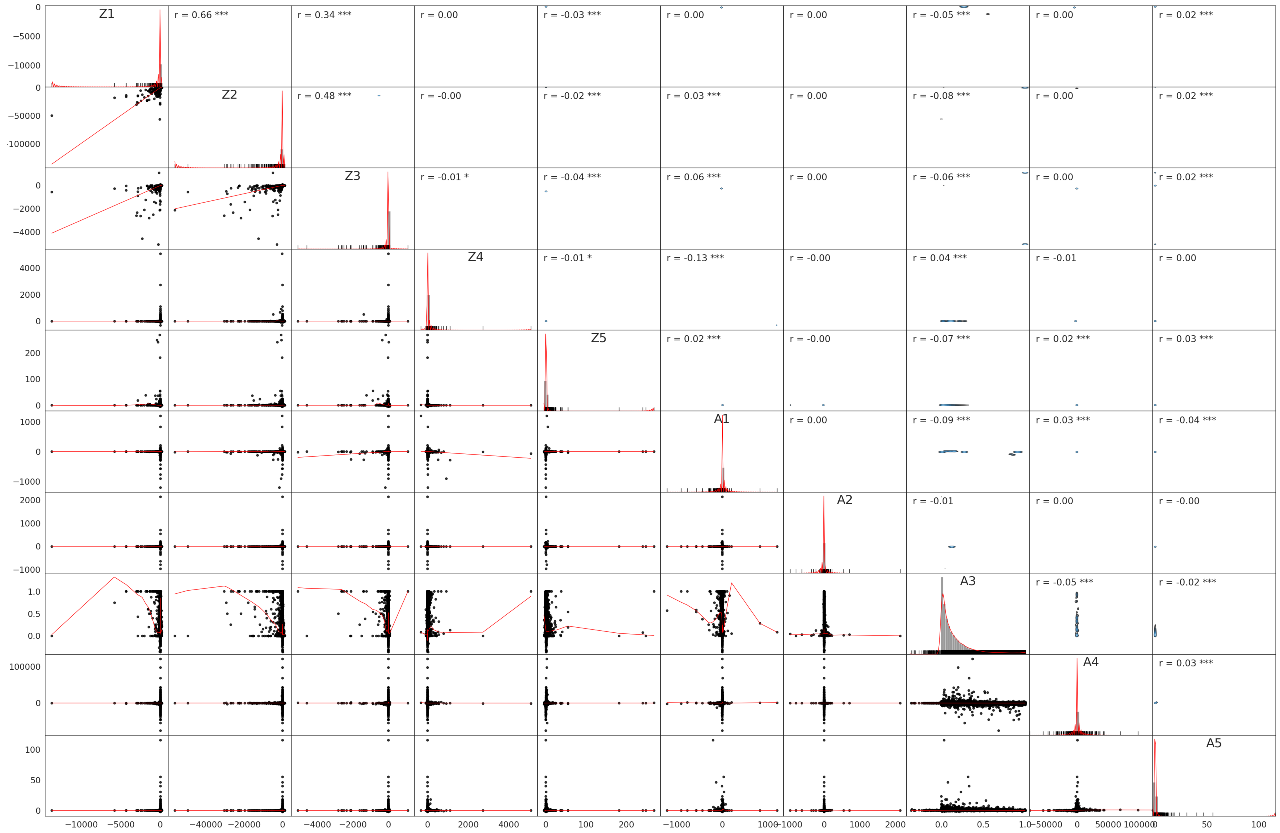


Figure B.1: The correlation of Altman's factors

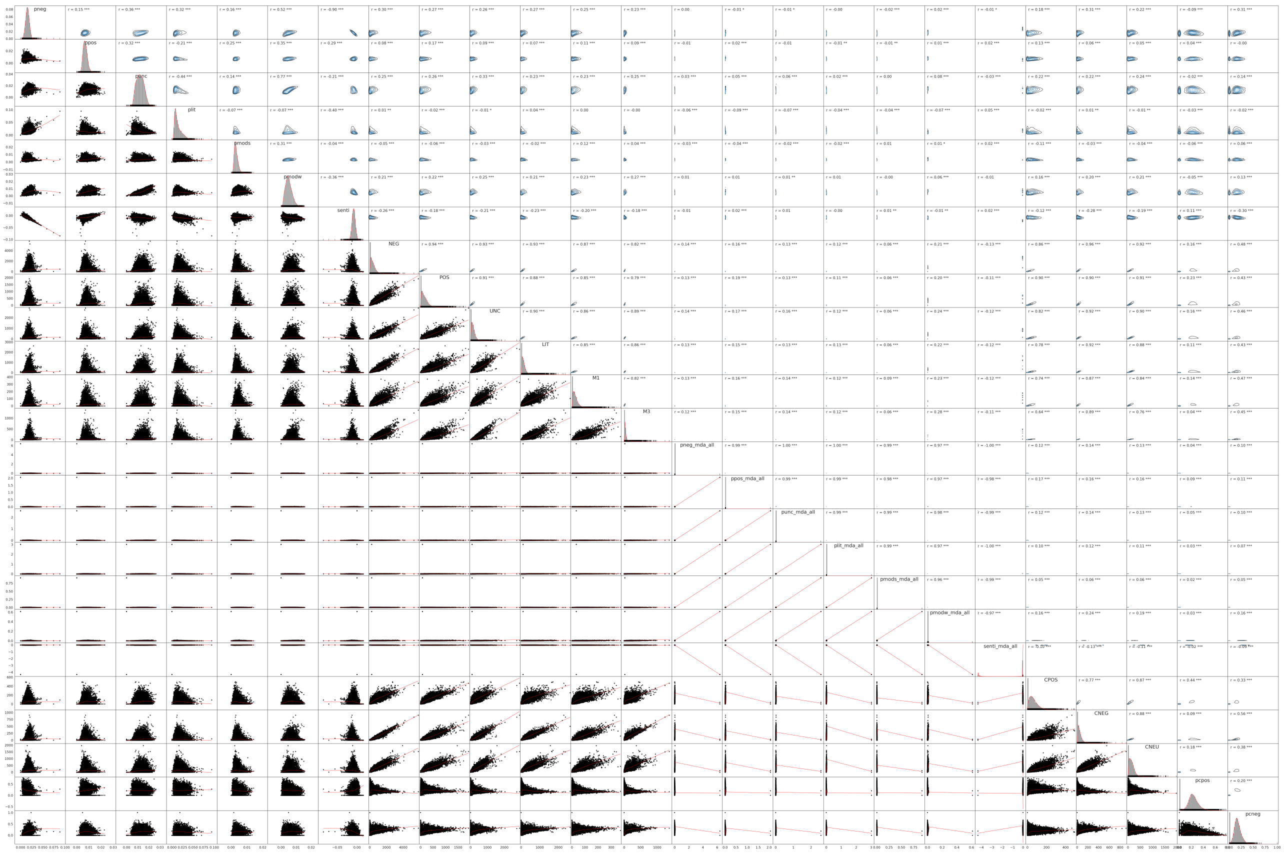


Figure B.2: The correlation of textual features

B.2 Lasso Paths of Regression Coefficients

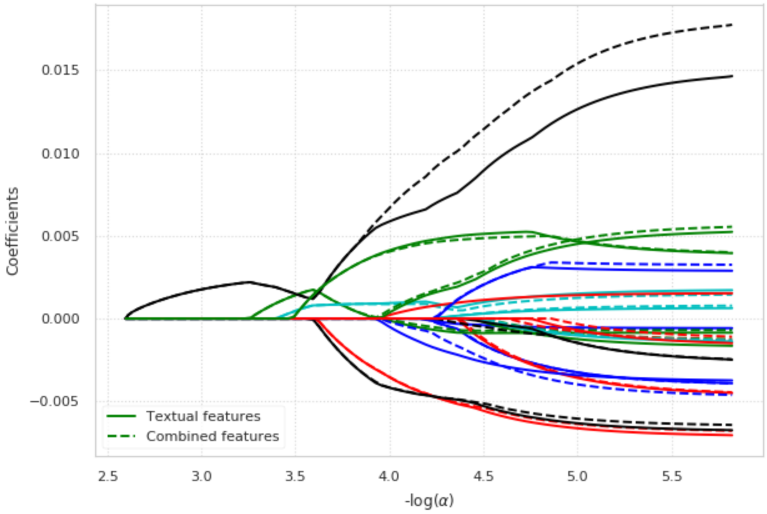


Figure B.3: SME sample

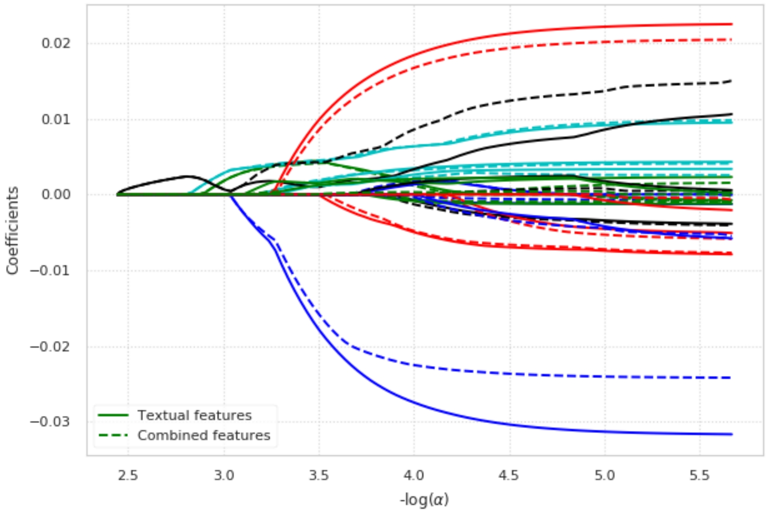


Figure B.4: non-SME sample

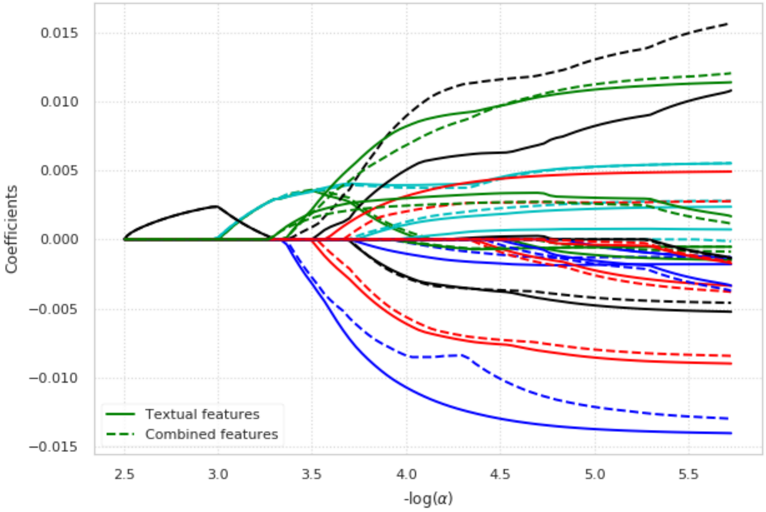


Figure B.5: All sample

B.3 One-year-ahead prediction performance

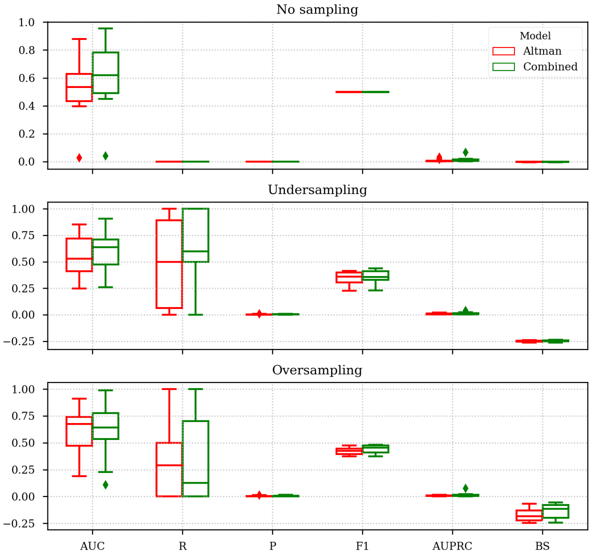


Figure B.6: The box plots of performance metrics, SME sample

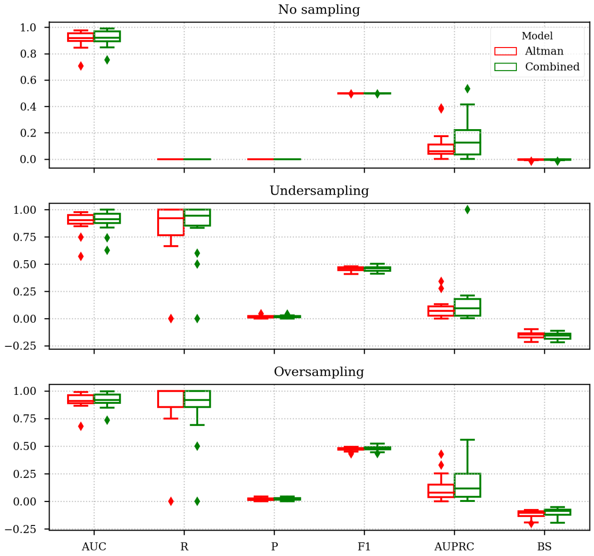


Figure B.7: The box plots of performance metrics, non-SME samples

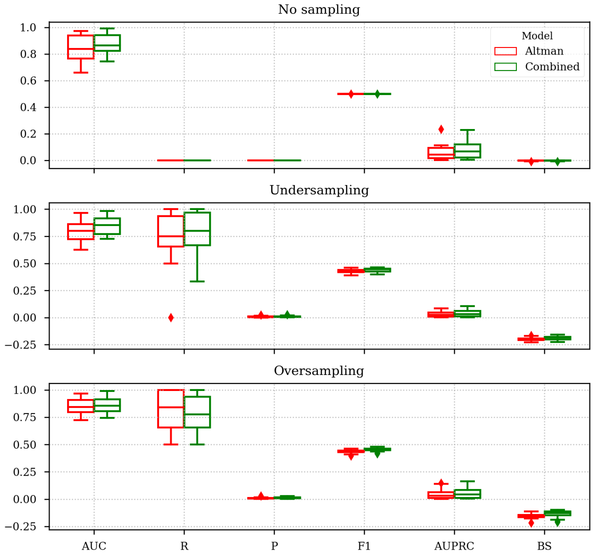


Figure B.8: The box plots of the performance metrics, all samples

Table B.2: The performance gain of one-year-ahead prediction (%)

	No sampling	Undersampling	Oversampling
Panel I: SME sample			
AUC	7.6	6.3	3.6
F1	0.0	1.7	1.9
R	0.0	9.4	-1.2
P	0.0	0.1	0.1
AUPRC	0.9	0.4	0.7
BS	0.0	0.5	3.5
Panel II: non-SME sample			
AUC	1.3	1.1	0.4
F1	0.0	0.3	1.4
R	0.0	4.5	4.1
P	0.0	0.1	0.8
AUPRC	5.5	5.3	2.0
BS	0.0	0.7	1.4
Panel III: All sample			
AUC	2.7	3.9	1.1
F1	0.0	1.1	1.3
R	0.0	5.4	-1.9
P	0.0	0.1	0.1
AUPRC	1.3	0.9	0.3
BS	0.0	1.4	2.0