PREDICTION OF CORPORATE DEFAULT USING LOGISTIC REGRESSION

by

Jingxuan Wu
BMS, Shandong University (2014)

and

Priyaadarshini Elango
BTech, Anna University (2011)

PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN FINANCE

In the Master of Science in Finance Program
of the
Faculty
of
Business Administration

SIMON FRASER UNIVERSITY

Fall 2019

Approval

Name:	**Jingxuan Wu and Priyaadarshini Elango**
Degree:	**Master of Science in Finance**
Title of Project:	**Prediction of corporate default using logistic regression.**

Supervisory Committee:

____

Dr Deniz Anginer
Senior Supervisor
Assistant Professor

____

Carlos da Costa
Second Reader
Lecturer

Date Approved: ____

Abstract

The main aim of the research is to examine the importance of Merton's (1974) distance-to-default measure in predicting corporate defaults. The data sample includes 75,667 companies from 1975 to 2007. We compare the predictive power of Merton's distance-to-default to accounting variables used in Ohlson (1980), Altman (1968), and a set of market measures used in Campbell et al. (2008).

The marginal effect is used to evaluate the efficiency of the independent variables to forecast corporate defaults. The relative or receiver operating characteristic (ROC) curve is used to show the accuracy of the model. The findings show that Merton distance to default improves the efficiency of the model and has a high marginal effect among the independent variables, as shown in the paper.

Keywords: Default; logistic regression; marginal effect; relative or receiver operating characteristic; Merton's Distance-to-Default.

Acknowledgments

We would like to sincerely thank our advisor Dr. Deniz Anginer, the Faculty of Business Administration at Simon Fraser University. Deniz has always answered our questions whenever we ran into a trouble spot in our research or writing. He consistently offered us guidance and provided room for individual learning.

We want to also express our heartfelt thanks and appreciation to our second reader Carlos da Costa, for his guidance and assistance.

Table of Contents

List of Figures

List of Tables

1: Introduction

The importance of default prediction models has increased after the global financial crisis. Multivariate discriminant analysis (MDA) and logistic regressions have often been used in prediction models. Statistical methods gauge the importance of a parameter and clarify the effects of predictor variables. Horta, Borges, Carvalho, and Alves (2011 and 2013) report say default prediction systems offer an analytical resource independent of personal factors for investors and credit managers. Therefore, the quantitative prediction models allow a robust assessment of the potential capacity of a company to continue to meet its financial commitments.

According to Bellovary, Giacomino, and Akers (2007), the forecast models, despite their differences, show high predictive potential, implying that they are valuable to auditors, managers, creditors, and analysts. Due to the loss of reliability of the parameters over time, Santos, Colauto, and Pinheiro (2009) emphasize the importance of upgrading such models. Balcaen and Ooghe (2004) point out that these errors exist mainly in simulations that include financial factors solely because they do not recognize macroeconomic conditions.

Korol and Korodi (2010) report that no single factor is mindful of a company's insolvency. There is a consensus on the presence of two sets of variables. The first includes endogenous causes, which happen inside a company and are related to inept asset allocation, to unproductive debt-equity structure, and inadequate company administration. The second factor is economic factors like the fiscal, monetary, and exchange rate policies, which the companies cannot control. In any case, such components influence companies' financial situation.

In this paper, our objective is to forecast default using the logistic regression with the Merton's (1974) distance-to-default variable. We also examine the incremental power of Merton's distance-to-default variable compared to accounting variables that have been previously used in the literature. As shown by Hillegeist et al. (2004), market variables play a vital role in predicting defaults. Merton model provides a way to combine key two key variables (market capitalization and equity volatility) along with book leverage in a single number. We use logistic regression to determine the importance of the Merton distance-to-default in default prediction after controlling a set of accounting variables used in Ohlson (1980), Altman (1968), and a set of market variables used in Campbell et al. (2008).

The rest of the paper is structured as follows: Section 2 includes an outline of existing literature on default prediction; Section 3 describes the sample and the variables used in the analyses; Section four describes the methodology used; Section 5 discusses the results. Finally, we describe the study's key conclusions and future approaches.

2: Literature Review

2.1 Univariate models for default prediction

The initial studies in the field of default prediction were the univariate models, models with a single variable. These studies were used individual company ratios and compared the values of these ratios for default and non-default companies. These studies served as a basis for future studies that have employed multivariate models, models that use more than one variable to predict the failure.

In 1930, the Bureau of Business analysis revealed the results of a study connected to the values of financial ratios of default industrial companies. The report studied 24 financial ratios, using data from 29 firms and meant to see similar characteristics of failed firms. Results disclosed eight financial ratios (Working Capital to Total Assets, Surplus and Reserves to Total Assets, Net Worth to Fixed Assets, Fixed Assets to Total Assets, the Current Ratio, Net Worth to Total Assets, Sales to Total Assets, and Cash to Total Assets), as factors of "weakening" financial situation of a company. According to the analysis, the primary consideration was found to be the ratio of working capital to total assets and then was the current ratio.

Fitzpatrick (1932) compared 13 financial ratios for 19 bankrupt companies and 19 non-bankrupt companies. He identified that in most of the cases, non-bankrupt companies showed favorable ratios compared to the customary level while insolvent companies showed unfavorable ratios. Results showed that the two most vital ratios were Net Worth to Debt and Net Profits to Net Worth. According to Fitzpatrick ought to concentrate less on the current ratio and quick ratio of companies with long-term liabilities.

Smith and Winakor (1935) identified the financial ratios of 183 companies operating in different industries. From the study, they found that the proportion of Working Capital to Total Assets was a better indicator for predicting financial problems than the Cash to Total Assets Ratio and the Quick Ratio. Further, they indicated that impending insolvency could be predicted based on the fall in the ratio of current assets to total assets.

Another study for predicting failure of companies is the one showed by Merwin (1942), that centered on small manufacturers. According to Merwin, bankrupt companies revealed signals of weakness 4 or 5 years before failure. Based on the analysis, three ratios resulted vital in the default of a firm: Net Working Capital to Total Assets, the Current Ratio, and Net Worth to Total Debt.

In default prediction, Beaver's (1966) univariate analysis is the most cited. He compared 29 financial ratios of bankrupted companies against non-bankrupted companies for the same five years period before default. His study was to find which among the ratios best indicates bankrupt companies from the non-bankrupt ones. He also identified how many years before default, we can see the differences between bankrupt and non-bankrupt companies. From the 29 ratios, the below six were found to have better predictability net income before amortization, depreciation, and depletion / total liabilities, net income / total assets, total debt / total assets, net-working capital / total assets, short-term assets / short-term liabilities, cash, short-term investments, and receivables / operating expenses without depreciation, depreciation, and depletion.)

2.2 Multivariate models for default prediction

After Beaver's univariate examination, several studies in default forecasting applied the statistical technique of MDA. MDA enables concurrent consideration of numerous financial variables on the purpose of developing a default prediction model. Though early bankruptcy prediction studies (Altman, 1968; Blum, 1974; Dambolena & Khoury; 1980; Deakin, 1972) used MDA to forecast failure, its application relies on three postulations for proper form. According to Lennox (1999), the three assumptions are:

(1) the independent variables are presumed to have a multivariate normal distribution;

(2) the samples of companies are drawn at random from their respective populations;

(3) If the restriction of equal cluster covariate matrices fulfilled, then MDA is optimal.

Altman (1968) first used MDA to predict the company's failure. In his study, he matched a sample of 33 manufacturing companies that filed for bankruptcy petitions under Chapter 11 with an example of 33 non-bankrupt manufacturing companies in terms of asset size and industry classification.

Altman's model studied five parameters (working capital/total assets, retained earnings/total assets, earnings before interest and taxes (EBIT)/total assets, market value of equity/par value of debt, and sales/total assets) and arrived at 79% prediction accuracy for the holdout sample one year before default. Deakin (1972) improved the efficiency of Altman's model by assessing an MDA model that considered 14 financial ratios. His MDA model had error rates for the holdout sample at 22%, 16%, 12%, 23%, and 15% for one to five years before the default, respectively. Blum (1974) employed the MDA model to forecast default and determined that his model could accurately predict 94% of bankruptcy cases one year before default. Dambolena and Khoury (1980) built MDA models and achieved prediction accuracy rates of 87%, 85%, and 78% for 1, 3, and 5 years before the default, respectively. In the analysis, they considered the stability of all financial ratios over time (measured by standard deviations) and the extent of these ratios as explanatory variables in the derivation of the MDA model.

2.3 Logistic models for default prediction

Since the early 1980s, researchers (Darayseh, Waples, & Tsoukalas, 2003; Lennox, 1999; Ohlson, 1980; Zavgren, 1985) have swapped their focus to the logit (logistic regression) model that has no constrictive assumptions for default forecast. Ohlson (1980) first assessed a logit model with nine independent variables and established that his model could accurately forecast over 92% of the bankrupt firms two years before default. Zavgren (1985) likewise employed the logit analysis for predicting default 1-5 years before the actual time of default of the companies. While the accuracy rate of his logit model for a one-year forecast was about the same as Ohlson's (1980) 92%, the error rates for more extended periods were like or marginally lower than those reported in the preceding default forecast models using MDA. Darayseh et al. (2003) built a logit model for default prediction using numerous economic variables in permutation with company-wise financial ratios. In their analysis, they compared 110 manufacturing companies that went bankrupt between 1990 and 1997 with 110 non-bankrupt companies based on total assets and industry classification. Their predicted model could make accurate predictions for 87.82% and 89.50% of the in-sample and holdout samples for one year before default. Some researchers (Collens and Green, 1982; Hamer, 1983; Lennox, 1999; Lo, 1986; Press & Wilson, 1978; Theodossiou, 1991) paralleled the usefulness of the popular statistical techniques used in default prediction. Their deduction of the analysis was diversified. Lo (1986) compared the accuracy of a logit model versus an MDA model in forecasting default. While the logit model was more robust than MDA in predictor variable estimation, both models gave out consistent results. In Theodossiou's (1991) study, he compared three statistical techniques the linear probability model, the logit model, and the probit model to identify the one with the most appealing performance in forecasting default in Greece. The outcome revealed that all three models were successful in forecasting default with precision rates over 90%.

Nevertheless, both logit and probit models were better than the linear probability model. Like Theodossiou's research (1991), Lennox's analysis (1999) studied the explanatory factors for failure for UK companies using three popular statistical techniques. He constructed an MDA, a logit, and a probit model to find default among the companies in the United Kingdom and compared the results of the three models in forecasting default. The estimations proved that the probit and logit models outperformed the discriminant model. Collens and Green (1982), Hamer (1983), Press and Wilson (1978) paralleled the performance of the logit model and the MDA model in forecasting default. The results showed that the predictability of the logit model is almost the same as that of MDA.

2.4 Merton model for default prediction

Unlike previous studies, the analysis by Hillegeist et al. (2004) identifies the likelihood of corporate defaults in the U.S. market using the probability of default from the Black-Scholes-Merton model. The research proves that the model gives notably more information than the two accounting-based bankruptcy models, namely Altman (1968) and Ohlson (1980). In contrast to the previous studies, the analysis uses relative information content tests to compare the out-of-sample performance of each default model. The sample includes 78,100 company-year observations and 756 initial bankruptcies between 1980-2000. The log-likelihood ratio tests suggest that the default probability predicted from the structural model contains more information that is important in predicting default than any of the accounting-based default models. Additionally, a comparison of each model Psuedo-R2 indicates that the structural model outperforms the original Altman and Ohlson models by 71% and 33%, respectively.

Therefore, we tried to estimate default using logistic and Merton models on the CRSP dataset between 1975 to 2007 to validate our prediction ability in specific conditions. The general forecastability of the models considers the influence of the elimination of variables in line with new trends.

3: Data

3.1 Sample selection

We included the dataset during the period 1975 to 2007[1]. We matched the information with firm-level accounting and price information obtained from COMPUSTAT- CRSP that belongs to the US market. We utilized the STATA software package for our research. We winsorized[2] the factors having extreme outliers at the one percentile and 99-percentile levels. Winsorization limits some extreme values in the data, thereby reducing the effect of possible spurious outliers.

We referred 1975 - 2007 corporate defaults (variable status) data from the Moody's Default Risk Services' Corporate Default database, SDC Platinum's Corporate Restructurings Database, Lynn M. LoPucki's Bankruptcy Research Database, and Shumway's (2001) list of defaults. In the data, we lagged the accounting variables by a quarter and the market variables by a month. Since the corporates announce their financials every quarter, we ensured the data in the prior quarter predicts the default by lagging the data.

3.2 Variable selection

We examined five factors (represented by Z1 to Z5) based on Edward I. Altman's paper published in July 2000 and seven factors (represented by O1 to O7) from James A. Ohlson's paper published in Spring 1980. We also included John Y. Campbell's factors (TLMTA, CASHMTA, NIMTAAV) and market-based variables used by CHS (2008) and

[1] Data used is same as per the research in Anginer and Yildizhan (2018)

[2] Factors that are winsorized include Z1,Z2,Z5,O2,O3,O4,O5,O6 and MERTON DD

Anginer and Yildizhan (2018) to predict defaults. The detailed variables description is as follows:

Z1: The working capital/total assets ratio is a measure of the liquidity, reveals whether the firm can pay for its short-term obligation or not. This ratio describes the relative proportion of net liquid assets to its total assets. The difference between a company's current assets and its current liabilities is called the working capital. Typically, a low or declining working capital to total asset ratio indicates the company may have too many current liabilities. Therefore, the firm must meet their losses by reducing their existing assets.

Z2: It is the retained earnings to total assets ratio. Retained earnings are the profits or losses that a firm can obtain after dividends or any other distribution to investors. The retained earnings to total assets ratio (Z2) can measure the extent of a company's leverage. Ordinarily, the companies with a low Z2 ratio financed capital through borrowings rather than retained earnings. Companies with a high Z2 ratio suggest positive profitability and the ability to endure losses.

Z3: Earnings before interest and tax (EBIT) to the total asset (Z3) is a ratio that measures the actual productivity of the firm's asset. It signals how effectively a company can use its assets to generate earnings. EBIT is the amount of profit independent of tax and leverage factors, and it is the net profit to focus on operating earnings when compared with other companies. Since the profitability of its assets determines the firm's ability to be in existence, this ratio is appropriate to deal with corporate failures. When the total liability exceeds the valuation of the asset, with the value determined by the efficiency of asset generating profit, the insolvency in a bankrupt may occur.

Z4: The market value of equity to book value of the total liabilities' ratio shows the capital structure of the firm. It indicates how much the firm's asset decrease can lead to a firm's insolvency (the total liabilities exceed the total assets). For the equity value, it is the market value of all shares of stock (preferred stock and common stock), and total liabilities include both short-term and long-term liabilities. The difference between the Z4 ratio to standard E/D ratio (equity and debt are book value) is that Z4 considers the market value dimension. The equity market value acts as a proxy for the firm's asset values. Our analysis used the log value of Z4 to scale down the value and prevent our model from being skewed due to large values.

Z5: Sales to total assets ratio is also called total assets turnover. This ratio illustrates a company's efficiency of managing the asset to generate revenue, and it would be better to generate income on as small a base of an asset as possible. The higher the ratio, the smaller the size of the investments required to obtain profit, and the higher the profitability of the company.

O1: It is the log of total assets to CPI. Total assets are as reported in dollars. The index year is of the prior year in the company's balance sheet date. The procedure assures a real-time implementation of the model.

O2: Total liabilities to total assets ratio is also called the debt ratio. This ratio indicates the debt-funded portion of the company's assets. The higher the O2 ratio, the higher the leverage a firm has, and the greater the risk will be associated with the operation of the company. Also, a high O2 ratio may illustrate that the firm's borrowing capacity is low. In turn, the low borrowing capacity will lower financial flexibility.

O3: Working Capital/Total Assets ratio is the same as the Z1 ratio. We did not exclude O3 here since we take the Ohlson variable as a group at one stage of our research.

O4: Current Liability to Current Asset ratio. This ratio (current ratio) is a liquidity ratio that measures the ability than a firm can pay for the short-term liability within one year. Current liabilities include account payable, wages payable, accrued tax, short-term debt, and related liabilities. Current assets listed on the accounting statements are cash, inventory, and account receivable or any other assets realized to cash within one year. If the O4 ratio is less than one, the company will have the capital to pay for its short-term obligations.

O5: Net income to total asset ratio is also called Return on Asset, and it explained how efficient the company could convert the asset it invests into earnings. The higher the O5 ratio, the better the company's performance, as the company can earn more money on fewer assets.

O6: Funds for Operations to Total Liability. Funds from operation (FFO) includes the money a company collects from its inventory sales and service it provides. FFO usually is calculated by Net Income + Depreciation + Amortization + Loss on sales of properties (or – Gains on sales of properties). The O6 ratio measures the ability that a firm can pay off its debt only using net operating income. The lower the O6 ratio, the more leverage the company has. If the O6 ratio is smaller than 1, this will indicate that the firm may need to sell some assets or finance additional debt to keep balance.

O7: (net income – last year's net income) / (|net income| + |last year's net income|). Net income here is the one for the most recent period. The denominator acts as a level indicator. The variable is thus intended to measure the change in net income.

TLMTA: Liabilities to Adjusted Total Assets. The value adjusted total assets means the total assets that are adjusted according to the market value to reflect the real fair market value of the assets. The adjusted total assets equal to the sum of the firm's Market Equity and the amount of the firm's Total Liability. The TLMTA ratio truly reflects the leverage in the firm according to market valuation, which performs better than the traditional book-value ratios.

CASHMTA: (Cash + Short-term Investment)/Adjusted Total Assets. This ratio is the ratio of the value of cash and short-term investments to the value of adjusted total assets. CASHMTA ratio helps investors capture the liquidity position of the firm. The higher the CASHMTA ratio is, the more the liquid assets the company will have in order to make interest payments. Thus, if the CASHMTA ratio is high, that means the company has enough time to respond to postpone or avoid bankruptcy.

NIMTAAV: It is a geometrically declining average of past quarterly values of the ratio of net income to adjusted total assets. The NIMTAAVG ratio is as follows:

$$NIMTAAVG_{t-1,t-12} = \frac{1-\phi^3}{1-\phi^{12}}(NIMTA_{t-1,t-3} + \cdots + \phi^9 NIMTA_{t-10,t-12})$$

where, $\phi = 2^{-\frac{1}{3}}$, which means the weight is halved quarterly. NIMTA is Net Income to Market-valued Total Assets:

$$NITMA = \frac{Net\ Income}{(Firm\ Market\ Equity + Total\ Liability)}$$

MB: Market-to-book ratio (MB) is also called price to book ratio. The ratio is the company's current market value divided by its book value. The market value is usually the current price of the firm's outstanding stock shares in the market. The book value is the amount to liquidating the value of the company's asset minus the cost of the liability.

EXTRETAVG: It is a geometrically declining average of monthly log excess stock returns on each firm's equity relative to the S&P 500 index. This ratio is as below:

$$EXRETAVG_{t-1,t-12} = \frac{1-\phi}{1-\phi^{12}}(EXRET_{t-1} + \cdots + \phi^{11}EXRET_{t-12})$$

Where, $\phi = 2^{-\frac{1}{3}}$, which means the weight is halved quarterly, and the EXRET is the monthly log excess returns relative to the S&P 500 index.

PRICE is the log price per share of each firm.

RSIZE is the ratio of the log ratio of the firm's market capitalization divided by the S&P 500 index. Market capitalization to the S&P index is the weight of each company in the index. The market capitalization is equal to the value of the current stock price multiplied by the company's outstanding shares.

MERTON DD: Merton's Distance-to-Default is the measure of the difference between the asset value of the firm and the face value of its debt, scaled by the standard deviation of the firm's asset value. We followed the Campbell, Hilscher, Szilagyi (2008), Hillegeist et al. (2004) and Anginer and Yildizhan (2018) to compute the Merton's distance-to-default method. The market value of equity is from the call option on the company's assets:

$$V_E = V_A e^{-dT} N(d_1) - X e^{-rT} N(d_2) + (1 - d^{-dT}) V_A$$

with $\quad d_1 = (\log\left(\frac{V_A}{X}\right) + \left(r - d + \frac{s_A^2}{x}\right)T)/(s_A\sqrt{T})$

and $\quad d_2 = d_1 - S_A\sqrt{T}$

where the $V_E$is the market value of firm equity, $V_A$is the value of the firm's asset, r is the risk-free rate, X is the face value of the debt which will mature at time T, d is the dividend rate, and s_A is the volatility of the value of the assets which is related to equity volatility(s_E).

$$s_E = (V_A e^{-dT} N(d_1) s_A)/V_E$$

After solving the above equations, we can find out the value of V_A and s_A. We obtain V_E from the market value of equity and short-term plus one-half long-term book debt to proxy for the face value of debt X. s_E is defined as the standard deviation of daily stock returns over the past three months, T is one year, and r equals to the one-year treasury bill rate. d, the dividend is the sum of the prior year's both common and preferred dividends. Besides, we use the Newton method to solve the two equations. The unknown variables are as follows:

$$V_A = V_E + X$$

$$s_A = s_E * V_E/(V_E + X)$$

Once asset values (V_A) is decided, the asst returns can be computed using the equation as in Hillegeist et al. (2004) mentioned:

$$m_t = \max[(V_{A,t} + d - V_{A,t-1})/(V_{A,t-1}, r]$$

As the expected returns cannot be negative, if the assets returns are less than zero, they set to the risk-free rate. Finally, the equation should be:

$$DD = \log\left(\frac{V_A}{X}\right) + \left(m - d - \frac{S_A^2}{2}\right)T/(S_A\sqrt{T})$$

4: Methodology

4.1 Logistic regression

Kwofie et al. (2015) used logistic regression to predict the probability of default for the loans issued by a Ghana-based microfinance company. We employed similar techniques to forecast default among corporations. Logistic regression analyses the relationship between multiple independent variables and a binary dependent variable. This model enabled the prediction of the desired result in our research “status” (default status) that has two possible variables – default/not default, where the two values are labeled “0” and “1”. The logit model solves the equation,

$$\mathrm{Y} = \ln\left(\frac{P}{1-P}\right) = a + bx + \varepsilon$$

The logistic distribution restricts the estimated probabilities to lie between 0 and 1.

$$Estimated\ probability\ (P) = \frac{e^{a+bx}}{1 + e^{a+bx}} = \frac{1}{1 + e^{-(a+bx)}}$$

Unlike linear regression models, logistic regression does not interpret the relationship between the dependent and independent variables. The Errors (ε) are generally not distributed because the dependent variable Y only has the outcomes 0 and 1. Additionally, the estimated probabilities P(Y) lies between 0 and 1.

The goal is to select the variables that can explain the data well. The null hypothesis behind the logistic regression is

$\boldsymbol{H_0}$: There is no relationship between dependent variables and independent variables

We need to reject the null hypothesis and show that the explanatory variables predict the default of the company.

We test the p-value at 99, 95, and 90 percentile levels. If $p < 0.01$, 0.05 or 0.1, reject the H_0 Hypothesis and the independent variables we test are significant at that specific confidence level. Other than p-value, we use McFadden pseudo R^2 to illustrate the proportion of the variation in the dependent variables that can be explained by the logistic regression model.

$$McFadden - R^2_{logistic} = 1 - \left(\frac{LL(a,b)}{LL\,(a)}\right) = 1 - \left(\frac{-2LL(a,b)}{-2LL(a)}\right)$$

For example, if the R^2 is 0.33, which means the logistic regression model consisted of independent variables can explain 33% of the dependent variable.

Petersen (2009) indicated that the fixed effect influences the relationship between the predictor variables and outcome variables. Fixed effects are variables that are consistent, which means any change they cause to the other variables is consistent. In our research, we considered the fixed effect of the financial year using the STATA function i. year. We removed the effect of economic factors (for example, global crisis; change in interest; GDP change; inflation), thereby assessing the net effect of the predictors on the outcome variable. We clustered the standard errors based on the companies using the STATA function cl(permno). We control the error term within each firm, which may be auto-correlated.

4.2 The marginal effect of the model

The research of Norton and Dowd (2018) published explained the impact of independent variables on dependent variables using margins. Instead of using odd ratio, we used the marginal (or incremental) effect to report the economic effect of each explanatory variable. For the given below logit model:

$$y_i = x_i \beta + \varepsilon_i$$

The marginal effect defined as the effect of a tiny change happening in a single continuous independent variable x_{1i} on the probability that $y_i = 1$

$$Incremental\ (Marginal\ effect) = \Pr(y_i = 1 |\ x_i = 1) - \Pr(y_i = 1\ |\ x_i, x_{1i} = 0)$$

For the logit model, the marginal effect of a continuous variable $x_{1i}$is

$$ME_i^{Logit} = \frac{\partial \Pr(y_i = 1|x_i)}{\partial x_{1i}} = \left(\frac{\beta_1}{\sigma}\right) * \ p_i * (1 - p_i)$$

σ is the standard deviation, ∂ is the percentage change of σ. The $\Pr(y_i = 1|x_i)$ cannot be evaluated as the distribution of ε is unknown. As an additional step, both ε and β by the standard deviation, σ. $\Pr(y_i = 1|x_i)$ transforms into a cumulative distribution function (CDF) of a standard logistic (logit) variable, which is easy to calculate for logistic distributions. We used MARGIN command in STATA to acquire the result of the average probability of an outcome (which is the default here).

4.3 Validation of the Model

4.3.1 Contingency tables

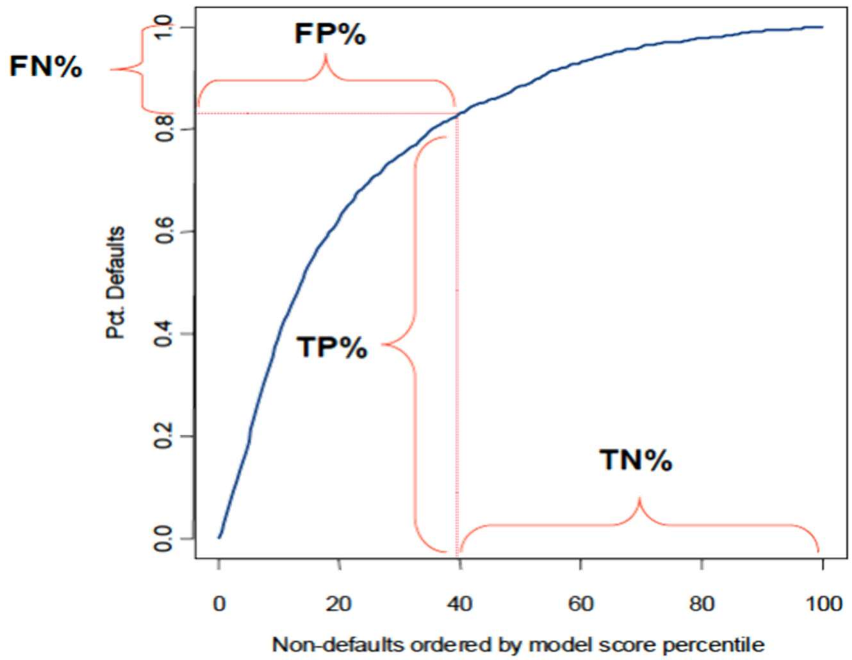
Stein (2002) validated the regression model in two dimensions: Calibration and Power. Calibration illustrates how well a model's projected probabilities match the actual default rates. Power describes how well the model can discriminate between Default and Non-default status. A good model has a higher percentage of default and a lower percentage of non-default in its “Bad” category and has a higher percentage of non-defaults and a smaller percentage of default in its “Good” category.

	Actual Default	Actual Non-Default
Bad	*TP*	*FP*
Good	*FN*	*TN*

As per the table above, the module produces only two ratings, Bad and Good. A True Positive (TP) is a predicted default that occurs. A True Negative (TN) is a predicted non-default that happens, which means the company does not default. The False positive (FP) is a predicted default that does not happen, and the False Negative (FN) is a predicted non-default, but the company defaults. So, the errors of the model are the FN and the FP, which listed in the cells above. A “perfect” model should have zero FN and FP, and the total number of defaults and non-defaults should be the total number of TP and TN. In that way, it can indicate that the model can correctly discriminate between the default companies and non-default companies.

4.3.2 ROC curve

Figure 1: *Schematic of a ROC*

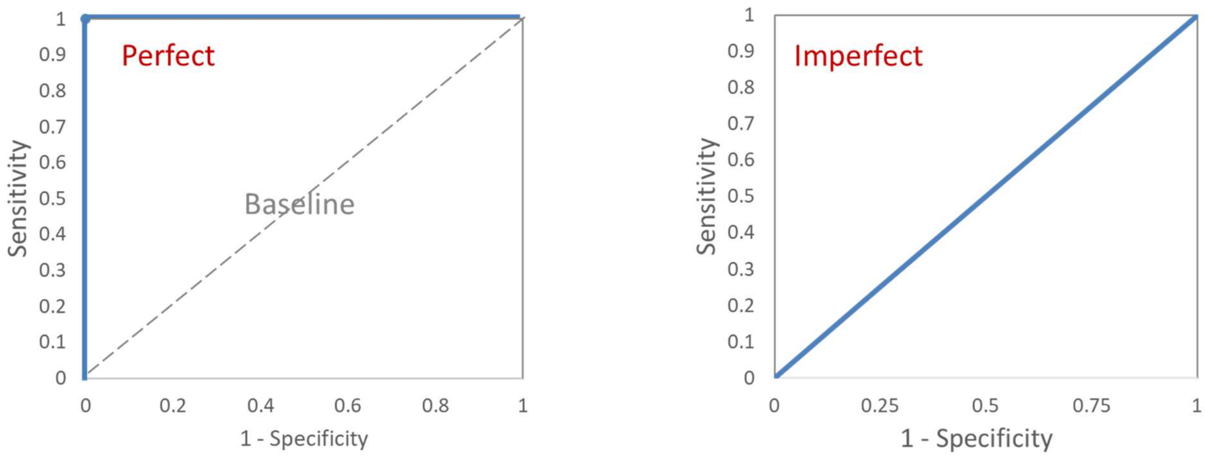


ROC (relative or receiver operating characteristic) is used to score all credits. The ROC plot will order the non-default companies from worst to best on the x-axis and plot the percentage of defaults excluded on the y-axis. Then, the graph associates the y-axis with every score on the x-axis as a cumulative percentage of defaults with a score equal to or worse than that score in the test data. In other words, the y-axis gives the percentage of defaults excluded as a function of the number of non-defaults excluded. The ROC evaluated by specificity and sensitivity. Sensitivity is a performance measure of the entire positive part, whereas specificity is a performance measure of the whole negative part of a dataset. The graph uses 1 – specificity on the x-axis and sensitivity on the y-axis.

$$\text{Sensitivity } (True\ Positive\ Rate) = \frac{TP}{TP+FN}$$

$$\text{Specificity } (True\ Negative\ Rate) = \frac{TN}{TN+F}$$

Figure 2: *Comparison of ROC for Perfect Vs. Imperfect Models*



5: Empirical Findings

5.1 Default prediction using Altman variables

Table 1: *Logistic Regression using Altman variables with and without winsorized Merton DD*

Table 1 reports result from logistic regressions of the default indicator on the predictor variables without (1) and with (2) winsorized Merton DD. The data used as per the variables described in detail in the Data section. The values in the parenthesis below the coefficient estimates show the absolute values of z-statistics. Each regression shows the McFadden pseudo R2 values. The 10%, 5%, and 1% statistical significance is represented by *, **, and ***, respectively.

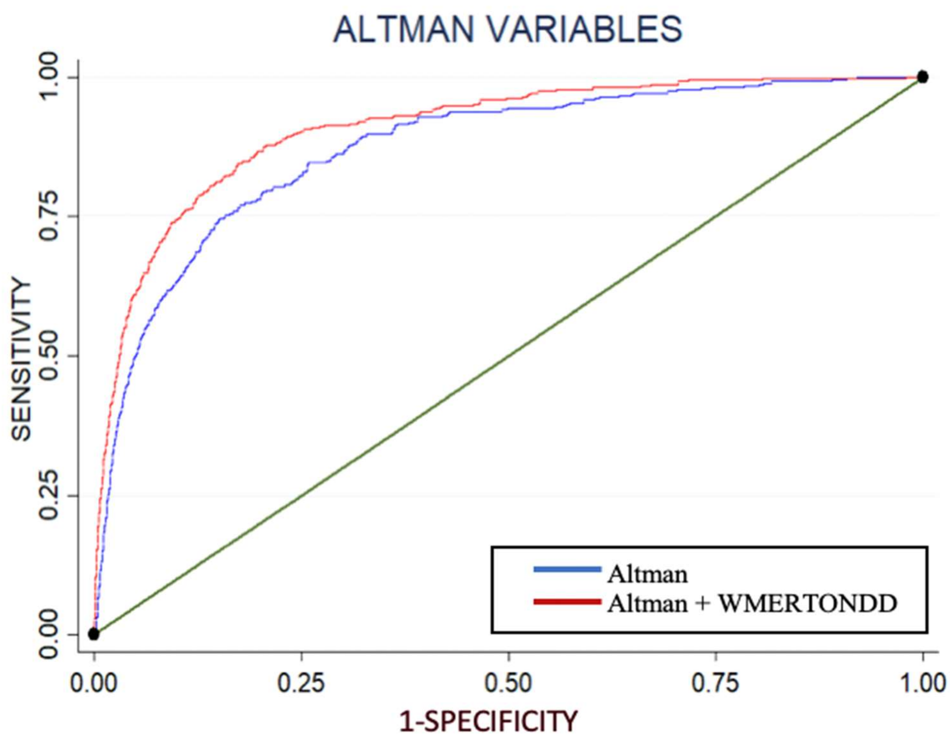
VARIABLES[3]	(1) 1975-2007	(2) 1975-2007
WZ1	1.006***	0.387*
	(0.247)	(0.235)
WZ2	-1.119***	-0.451**
	(0.153)	(0.180)
Z3	-9.508***	-8.431***
	(1.298)	(1.028)
WZ5	-4.075***	-4.453***
	(1.139)	(1.218)
LOGZ4	-0.638***	-0.585***
	(0.049)	(0.046)
WMERTONDD		-0.527***
		(0.046)
CONSTANT	-6.552***	-4.381***
	(0.489)	(0.489)
Observations	75,667	75,667
Pseudo R^2	0.161	0.250

Based on the table above, we observe that R^2 is higher for the model with winsorized Merton DD, and Altman variables Z3, Z4, and Z5 have significance in both the model. The model validation was done by comparing the ROC (as shown in figure 2), which reflects the same as the table as the area under the curve for the model with and without winsorized

[3] WZ1, WZ2, WZ5 and WMERTONDD are winsorized variables

Merton DD is 0.9086 and 0.8723. The results demonstrate that the model containing winsorized Merton DD has better accuracy in predicting default.

Figure 3: *ROC Comparison for Altman variables with and without winsorized Merton DD*



5.2 Default prediction using Ohlson variables

Table 2: *Logistic Regression using Ohlson variables with and without winsorized Merton DD*

Table 2 reports result from logistic regressions of the default indicator on the predictor variables without (1) and with (2) winsorized Merton DD. The data used as per the variables described in detail in the Data section. The values in the parenthesis below the coefficient estimates show the absolute values of z-statistics. Each regression shows the McFadden pseudo R2 values. The 10%, 5%, and 1% statistical significance is represented by *, **, and ***, respectively.

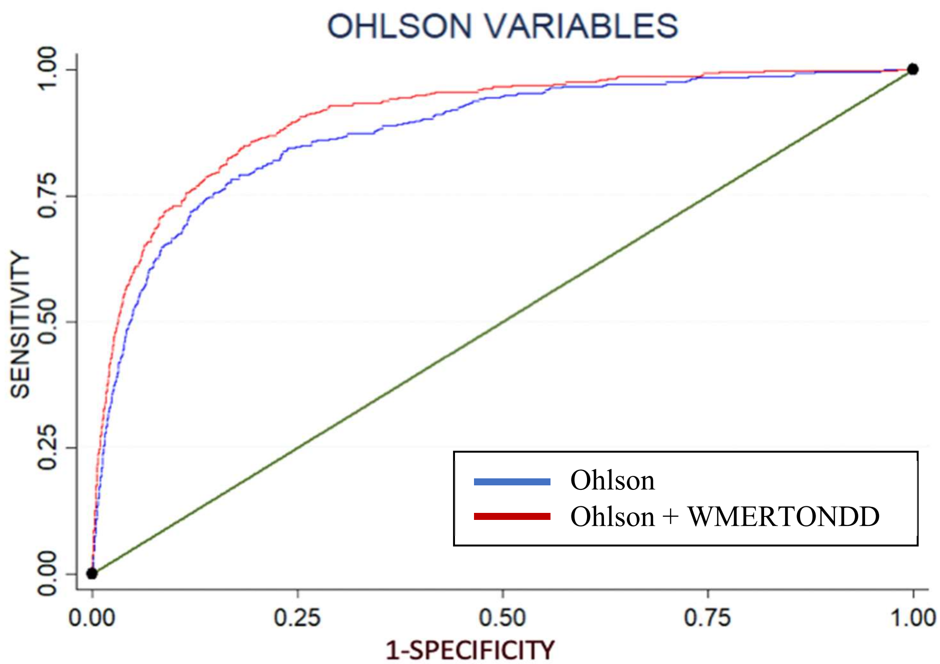
VARIABLES[4]	(1) 1975-2007	(2) 1975-2007
O1	-0.130***	0.162***
	(0.027)	(0.039)
WO2	13.167***	6.492***
	(0.926)	(0.968)
WO3	0.999***	0.887***
	(0.241)	(0.261)
WO4	0.503***	0.425***
	(0.076)	(0.077)
WO5	-16.131***	-10.138***
	(2.627)	(3.002)
WO6	-0.623	-0.304
	(0.535)	(0.597)
O7	0.077	0.011
	(0.090)	(0.086)
WMERTONDD		-0.572***
		(0.060)
CONSTANT	-10.638***	-7.562***
	(0.614)	(0.617)
Observations	75,667	75,667
Pseudo R^2	0.192	0.245

Based on the table above, we observe that R^2 is higher for the model with winsorized Merton DD, and Ohlson variables O1, O2, O3, O4, and O5 have significance in both the model. The model validation was done by comparing the ROC (as shown in figure 3), which reflects the same as the table as the area under the curve for the model with and

[4] WO2, WO3, WO4, WO5, WO6 and WMERTONDD are winsorized variables

without winsorized Merton DD is 0.9075 and 0.8766. The results demonstrate that the model containing winsorized Merton DD has better accuracy in predicting default.

Figure 4: *ROC Comparison for Ohlson variables with and without winsorized Merton DD*



5.3 Default prediction using Campbell variables

Table 3: *Logistic Regression using Campbell variables with and without winsorized Merton DD*

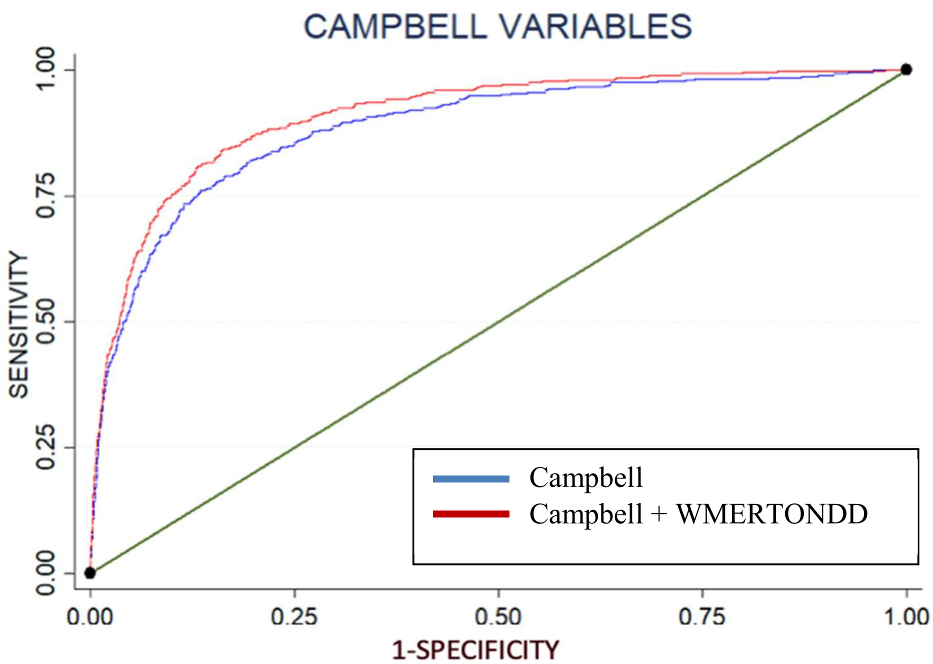
Table 3 reports result from logistic regressions of the default indicator on the predictor variables without (1) and with (2) winsorized Merton DD. The data used as per the variables described in detail in the Data section. The values in the parenthesis below the coefficient estimates show the absolute values of z-statistics. Each regression shows the McFadden pseudo R2 values. The 10%, 5%, and 1% statistical significance is represented by *, **, and ***, respectively.

VARIABLES[5]	(1) 1975-2007	(2) 1975-2007
NIMTAVG	-45.574***	-29.252***
	(2.597)	(2.935)
CASHMTA	-2.315***	-1.684***
	(0.648)	(0.636)
TLMTA	5.380***	3.897***
	(0.313)	(0.331)
WMERTONDD		-0.412***
		(0.047)
CONSTANT	-9.049***	-6.681***
	(0.512)	(0.553)
Observations	75,667	75,667
Pseudo R^2	0.220	0.256

Based on the table above, we observe that R^2 is higher for the model with winsorized Merton DD, and all Campbell accounting variables have significance in both the model. The model validation was done by comparing the ROC (as shown in figure 4), which reflects the same as the table as the area under the curve for the model with and without winsorized Merton DD is 0.9104 and 0.8860. The results demonstrate that the model containing winsorized Merton DD has better accuracy in predicting default.

[5] WMERTONDD is winsorized MERTONDD variable

Figure 5: *ROC Comparison for Campbell variables with and without winsorized Merton DD*



5.4 Default prediction using Market variables

Table 4: *Logistic Regression using Market variables with and without winsorized Merton DD*

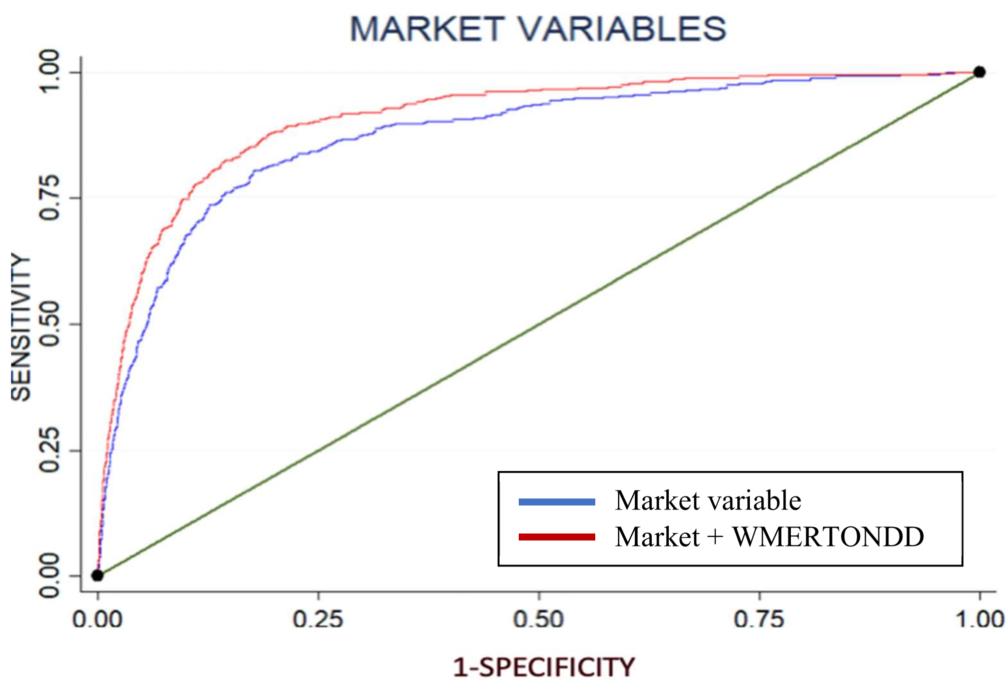
Table 4 reports result from logistic regressions of the default indicator on the predictor variables without (1) and with (2) winsorized Merton DD. The data used as per the variables described in detail in the Data section. The values in the parenthesis below the coefficient estimates show the absolute values of z-statistics. Each regression shows the McFadden pseudo R2 values. The 10%, 5%, and 1% statistical significance is represented by *, **, and ***, respectively.

VARIABLES[6]	(1) 1975-2007	(2) 1975-2007
MB	-0.267***	-0.134**
	(0.064)	(0.055)
PRICE	-1.177***	-0.556***
	(0.124)	(0.119)
RSIZ	0.209***	0.311***
	(0.048)	(0.050)
EXRETAVG	-19.853***	-15.182***
	(1.975)	(1.600)
WMERTONDD		-0.571***
		(0.059)
CONSTANT	-2.111**	-0.018
	(0.864)	(0.879)
Observations	75,667	75,667
Pseudo R^2	0.193	0.247

Based on the table above, we observe that R^2 is higher for the model with winsorized Merton DD, and all market variables have significance in both the model. The model validation was done by comparing the ROC (as shown in figure 5), which reflects the same as the table as the area under the curve for the model with and without winsorized Merton DD is 0.9097 and 0.8739. The results demonstrate that the model containing winsorized Merton DD has better accuracy in predicting default.

[6] WMERTONDD is winsorized MERTONDD variable

Figure 6: *ROC Comparison for Market variables with and without winsorized Merton DD*



5.5 Consolidated model for default prediction

Table 5: *Logistic Regression using all the variables with and without winsorized Merton DD*

Table 5 reports result from logistic regressions of the default indicator on the predictor variables without (1) and with (2) winsorized Merton DD. The data used as per the variables described in detail in the Data section. The values in the parenthesis below the coefficient estimates show the absolute values of z-statistics. Each regression shows the McFadden pseudo R2 values. The 10%, 5%, and 1% statistical significance is represented by *, **, and ***, respectively.

VARIABLES[7]	(1) 1975-2007	(2) 1975-2007
O1	0.180	0.307**
	(0.131)	(0.148)
WO2	-12.883***	-16.580***
	(3.237)	(3.527)
WO4	0.265***	0.260***
	(0.086)	(0.085)
WO5	3.362	3.490
	(3.917)	(4.123)
WO6	1.289**	1.645**
	(0.658)	(0.811)
O7	0.038	0.033
	(0.093)	(0.092)
WZ1	1.158***	1.022***
	(0.303)	(0.312)
WZ2	-0.155	-0.091
	(0.253)	(0.268)
Z3	-8.651***	-9.807***
	(1.846)	(2.062)
WZ5	-2.560*	-2.542*
	(1.424)	(1.414)
LOGZ4	0.175	0.074
	(0.132)	(0.170)
NIMTAVG	-24.544***	-21.010***
	(4.690)	(4.699)
CASHMTA	-2.912***	-2.750***
	(0.726)	(0.727)
TLMTA	10.034***	9.380***
	(1.634)	(1.865)
MB	0.075	0.046
	(0.052)	(0.052)
PRICE	-0.599***	-0.287**
	(0.131)	(0.129)
RSIZ	-0.079	-0.150
	(0.148)	(0.164)
EXRETAVG	-16.181***	-14.023***
	(2.013)	(1.771)
WMERTONDD		-0.331***
		(0.056)
CONSTANT	-8.792***	-7.717***
	(2.375)	(2.651)
Observations	75,667	75,667
Pseudo R^2	0.284	0.300

[7] WZ1, WZ2, WZ5, WO2, WO4, WO5, WO6 and WMERTONDD are winsorized variables

Based on the table above, we observe that R^2 is higher for the model with winsorized Merton DD, and several variables have significance in both the model. The model validation was done by comparing the ROC (as shown in figure 6), which reflects the same as the table as the area under the curve for the model with and without winsorized Merton DD is 0.9253 and 0.9160. The results demonstrate that the model containing winsorized Merton DD has better accuracy in predicting default.

Figure 7: *ROC Comparison for all variables with and without winsorized Merton DD*

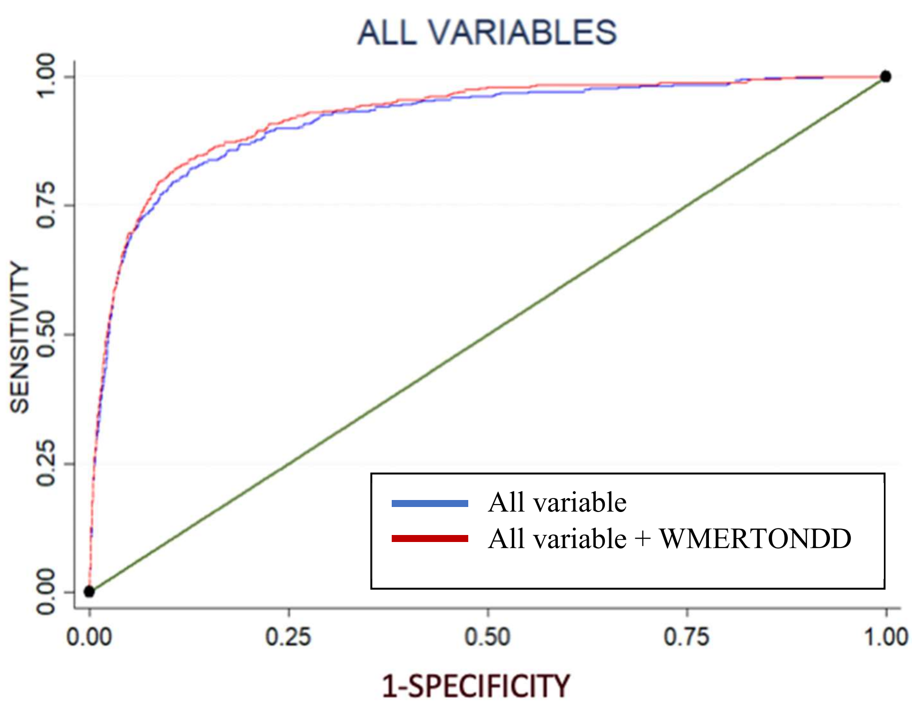


Figure 8: *ROC Comparison across models with winsorized Merton DD*

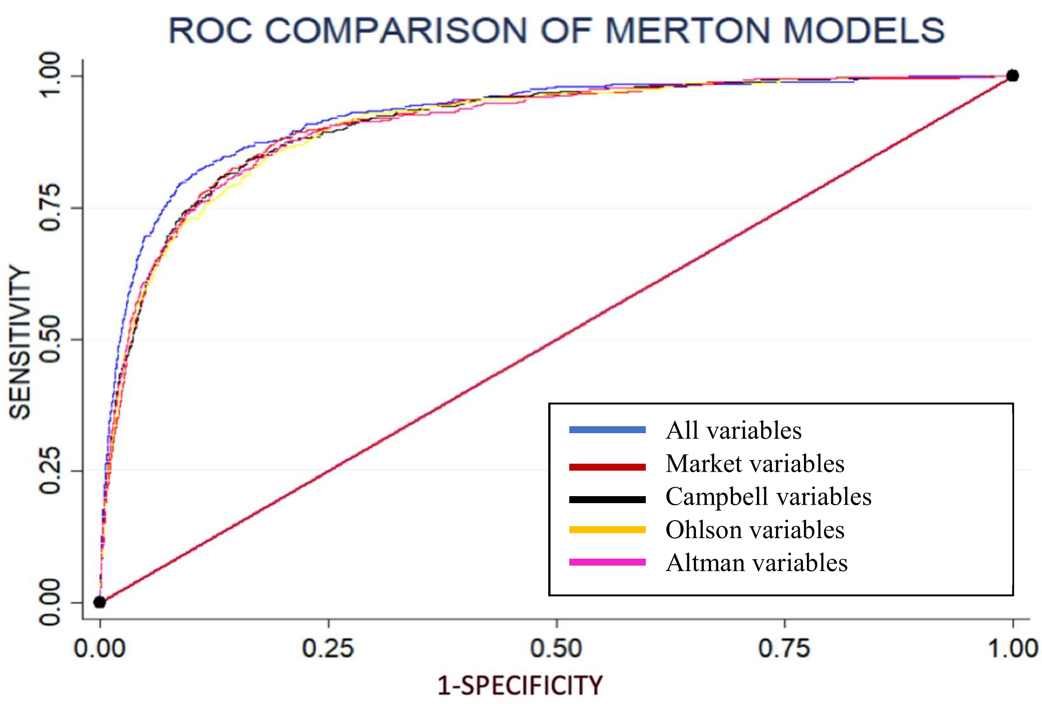


Figure 7 shows the comparison of ROC from all the models with winsorized Merton DD. Since the models with winsorized Merton DD have better forecasting and accuracy, we studied which among the five models have better accuracy. The model containing all the variables had the highest R^2 and AUC, which was used for the study of the marginal effect of the predictor variables.

5.6 Marginal Effect

Table 6: *Marginal Effect at Median and 75 Percentile of the Independent Variables*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	WZ2	Z3	WZ5	LOGZ4	O1	WO2	WO3	WO4	WO5	WO6
At Median	0.00514***	0.00457***	0.00581***	0.00538***	0.00579***	0.0351***	0.00507***	0.00479***	0.00555***	0.00600***
	(0.000314)	(0.000250)	(0.00151)	(0.000282)	(0.000658)	(0.0113)	(0.000251)	(0.000271)	(0.000511)	(0.000527)
At 75 percentile	0.00510***	0.00432***	0.00617**	0.00479***	0.00884***	0.0130***	0.00571***	0.00510***	0.00563***	0.00636***
	(0.000389)	(0.000262)	(0.00242)	(0.000325)	(0.00268)	(0.00281)	(0.000320)	(0.000253)	(0.000593)	(0.000711)
Difference	(0.00004)	(0.00025)	0.00036	(0.00059)	0.00305	(0.02210)	0.00064	0.00031	0.00008	0.00036
Change Ratio	-0.78%	-5.47%	6.20%	-10.97%	52.68%	-62.96%	12.62%	6.47%	1.44%	6.00%
Observations	75,667	75,667	75,667	75,667	75,667	75,667	75,667	75,667	75,667	75,667

	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
VARIABLES	O7	NIMTAVG	CASHMTA	TLMTA	MB	PRICE	RSIZ	EXRETAVG	WMERTONDD
At Median	0.00523***	0.00366***	0.00566***	0.00423**	0.00529***	0.00374***	0.00444***	0.00279***	0.00220***
	(0.000259)	(0.000330)	(0.000302)	(0.00198)	(0.000272)	(0.000587)	(0.000703)	(0.000264)	(0.000350)
At 75 percentile	0.00527***	0.00331***	0.00484***	0.0221	0.00553***	0.00363***	0.00361***	0.00199***	0.000711***
	(0.000313)	(0.000364)	(0.000256)	(0.0147)	(0.000465)	(0.000617)	(0.00139)	(0.000263)	(0.000246)
Difference	0.000040	(0.000350)	(0.000820)	0.017870	0.000240	(0.000110)	(0.000830)	(0.000800)	(0.001489)
Change Ratio	0.76%	-9.56%	-14.49%	422.46%	4.54%	-2.94%	-18.69%	-28.67%	-67.68%
Observations	75,667	75,667	75,667	75,667	75,667	75,667	75,667	75,667	75,667

Table-6 represents the marginal effect of the independent variables on the dependent variables. The 10%, 5%, and 1% statistical significance is represented by *, **, and ***, respectively. The coefficient estimates at median and 75-percentile are the average predicted probabilities when the independent variables are at given two levels. The values in the parenthesis below the coefficient estimates show the absolute values of z-statistics. The change ratio represents the difference when the estimates shift from the median to the 75-percentile value. We did not include Z1 because it is the same as WO3.

From the change ratio, we observe that the LOGZ4, O1, WO2, WO3, CASHMTA, TLMTA, RSIZ, EXRETAVG, and winsorized Merton DD have a significant impact on the default indicator. For example, when winsorized Merton DD marginal coefficients change from the median level to the 75- percentile level, the default probability will decrease by 67.68%. Whereas in the case of TLMTA, we observe that the default probability increases by 400 times when the marginal effect shifts from the median to 75-percentile.

6: Conclusion

In our study, we predicted the corporate defaults by taking a sample of 75,667 corporate data between the period of 1975 to 2007 using logistic regression and Merton's distance to default. Previous literature had forecasted default with different company-specific financial and macroeconomic variables using MDA and logistic regression. In the case of the Merton model, distance to default variable combines crucial market variables to predict the probability of default. Distance to default is calculated as the difference between the asset value of the firm and the face value of its debt, scaled by the standard deviation of the firm's asset value. We studied different models to determine the importance of the Merton distance-to-default in default prediction after controlling the accounting variables used in Ohlson (1980), Altman (1968) and set of market variables used in Campbell et al. (2008).

We observed that the Merton distance to default and TLMTA (leverage to the market value of assets) have considerable marginal effect compared to other variables. The effect is significant because when the distance to default decreases or when the leverage increases, there is an increase in credit risk, which leads to increased bankruptcy. The regression containing both accounting and Merton variables had the accuracy improved when compared without the Merton distance to default, which is supported empirically and by ROC comparison. The scope of this study can be improved by considering data beyond the period 2007, which might lower the predictability. Nevertheless, the loss is expected to be minimal since the model includes variables for macroeconomic effects over time.

Bibliography

Works Cited

Alaminos D, del Castillo A. and Fernández MÁ. (2016). A Global Model for Bankruptcy Prediction, PLoS ONE 11, p.e0166693

Altman, Edward I. (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, Journal of Finance, 23, 589–609.

Altman, Edward I. (2000) Predicting Financial Distress of Companies: Revising The Z-score and ZETA® Models 2013, 428–456

Anginer, Deniz & Yıldızhan, Çelim (2018), Is There a Distress Risk Anomaly? Pricing of Systematic Default Risk in the Cross Section of Equity Returns, Review of Finance, 22, 633–660

Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. *The British Accounting Review*, 38(1), 63-93.

Beaver, W. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 5: 71-111.

Bellovary, J. L., Giacomino, D. E., & Akers, M. D. (2007). A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial Education*, 33, 1-42.

Bharath S., and Shumway T. (2008) Forecasting default with the KMV Merton model, *Review of Financial Studies*, 21, 1339–1369.

Blum, M. (1974). Failing company discriminant analysis. *Journal of Accounting Research*, 12(10), 1-25.

Campbell J. Y., Hilscher J., and Szilagyi J. (2008). In search of distress risk. *Journal of Finance*, 63, 2899–2939.

Collens, R. A., & Green, R. D. (1982). Statistical methods for bankruptcy forecasting. *Journal of Economics and Business*, 34(4), 349-354.

Dambolena, I. G., & Khoury, S. J. (1980). Ratio stability and corporate failure. *Journal of Finance*, 35(4), 1017-1026.

Daniel, Brindescu-Olariu & Ionut, Golet (2013) Prediction of Corporate Bankruptcy in Romania through the use of Logistic Regression, Annals of Faculty of Economics, 1, 976-986

Darayseh, M., Waples, E., & Tsoukalas, D. (2003). Corporate failure for manufacturing industries using firm specifics and economic environment with logit analysis. *Managerial Finance*, 29(8), 23-37.

Deakin, E. B. (1972). A discriminant analysis of predictors of business failure. *Journal of Accounting Research*, 10(2), 167-179.

Fitzpatrick, P. (1932). A comparison of ratios of successful industrial enterprises with those of failed companies. The Certified Public Accountant (October, November, December): 598-605, 656-662, and 727-731, respectively.

Hamer, M. (1983). Failure prediction: Sensitivity of classification accuracy to alternative statistical methods and variable sets. *Journal of Accounting and Public Policy*, 2(4), 287-307.

Hillegeist, S. A., E. Keating, D. Cram, and K. Lundstedt (2004). Assessing the Probability of Bankruptcy, *Review of Accounting Studies,* 9: 5-34.

Horta, R. A. M., Alves, F. J. S., & Carvalho, F. A. A. (2013). Seleção de atributos na previsão de insolvência: aplicação e avaliação usando dados brasileiros recentes. *Revista de Administração Mackenzie*, 15(1), 125-151.

Horta, R. A. M., Borges, C. C. H., Carvalho, F. A. A., & Alves, F. J. S. (2011). Previsão de insolvência: uma estratégia para balanceamento da base de dados utilizando variáveis contábeis de empresas brasileiras. *Sociedade, Contabilidade e Gestão*, 6(2), 21-36.

Kim, Hyunjoon & Gu, Zheng (2009) Financial features of dividend-paying firms in the hospitality industry: A logistic regression analysis, International Journal of Hospitality Management, 28, 359-366

Korol, T., & Korodi, A. (2010). Predicting bankruptcy with the use of macroeconomic variables. *Economic Computation and Economic Cybernetics Studies and Research*, 44(1), 201-221.

Kwofie, Charles, Owusu-Ansah, Caleb, & Boadi, Caleb. (2015). *Research Journal of Mathematics and Statistics*, 7, 46-52

Lennox, C. (1999). Identifying failing companies: A re-evaluation of the logit, probit, and DA approaches. *Journal of Economics and Business*, 51(4), 347-364.

Lo, A. W. (1986). Logit versus discriminant analysis: A specification test and application to corporate bankruptcies. *Journal of Econometrics*, 31(2), 151-178.

Merwin, C. (1942). Financing small corporations in five manufacturing industries. New York: National Bureau of Economic Research.

Merton R. C. (1974). On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance*, 29, 449–470.

Norton, Edward C. and Dowd., Bryan E. (2018). Log Odds and the Interpretation of Logit Models. *Health services research*, 53, 859-878

Norton, Edward C., Dowd, Bryan E., & Maciejewski., Matthew L. (2019). Marginal Effects—Quantifying the Effect of Changes in Risk Factors in Logistic Regression Models, JAMA 321, 1304-1305

Ohlson, James A. (1908) Financial Ratios and the Probabilistic Prediction of Bankruptcy, Journal of Accounting Research, 18, 109-131

Petersen., Mitchell A (2009). Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches. *The Review of Financial Studies*, 22, 435-480

Pinheiro, L. E. T., Santos, C. P., Colauto, R. D., & Pinheiro, J.L. (2009). Validação de modelos brasileiros de previsão de insolvência. Contabilidade Vista & Revista, 18(4), 83-103.

Press, J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73(364), 699-705.

Shumway T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *Journal of Business* 74, 101–124.

Smith, R., Winakor, A. (1935). Changes in Financial Structure of Unsuccessful Industrial Corporations. *Bureau of Business Research*, Bulletin No. 51. Urbana: University of Illinois Press.

Stein, R. M. (2002). Benchmarking default prediction models: Pitfalls and remedies in model validation. *Moody's KMV*, New York, 20305.

Theodossiou, P. (1991). Alternative models for assessing the financial condition of business in Greece. *Journal of Business and Accounting*, 18(5), 697-720.

Zavgren, C. V. (1985). Assessing the vulnerability to failure of American industrial firms. *Journal of Business and Accounting*, 12(1), 19-45.