ANL 488 FINAL PROJECT REPORT

Prediction of Corporate Bankruptcy with Decision Trees

Submitted by

HO ZHONG TA BENJAMIN

SCHOOL OF BUSINESS

Singapore University of

Social Sciences

Presented to Singapore University of Social Sciences in partial fulfilment of the requirements for the Degree of Bachelor of Science in Business Analytics

2021

Abstract

The spike in corporate bankruptcy has been a huge concern to creditors, lenders, and shareholders across the globe, making it imperative to predict corporate bankruptcy early. The existing predictive solutions revolves the use of statistical models like Discriminant Analysis and advanced predictive models like Neural Network. The issue with statistical models is that they are not as robust and requires multiple assumptions, while advance predictive models are known to be not explainable.

Using financial ratios from the global energy sector between 2011 and 2018, this study aimed to address the issues of existing methodologies of corporate bankruptcy and propose the use of Decision Trees as viable alternative as they perform better than statistical models and more explainable than neural network, overcoming issues faced by either model.

Using a mixture of recall, AUC and lift chart as evaluation methods, the performance of four different types of decision tree models were compared and the champion model (C5.0) was compared with a logistic regression model and a neural network in terms of performance and explainability. The decision tree model outperformed both models with an outstanding recall rate of 93.83%, AUC of 0.933 and higher lift values throughout dataset. The decision tree has proven to be more explainable than neural network and more intuitive than logistic regression. The model suggests that financial ratios such as Earning per Share, Operating Cash Flow per Share, Debt Ratio and Cash Flow from Operations over Sales are important predictors of corporate bankruptcy.

The C5.0 DT has proven its performance and explainability against both models and intuitiveness of use and thus proven to be a viable alternative to existing corporate bankruptcy prediction model.

Table of Contents

1. Introduction

The significant spike in corporate bankruptcy has become increasing worrying for creditors, lenders, and shareholders alike. Corporate bankruptcy occurs when a judge decides that a company is unable to repay the debt to debtors as it falls due (Levratto, 2013). It is a legal way of resolving financial distress by firms who are financially insolvent and usually in default (Agrawal & Maheshwari, 2018). For creditors, they will not be able to recover their loans or bonds, which will drive up the interest rates to future lenders in the similar industry, potentially affecting their future business plans (Financial Times, 2020). As shareholders are the last to be paid in the case of bankruptcy, they are potentially losing their investments if their company declares bankruptcy (U.S. Securities and Exchange Commision, 2009).

Other than affecting the lives of firms and individuals alike, corporate bankruptcy has devastating consequences on the economy, which is evident from the bankruptcy of Lehman Brothers which kickstarted the global financial crisis of 2008 (Wiggins, Piontek, & Metrick, 2014). The impact of multiple corporate bankruptcy will cause rising unemployment rates due to the drop in job supplies and workers being laid off from their work (Forbes, 2020), and it will also lead to the increase in bank interest rates due to the drop in banks' profitability and capital (Forbes, 2020). The numerous negative consequences of corporate bankruptcy make it essential to prevent its occurrence. Thus, it is imperative that any sign of corporate distress should be predicted immediately.

Several statistical methods, such as Discriminant Analysis, and predictive models, like Neural Network, are used to predict corporate bankruptcy. The issue faced with statistical methods is the assumptions required, such as the assumption of normality, which might not be representative of the dataset used (Eisenbeis, 1977) and that it has lesser predictive capability than machine learning models. While on the other end, machine learning methods such as Neural Network are often considered "Black Box" which provides an approximation without

any further insights on the formulae and functions used to reach the approximation (Bathaee, 2018). Financial ratios are used as predictive variables for these models to predict corporate bankruptcy.

To provide early warning on companies in corporate distress, it is essential to find out the red flags that the companies are showing prior to bankruptcy. This requires a predictive model which is more explainable and yet have more predictive capability than the conventional statistical models. Thus, this paper proposes the use of decision trees models for the prediction of corporate bankruptcy. The decision tree model can adapt to non-linear data, does not require normalisation and is not a black box as it able to produce a decision tree which is able to explain the functions and the reasons for the approximation.

The study will be done using financial ratios of companies in the energy sector. The rationale behind the choice of industry is due to the alarmingly high rates of corporate bankruptcy found in the energy sector. Based on research done by the S&P global ratings, energy sector has been the industry with the highest global default rates since 2014 (S&P Global Ratings, 2020). The energy sector consists of three main types of companies that makes up the industry's supply chain, namely the Upstream, Midstream and Downstream companies (Saad, Udin, & Hasnan, 2014). The upstream companies deal with the exploration and production of oil and gas. The midstream companies deal with the transportation and storage of oil and gas products. Lastly, the downstream companies deal with the refining and procession of crude oil and gas into usable products, like fuel and chemicals. While upstream companies' profitability and liquidity heavily rely on the oil prices, the downstream companies rely on the demand of oil and gas products. However, the entire sector, regardless of its type, has been racking up corporate defaults due to oversupply from the upstream companies (Wood Mackenzie, 2020) and drop in demand of fuels (Forbes, 2020). Given the high risk of this industry, there is a need to provide

early warning on this sector to mitigate the risk of corporate bankruptcy. Thus, the energy sector is chosen as a focal point to the study.

This paper reviews and addresses issues with existing methods of corporate bankruptcy prediction. Balancing between predictive capability and model explainability, this research proposes the use of decision tree models to predict corporate bankruptcy as opposed to existing predictive methods. Financial data on companies in the energy sector will be used for this study.

The remainder of this paper is structured as follows. The next section entails an extensive review of literature on predictive modelling methodologies. The third section provides more insights on the data to be used, the data preparation and the data sampling stage. The fourth section will entail the prediction of bankruptcy using different decision tree models, and the champion model will be compared with other predictive methods to determine its performance and explainability. Lastly, the paper is concluded and discusses areas of further research in the last section.

2. Literature Review

2.1 Usage of Financial Ratios as Predictors of Bankruptcy

Initial studies on bankruptcy prediction had a heavy focus on the use of financial ratios found from the companies' financial statement. Fitzpatrick (1932) compared financial ratios of failed and successful firms and concluded that successful companies tend to have more favourable financial ratios compared to those who failed. Smith and Winakor (1935) confirmed Fitzpatrick's observations and added that the financial structure as shown by the Current Assets to Total Assets ratio dropped as firm approaches bankruptcy. These studies proved that financial ratios are significant predictors of bankruptcy and a firm's financial health.

2.2 Statistical Models for Corporate Bankruptcy Prediction

These studies on ratio analysis are considered univariate, which are models with a single variable. These studies focus on studying a single variable's impact on the bankruptcy prediction. The most important study on univariate model is that of Beaver's (1966) which he tested each ratio's individual predictive ability in classifying bankrupt and non-bankrupt companies. He further touted the idea of using multiple ratios simultaneously, which set the groundwork for multivariate models.

After Beaver's study, Altman (1968) proposed the first multivariate bankruptcy prediction model using Multivariate Discriminant Analysis (MDA). Based on the firm's financial ratios, the model classifies the companies into two different groups, either bankrupt or non-bankrupt. Altman used a "Z-score" which predicted possible bankruptcy if the firm falls within a certain score range. The accuracy of the model was remarkable, thus making MDA an important multivariate model for bankruptcy prediction.

However, Ohlson (1980) pointed out several cons of using the MDA. Firstly, statistical requirement such as the assumptions of normality goes against the use of dummy independent variables. Secondly, the output of MDA is a score of an ordinal ranking, which provides for little interpretation. Ohlson proposed the use of a conditional logistic regression model, otherwise known as logit model, which was introduced by Martin (1977). Logit models are used to predict the probability of binary outcomes, which in this case is between bankrupt or not bankrupt. This model considers the probability of bankruptcy, and according to Mihalovič (2016) provides better predictive capability than MDA.

2.3 Machine Learning Models for Corporate Bankruptcy Prediction

Models mentioned prior were generally statistical models, which have multiple assumptions such as the assumption of normality for MDA and linearity for continuous variables in logit model. These restrictions, along with the increased predictive capability (Barboza, Kimura, & Altman, 2017), made machine learning models the primary method used in predictive studies post-statistical models (Bellovary, Giacomino, & Akers, 2007).

2.3.1 Neural Network

The first of these models used in the study of corporate bankruptcy is the Neural Network (NN). NN is emulates human learning in the form of human pattern recognition function (Anandarajan, Lee, & Anandarajan., 2004). Messier and Hansen (1988) proposed a Neural Network model which managed to achieve 100% accuracy in predicting bankruptcy. El-temtamy (1995) compared the performance between neural network models and logit models and discovered that neural network outperformed logit models in bankruptcy prediction. Thus, this showed that neural network and machine learning algorithms in general has better predictability than statistical models of the past.

2.3.2 Support Vector Machines

Another machine learning model used for corporate bankruptcy prediction is the Support Vector Machines (SVM). SVM is a classification algorithm used to obtain a decision surface, which is a hyperplane which has the maximum distance between data points of both classes to be classified (Hamel, 2009). Hamel (2009) states that this algorithm reduces the probability of misclassification, thus improving accuracy. Huang et al. (2004) proposed the use of SVM over NN for the use of corporate credit rating analysis and found that SVM had slightly better explanatory and predictive power. Kim (2010) used SVM for bankruptcy prediction of SMEs and discovered that it had better accuracy than NN and logit regression. Thus, it is understood that that machine learning models such as NN and SVM has generally higher predictive performance as compared to statistical models.

Despite their superior performance, NN and SVM are known to be black boxes which do not provide meaningful answers to the cause of corporate bankruptcy (Benítez, Castro, & Requena, 1997). Kim et al. (2020) noted that these models do not provide as much explainability as opposed to its superior predictive capability. Explainability is defined by Gilpin et al. (2018) as the extent of knowing how an input influences the outcome of a model. Thus, models like NN and SVM are black boxes that does not provide any additional meaningful insight other than bankruptcy or non-bankruptcy.

2.3.3 Decision Trees as Viable Alternative

A machine learning method, the Decision Tree (DT) model provides a solution to the black box problem without sacrificing predictive capabilities. According to Kim et al. (2020), DT is used to solve classification and regression problems, like in the case of bankruptcy prediction, by charting decision rules in a tree structure. This tree structure has multiple nodes which represents an individual factor's probability of predicting an outcome as it goes down the

structure. This method provides insight on predictor's importance and the final decision tree is more explainable as compared to NN and SVM. In terms of predictive capabilities, Olson et al (2012) suggests that DT was able to provide more accurate results as compared to NN and SVM. Golbayani et al (2020) seconded that results and stated that decision trees had superior performance as compared to NN and SVM. Thus, it can be noted that DT has better predictive performance and better explainability as compared to NN and SVM.

Despite DT proving superior to NN and SVM, there is a lack of studies on the use of DT on corporate bankruptcy prediction. This research gap has inspired the possibility of using a DT model to predict corporate bankruptcy and to understand the factors that causes corporate bankruptcy.

3. Data Understanding and Preparation

3.1 Dependent and Independent Variables

The data used in the study are financial ratios of listed energy companies in the North America region. Factset, a software akin to Bloomberg, is used to collect these data. The companies which are bankrupt are removed from the dataset after their bankruptcy year. From Factset, Table 1 below shows the number of listed firms in the energy sector in years 2011 to 2018.

Year	Number of Companies
2011	778
2012	725
2013	676
2014	631
2015	563
2016	521
2017	497
2018	461
Total	**4,852**

Table 1: Number of Listed Firms in Energy Sector in 2011 to 2018

A total of 51 financial ratios spread across six different categories were extracted for each of the listed firms. The financial ratios and their corresponding categories are listed in Table 2 below. The description of each ratio is provided in Appendix A.

Liquidity Ratios	Cash Flow Ratios	Leverage Ratios	Activity Ratios	Profitability Ratios	Valuation Ratios
Current Ratio	Operating Cash Flow per Share	Debt Ratio	Inventory Turnover	Net Profit Margin	Price to Earnings Ratio
Quick Ratio	Free Cash Flow (FCF) per Share	Debt to Equity Ratio	Days' Inventory on Hand Ratio	Gross Profit Margin	Price to Book Ratio
Cash Ratio	Cash Flow from Operations (CFO) / Sales	Debt to Capital Ratio	Receivables Turnover	Operating Margin Ratio	Dividend Payout Ratio
Cash Conversion Cycle	FCF /Sales	Times Interest Earned Ratio	Days' Sales Outstanding (DSO) Ratio	Return on Assets	Dividend Yield Ratio
Working Capital/Total Assets	CFO / Total Assets	Equity Multiplier	Payables Turnover Ratio	Return on Capital Employed	Retention Ratio
	CFO / Short Term Debt	Retained Earnings / Current Liabilities	Days Payable Outstanding	Return on Equity	Price to Cash Flow Ratio
	FCF / Current Liabilities	Equity / Total Assets	Fixed Asset Turnover	Earnings Per Share	Price /Sales
	FCF / Short Term Debt	Equity / Fixed Assets	Working Capital Turnover Ratio	EBITDA/Total Asset	
		Interest Expense / Debt	Sales/Total Asset	Equity Growth	
			Sales Growth		
			Net Income Growth		
			Asset Turnover Ratio		
			Asset Impairment/Total Asset		

Table 2: Financial Ratios Extracted

Thus, the final uncleaned dataset has a total of 4,852 entries throughout 2011 to 2018, with 51 different financial ratios being used for this study.

The dataset's data quality was analysed using IBM SPSS modeller. The dataset's data quality report is as shown in Figure 1 below. From Figure 1, it can be noted that the dataset has significant amount of missing data.

Complete fields (%): 74.51% Complete records (%): 0.12%

Field	Measurement	% Complete	Valid Records
Current Ratio	Continuous	100	4852
Quick Ratio	Continuous	100	4852
Cash Ratio	Continuous	100	4852
Working Cap...	Continuous	100	4852
Operating Ca...	Continuous	100	4852
Free Cash Fl...	Continuous	100	4852
CFO / Sales	Continuous	100	4852
Free Cash Fl...	Continuous	100	4852
CFO / Total A...	Continuous	100	4852
FCF / Current...	Continuous	100	4852
Debt Ratio	Continuous	100	4852
Debt to Equit...	Continuous	100	4852
Debt to Capit...	Continuous	100	4852
Times Intere...	Continuous	100	4852
Equity Multipl...	Continuous	100	4852
Retained Ear...	Continuous	100	4852
Equity/Total A...	Continuous	100	4852
Equity/Fixed ...	Continuous	100	4852
Interest Expe...	Continuous	100	4852
Receivables ...	Continuous	100	4852
Days Sales ...	Continuous	100	4852
Payables Tur...	Continuous	100	4852
Days Payabl...	Continuous	100	4852
Fixed Asset T...	Continuous	100	4852
Sales/Total A...	Continuous	100	4852
Sales Growth	Continuous	100	4852
Asset Turnov...	Continuous	100	4852
Net Profit Mar...	Continuous	100	4852
Gross Profit ...	Continuous	100	4852
Operating Ma...	Continuous	100	4852
Return on As...	Continuous	100	4852
Return on Eq...	Continuous	100	4852
Earning Per ...	Continuous	100	4852
EBITDA/Total...	Continuous	100	4852
Equity Growth	Continuous	100	4852
Price to Book...	Continuous	100	4852
Dividend Yiel...	Continuous	100	4852

Bankruptcy S...	Flag	100	4852
Price /Sales	Continuous	67.354	3268
Net Income ...	Continuous	60.82	2951
CFO / Short T...	Continuous	49.629	2408
FCF / Current...	Continuous	48.413	2349
Price to Cas...	Continuous	46.393	2251
Working Cap...	Continuous	45.981	2231
Inventory Tur...	Continuous	30.791	1494
Cash Conv C...	Continuous	29.431	1428
Dividend Pay...	Continuous	26.216	1272
Retention Ra...	Continuous	26.216	1272
Price to Earni...	Continuous	25.433	1234
Return on Ca...	Continuous	21.538	1045
Asset Impair...	Nominal	6.43	312

Figure 1: Data Quality Report

3.2 Issues faced with Dataset

There are three main issues faced with the current dataset. Firstly, there are missing data within the dataset. Since 51 factors are used throughout all the companies and throughout the entire duration, there might be cases which the companies ceased to exist after bankruptcy, or that certain data are not available due to the differences in financial ratios available. From the data quality report, it can be noted that none of the fields have 100% complete data. This means that we are unable to remove the fields with incomplete data. To overcome this issue, the dataset will be filtered to only include fields with a minimum percentage of 70% complete data.

Secondly, the bankruptcy field does not indicate the bankruptcy year of the company. The status only reflects the current bankruptcy status of the company. Thus, companies who are bankrupt are stated as bankrupt for all the years that are present in the dataset. After further study on the individual bankrupt cases, it can be noted that the last financial ratios given for a bankrupt company, is generally one or two years before the actual bankruptcy ruling. An assumption is required to overcome this issue, that is that the last filing of financial ratio is representative of a company facing financial distress, thus causing the corporate bankruptcy.

The years prior to the last filing will have their bankruptcy status changed to “non-bankrupt” instead.

Lastly, as our dataset consists of listed companies and the occurrence of bankruptcy amongst listed companies are rare, there is insufficient bankruptcy data to train the predictive model. This lack of representation of bankrupt cases leads to the class imbalance, which significantly reduces the performance of predictive models (Chawla, 2005). This class imbalance is detrimental to the model’s performance as, the model will tend to favour the prediction performance of the majority class, at the cost of more false negatives (Hasanin, Khoshgoftaar, Leevy, & Seliya, 2019). In our case, this will cause the model to predict more non-bankrupt cases more accurately over bankrupt cases, which is detrimental to our study as we are aiming to predict bankrupt cases. Thus, two methods will be used to combat this class imbalance. Firstly, the dataset will be sampled to increase the proportion of bankrupt cases in the data preparation phase. Secondly, there will be misclassification cost assigned to the training dataset during the modelling phase, which has proven to improve the predictive performance of minority class samples (Pazzani, et al., 1994).

3.3 Data Preparation

The data preparation process will be done using Python 3.8 on Jupyter Notebook. The following flowchart summarises the data preparation processes.

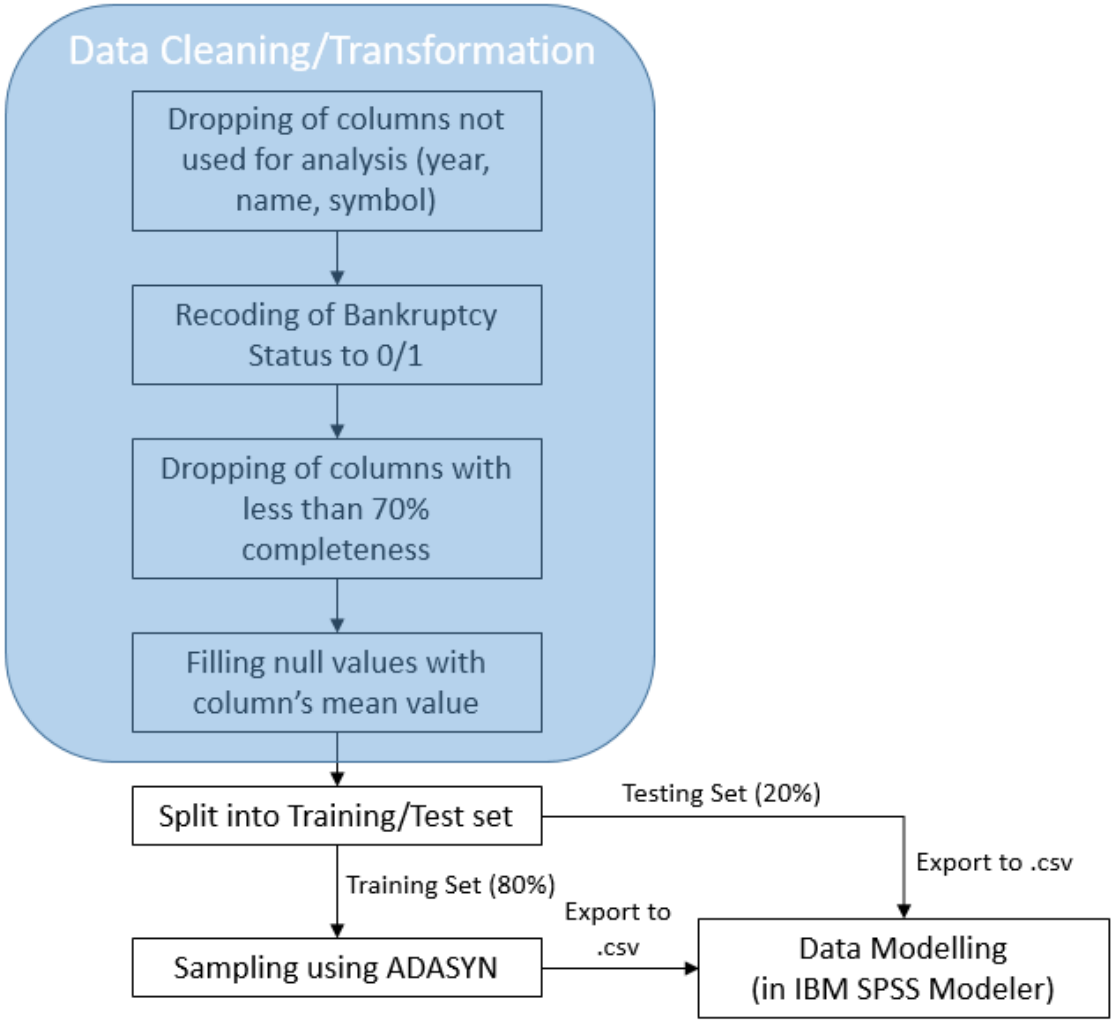

Figure 2: Flowchart of data preparation, transformation, and sampling steps

The three main processes for the data preparation are the data cleaning/transformation, splitting of dataset and data sampling.

3.3.1 Data Cleaning/Transformation

Prior to data cleaning, the data columns which are not used for analysis are dropped, such as year and name. The target variable, Bankruptcy, will be recoded to 1 for bankrupt and 0 for not bankrupt.

From the data quality report shown in Figure 1 above, it is known that 13 data columns have less than 70% completeness. Thus, these columns will be removed. Table 3 below shows the remaining 38 data fields.

Liquidity Ratios	**Cash Flow Ratios**	**Leverage Ratios**	**Activity Ratios**	**Profitability Ratios**	**Valuation Ratios**
Current Ratio	Cash Flow from Operations (CFO) / Sales	Debt Ratio	Receivables Turnover	Net Profit Margin	Price to Book Ratio
Quick Ratio	CFO / Total Assets	Debt to Equity Ratio	Days' Sales Outstanding (DSO) Ratio	Gross Profit Margin	Price /Sales
Cash Ratio	FCF / Current Liabilities	Times Interest Earned Ratio	Payables Turnover Ratio	Operating Margin Ratio	
Working Capital/Total Assets		Equity Multiplier	Days Payable Outstanding	Return on Assets	
		Retained Earnings / Current Liabilities	Sales/Total Asset	Return on Equity	
		Equity / Total Assets	Sales Growth	Earnings Per Share	
		Equity / Fixed Assets		EBITDA/Total Asset	
				Equity Growth	

Table 3: Remaining financial ratios to be used for modelling

The remaining columns will have their missing value replaced with the mean value of their specific column. The cleaned dataset will then be split 80:20, with 80% being used as the training dataset and 20% used as testing dataset.

The training dataset will be then be used for the data sampling stage while the testing dataset will be left intact and exported into Tab-Separated Values (TSV) format to be used in IBM SPSS Modeller. Table 4 below shows the number of records for pre-sampled training set and testing set.

	Initial Dataset	Training Dataset	Test Dataset
Non-bankrupt	4,492	3,602	890
Bankrupt	360	279	81
Total	**4,852**	**3,881**	**971**

Table 4*: Number of records pre-sampled training set and test dataset*

The codes used for the data cleaning/transformation can be found in Appendix B.

3.3.2 Data Sampling

The training dataset requires sampling to combat the class imbalance problem. Conventional sampling methods such as random oversampling and undersampling are explored.

Random oversampling refers to the replication of minority class data points to alleviate the class imbalance (Kaur & Gosain, 2018). The advantage of this approach is that there will be no loss of data as the information of the original dataset remains intact. However, as the method replicates the same points, it might lead to possible overfitting of the predictive models as the model learns very specific regions of minority classes as they get more prominent. Furthermore, in this study, the minority case is approximately 7.75% of the training dataset and replication of these dataset will very likely lead to overfitting of the model and is thus detrimental to the study. Thus, oversampling is not satisfactory to be used for the data sampling of training set.

Random undersampling refers to the reduction of majority class data points randomly, thus allowing the minority class to be more prominent, leading to reduced bias (Kaur & Gosain, 2018). However, this leads to loss of data as the majority class data is disposed. In this study, the training dataset only has 3,881 cases and reducing the amount of dataset will lead to the

deprivation of data for learning and thus detrimental to the model's performance. Thus, the undersampling is not a viable option in this case.

In more recent studies, the adaptive synthetic (ADASYN) sampling approach has been explored and deemed as a viable option to combat class imbalance (Le, Lee, Park, & Baik, 2018). ADASYN is a synthetic sampling approach inspired by the Synthetic Minority Oversampling Technique (SMOTE) technique. As the SMOTE technique is limited to creating equal numbers of synthetic sample per minority data point, ADASYN adaptively generate different number of synthetic data based on the weighted distribution of minority class samples based on their difficulty to learn (He, Bai, Garcia, & Li, 2008), overcoming the limitation of SMOTE. By doing that, ADASYN will generate more minority class samples which are harder to learn. This will reduce the bias caused by class imbalance as it shifts the classification decision boundary adaptively towards the harder to learn samples, allowing the predictive models to classify the minority classes better.

For this study, the training dataset will be sampled using the ADASYN technique. The sampling will be done on Jupyter Notebook using Python 3.8. The sampled dataset will then be export into Tab Separated Value (TSV) format and modelling will be done using the IBM SPSS modeler. The codes used for data sampling can be found in Appendix B.

The training dataset has 279 bankrupt cases, approximately 7.19% of the dataset. Using ADASYN, the bankruptcy cases are oversampled to 911 cases, making up approximately 20.17% of the dataset. The dataset is sampled at approximately 20% to have adequate amount of bankruptcy cases for model training but still representative of the minority proportion of the original dataset. Figure 3 below shows the distribution of bankruptcy cases for the sampled training dataset which will be exported and used for modelling.

Balance of positive and negative classes (%):

```
0    79.813871
1    20.186129
Name: Bankruptcy Status, dtype: float64
```

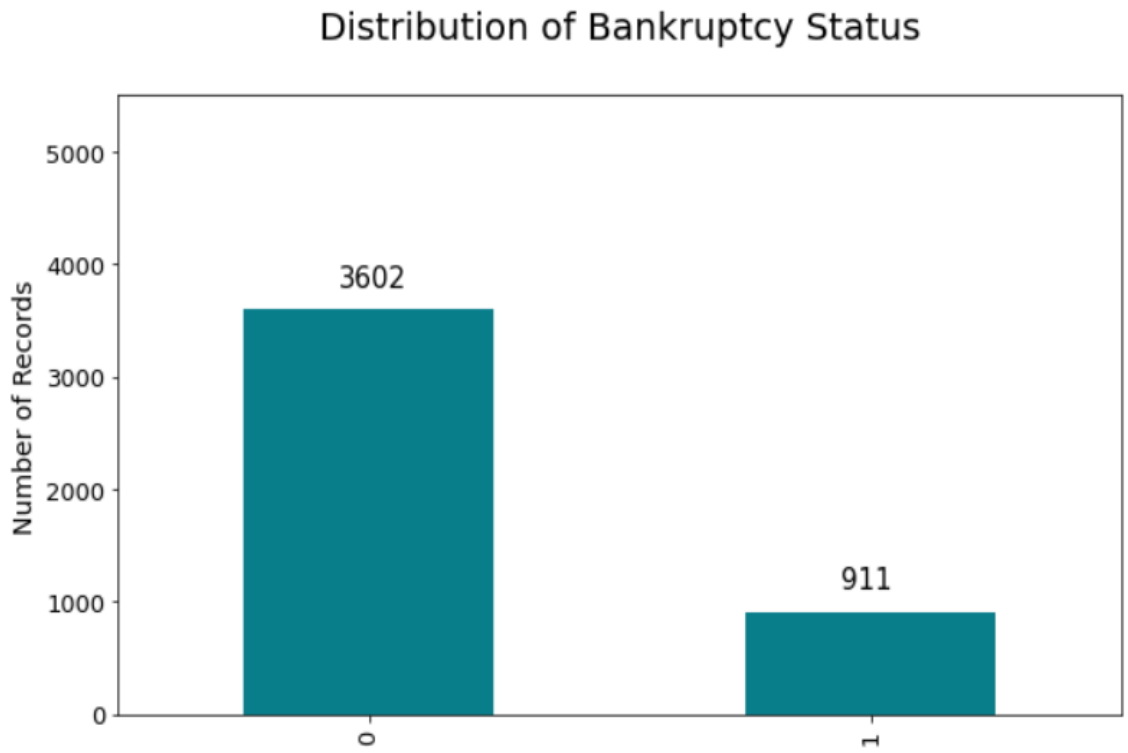

Figure 3: Distribution of bankruptcy cases for the sampled training dataset

Figure 4 below shows the proportion and the final number of data for the training dataset and test dataset as compared to the initial dataset. The test dataset is kept at a similar proportion with the initial dataset and not sampled.

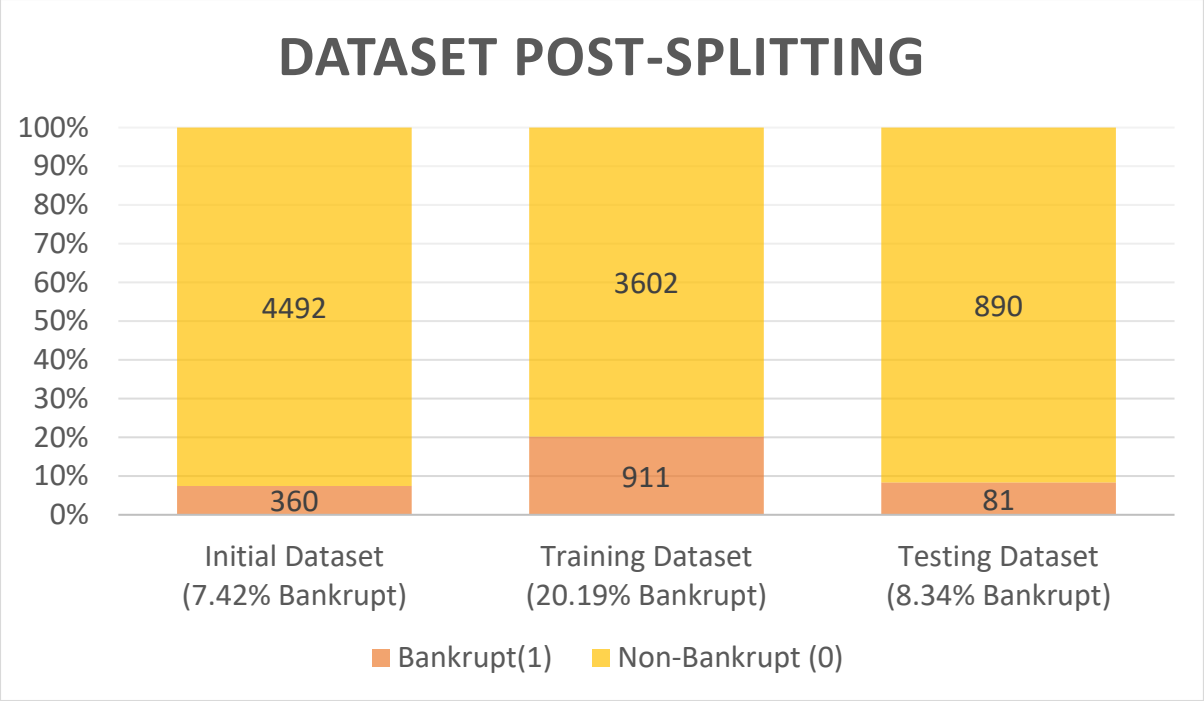

Figure 4: Final proportion and number of records for initial, training dataset and test dataset

4. Modelling and Evaluation/Discussion

4.1 Methodology

The existing predictive methods in the industry currently are the Altman Z-score (MDA) model and Logit model. Despite having high model explainability, there are a few issues pertaining to these models. As discussed in the literature review section, both the MDA and Logit model are statistical in nature and they require the assumptions of normality and non-multicollinearity. There might not be normality and non-multicollinearity for real-world datasets and the data transformation used to fit either the normality constraint or the removal of multicollinear variables will cause data loss. Furthermore, it was proven that machine learning methods have been able to produce models with higher predictive capability.

This paper proposes the use of DT models to predict corporate bankruptcy. The research problem revolves around the need for a highly predictive machine learning model while providing high level of explainability as opposed to a black box model. From the Literature Review section, it can be concluded that DT model has generally higher accuracy as compared to SVM and NN, while providing higher explainability. Thus, the DT model is the proposed method to be used for this paper. To determine the DT model's capability, a Logistic Regression and Neural Network model will be generated using the same training dataset and tested with the same test dataset. The flow of modelling will be as follows: comparison of DT models to determine the champion DT model, the champion DT model will be used to compare against the Logistic Regression and Neural Network model.

4.2 Evaluation Methods

To accurately evaluate the performance of DT for the purpose of corporate bankruptcy, the DT models using different algorithms and settings are evaluated. The optimised DT model will then be used to compare with other predictive methods to evaluate its performance, namely the

Logit Regression model and NN model, to compare the model's predictive capability and explainability. They will be compared using the Recall, Area under the Receiver Operating Characteristic (ROC) curve (AUC), and the Lift Charts. The optimisation of models will be done using the settings that yield the highest recall with the least drop in AUC.

4.2.1 Recall

Recall is a confusion metric that reflects the proportion of total true positive values over actual positive cases. To further understand this metric, the confusion matrix must be elaborated. The confusion matrix is a two-dimensional table that summarises the performance of a trained classification model as compared to test data (Ting, 2011). Figure 5 below show the different parts of a confusion matrix.

	Predicted: Non-bankrupt	Predicted: Bankrupt
Actual: Non-bankrupt	**True Negative** – Firm is non-bankrupt and predicted as non-bankrupt	**False Positive** – Firm is not bankrupt but predicted as bankrupt
Actual: Bankrupt	**False Negative** – Firm is bankrupt but predicted as non-bankrupt	**True Positive** – Firm is bankrupt and predicted as bankrupt

Figure 5: Confusion Matrix Example

The "Actual" column refers to the actual outcome based on the test data, while the "Predicted" row refers to the predicted outcome as generated by the model. By comparing the actual results with the predicted results by the model, there will be four outcomes as shown in Figure 5 above.

True Positive (TP) refers to the outcome where the positive event is predicted correctly as per the test data, while True Negative (TN) refers to outcome where the negative event is predicted corrected as per the test data. False Positive (FP), otherwise known as Type I error, refers to positive event which are wrongly predicted. FP are events which are actually negative but predicted to be positive. In contrast to this, False Negative (FN), otherwise known as Type II

Error, are events that are actually positive, but predicted to be positive (Towards Data Science, 2018). For this paper, a positive case will refer to if the company is likely to bankrupt, while the negative will refer to not likely to be bankrupt.

There will be five confusion metrics generated to showcase the performance of each model based on the confusion matrix. They are accuracy, misclassification, precision, recall, and specificity. Accuracy measures the extent of correctly predicting events (TP) and non-events (TP) (Koh, 2005). Conversely, misclassification measures the extent of incorrect predictions of events (FP) and non-events (FN). Recall measures solely the extent of true positive prediction (TP) as compared to all actual positive prediction (TP + FN), while specificity measures the extent of actual negative prediction (TN) as compared to all negative prediction (TN + FP) (Chan, et al., 2006). Precision refers to the extent of events that occur based on the predictive model, which can be either positive or negative. These metrics are summarised in Figure 6 below.

	Formula
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Misclassification	$\frac{FP + FN}{TP + TN + FP + FN}$
Recall	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
Precision (Positive)	$\frac{TP}{TP + FP}$
Precision (Negative)	$\frac{TN}{TN + FN}$

Figure 6: Confusion Metrics

For this paper, the positive event is set as “Bankrupt”, while negative is “Non-bankrupt”. Since this paper focuses on the prediction of corporate bankruptcy, the evaluation of proposed models

should focus heavily on the recall as it emphasises on the correctly predicting actual bankrupt cases as compared to non-bankrupt cases. As more samples are non-bankrupt in the dataset, model will learn to correctly classify non-bankrupt better than bankrupt, leading to the increase of overall accuracy as the model learns to classify non-bankrupt cases better than bankrupt cases. However, the recall value will reflect the actual extent of whether the model is able to distinguish bankrupt cases from False Negatives. Furthermore, a wrong prediction in the case of a False Negative will mean that a firm in financial distress is predicted as non-bankrupt, possibly leading to a bad loan. Thus, there is ample incentive to focus on Recall to evaluate the performance of the model.

4.2.2 AUC

The AUC is a summary statistic of the Receiver Operating Characteristics (ROC) curve. The ROC curve is also used to measure the discriminatory power of a classification model (Hajian-Tilaki, 2013). The curve plots the True Positive Rate (TPR) vs. the False Positive Rate (FPR). Figure 7 below shows an example of the ROC curve.

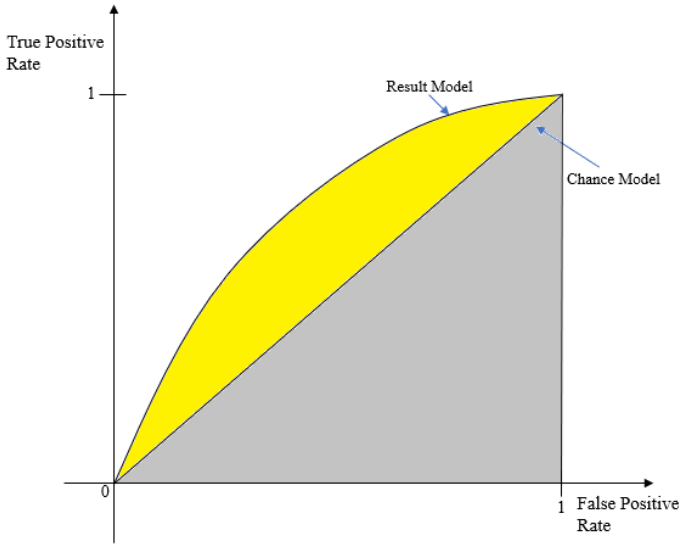

Figure 7: Example of ROC Curve

The linear line stands for the chance model, which classifies the cases randomly with zero information gain. If the line is above the chance model, as shown in Figure 7 as “Result Model”, the model will have more predictive capability than chosen in random.

The summary statistic of the ROC curve, the AUC measures the total two-dimensional area below the ROC curve, as shown by the entire area in yellow and grey. The AUC measures the performance of the model across all thresholds of classification. It can be interpreted as the probability of the model classifying a positive result rather than a negative result. The AUC values range from 0 to 1 and the higher the AUC, the stronger the predictive model. A model with an AUC of 1 is a model that predicts True Positive cases 100% of the time and while a model with AUC of 0 is a model that predicts False Positive 100% of the time.

Both evaluation methods will be used to evaluate the performance of the proposed DT, Logit and NN model. The model with the best AUC/ROC score will be considered the model with the highest predictive power.

4.2.3 Lift Chart

The third evaluation method will be the use of lift charts. The lift chart is used to evaluate and compare the performance of predictive machine learning models predicting the same variables (Witten & Frank, 2002). The chart is a graphical representation of the improvement in response when a predictive model is used as compared to a random guess (Horvat, Jovic, & Ivošević, 2020). Figure 8 below shows an example of the lift chart.

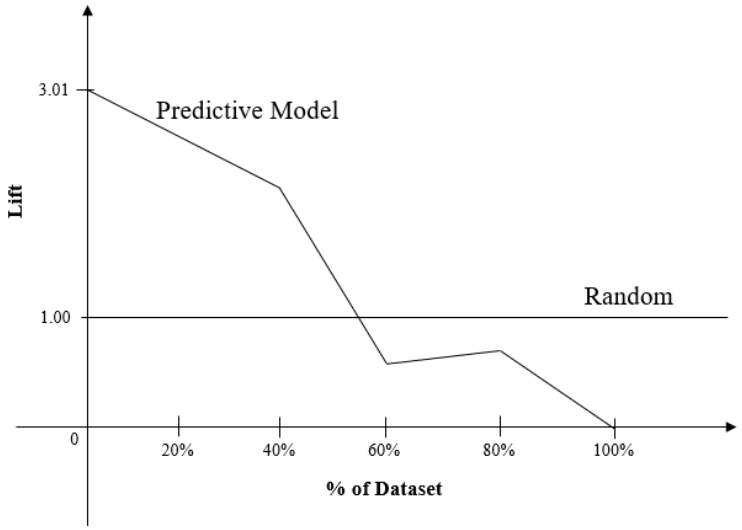

Figure 8: Example of a Lift Chart

IBM SPSS can produce a lift chart using the evaluation node. Using this function, the proposed models will be evaluated and compared. The model with the highest lift value will be considered the best.

4.3 Decision Tree

The DT model is a supervised classification model that consists of two different parts, the training stage, and the testing stage. In the training stage, the model will be developed based on the training dataset, which will learn to classify the dependent variables based on independent variables provided within the training dataset. Using this trained model. The testing data, which is unseen to the model, will be introduced to assess the model's accuracy. If the trained model achieves a satisfactory accuracy, it can be then deployed on other dataset for prediction.

The top of the DT is referred to as the Root Node, which contains the entire dataset. DT works by classifying the examples given in the dataset, and sort them down the tree, splitting them into subset of the data above it. If the leaves do not split further into any subset, that node is called a terminal node, which contains the predictive outcome (Granström & Abrahamsson, 2019). Each split is defined by an attribute as given as an independent variable fed to the model. For each attribute, the node is split into subsets that better represent the predictive outcome. This split is then repeated and stopped when all the observations in a subset are further splitting of the data stops adding valid to the prediction (Devi & Nirmala, 2013). This will provide a visual representation of how the model will classify the dependent variable based on the attributes of the independent variable. An example of this is shown in Figure 9 below.

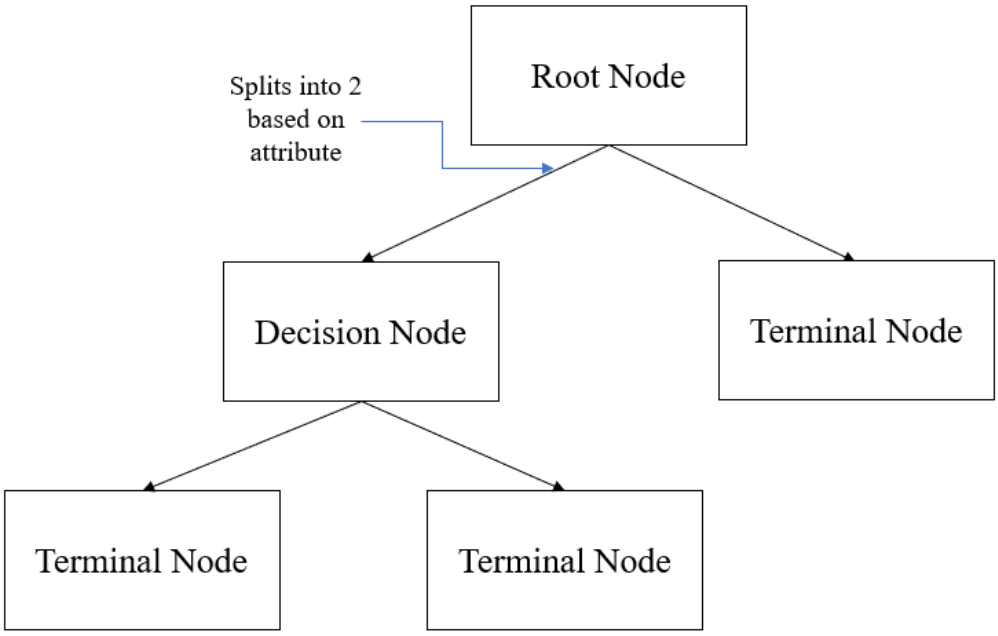

Figure 9: Example of a Decision Tree

For this paper, the Gini Index will be used for the splitting of each partition. The Gini index measures the probability of a variable being wrongly classified while selected at random. If the node has a Gini value of close to 1, it means that it has higher chance of being wrongly classified, otherwise called “impure”. If the node has a Gini close to 0, it is considered pure

and more homogenous. Thus, the impure nodes will be further split into two subsets to ensure that the resulting terminal nodes are as pure as possible.

Four types of DT are used for this study. They are: CART, CHAID, C5.0 and Quest. The dataset will be modelled using all four types of DT and the models' performance will be evaluated and ranked. For this paper, we used the IBM SPSS modeller for all modelling methods.

For each DT Model, there are several parameters to be optimised. The misclassification costs are set to better identify false negatives and reduce them to produce models with higher recall rates. The respective False Negative misclassification costs for each model as follows: CART is set as 10, CHAID is set at 3, C.5 and QUEST are set at 5. The stopping rules for CART, C5.0 and QUEST are set as default, while C5.0 is set as absolute value, with minimum in parent branch set as 400 while child is 100. All models except CHAID has pruning option activated and set at 100 for C5.0. CHAID do not have pruning option available. The model will be trained using the training dataset and the resulting model will be tested using the testing dataset.

4.4 Comparison models: Logistic Regression and Neural Network

For comparison, a Logistic Regression model and a Neural Network model will be trained using the same training dataset and the resulting model tested using the test dataset. The champion DT Model will be used to compare with the Logit and NN model. Like the DT models, the models are optimised to obtain the highest recall with lowest drop in AUC.

For the Logit model, the optimised parameters are as follows: Binomial procedure selected, Forwards Stepwise method used, and removal criterion is set as Likelihood ratio. For the NN model, the optimised parameters are as follows: Multilayer Perceptron setting used, with number of hidden layers set as 1, stopping rule set as error not being able to decrease further. The model with the highest recall and AUC, with the best performance in Lift Chart will be selected as the best model for bankruptcy prediction.

4.5 Results

The results are split into the selection of champion of DT model (covered in 6.1) and the subsequent comparison with the NN and Logit model (covered in 6.2)

4.5.1 Results of DT models

		Recall	AUC	Accuracy (for reference)
CART	Training	93.74%	0.887	84.89%
	Testing	82.72%	0.837	82.80%
C5.0	Training	99%	0.944	90.41%
	Testing	**93.83%**	**0.933**	**89.29%**
CHAID	Training	87.49%	0.945	86.55%
	Testing	76.54%	0.918	84.65%
QUEST	Training	75.02%	0.761	68.14%
	Testing	70.37%	0.699	65.71%

Figure 10: Results of DT models

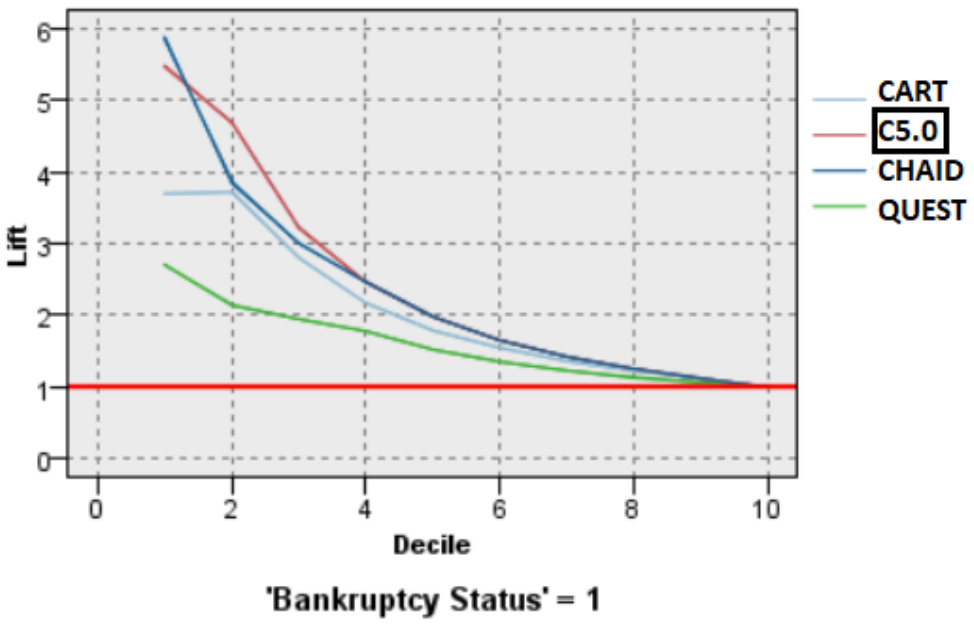

Figure 11: Lift chart of DT models

From Figure 10 above, it can be noted that C5.0 DT has the highest recall and AUC for the testing dataset. In Figure 11 above, it is noted that C5.0 generally has the highest lift value throughout the dataset except the first decile. Thus, it can be concluded that the C5.0 DT is the champion DT model and will be used to predict the bankruptcy of companies in the energy sector.

4.5.2 Discussion of observations from the champion model (DT C5.0)

Figure 12 shows the decision tree generated by the champion DT model. The decision tree has a tree depth of eight.

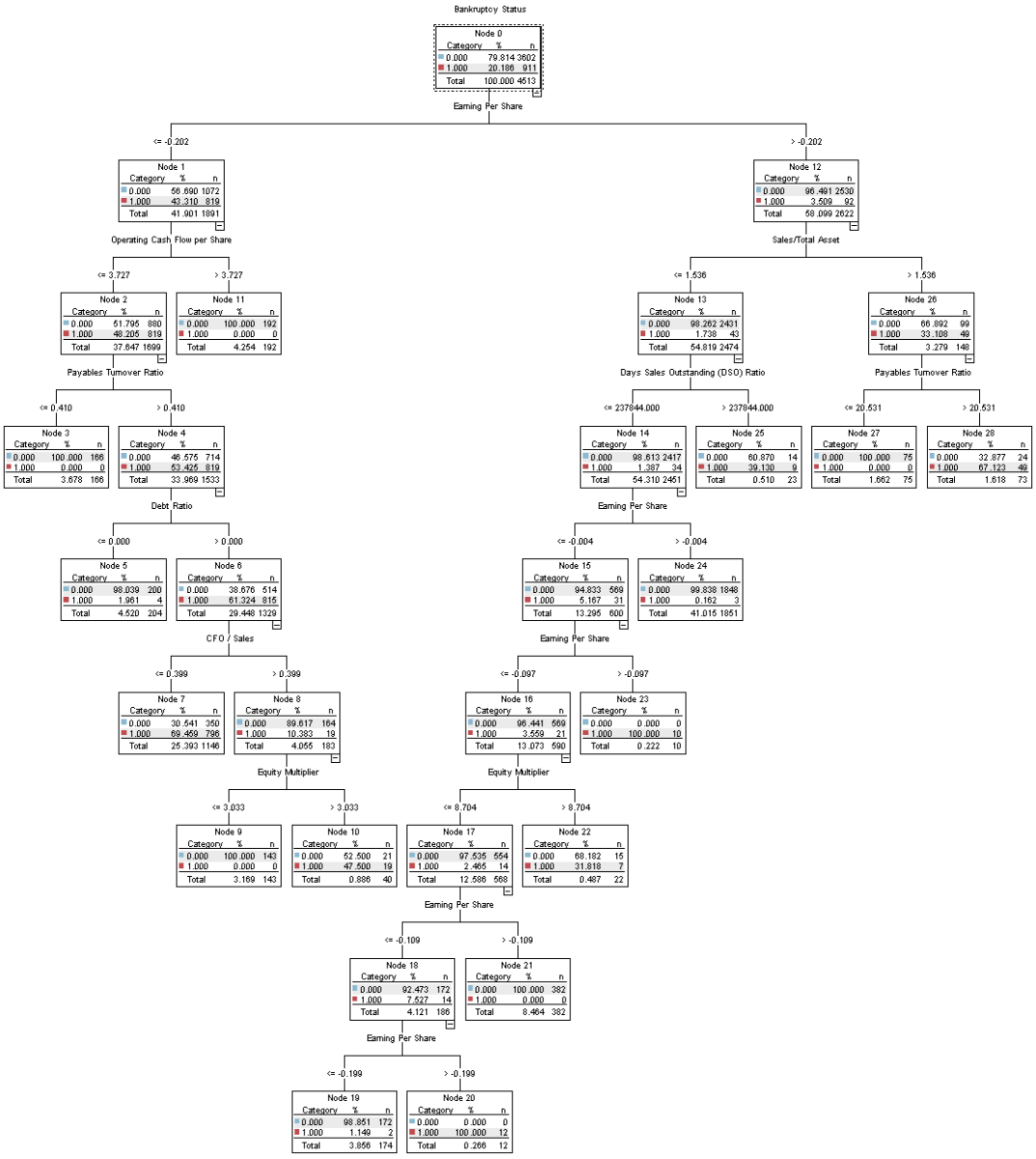

Figure 12: Generated decision tree

Figure 13 shows the predictor importance chart of the model. From Figure 13, the most important variable associated with bankruptcy prediction is earning per share.

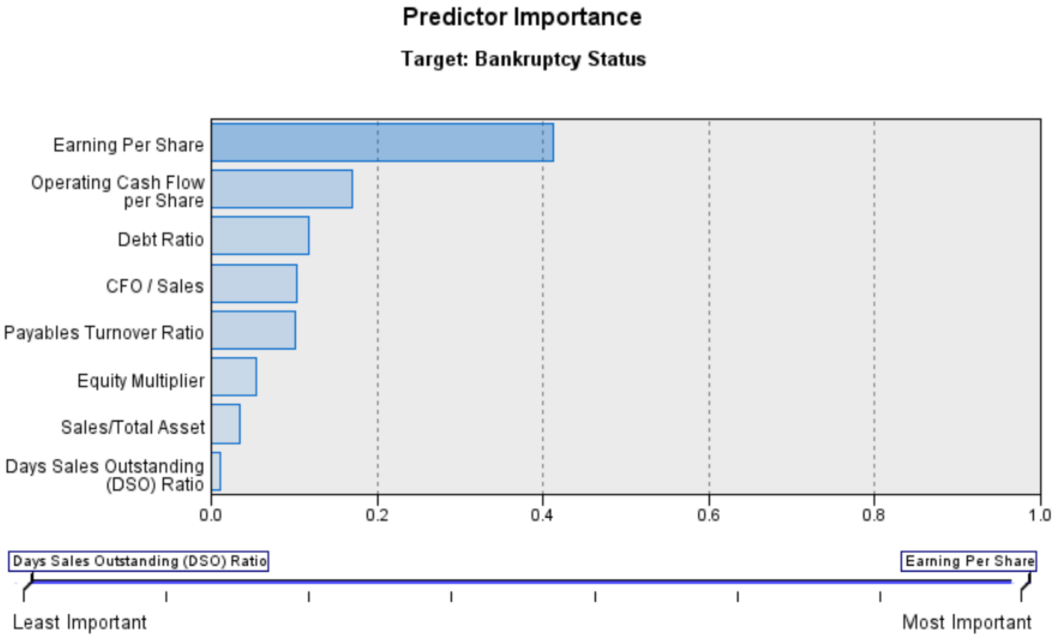

Figure 13: Predictor importance of DT model

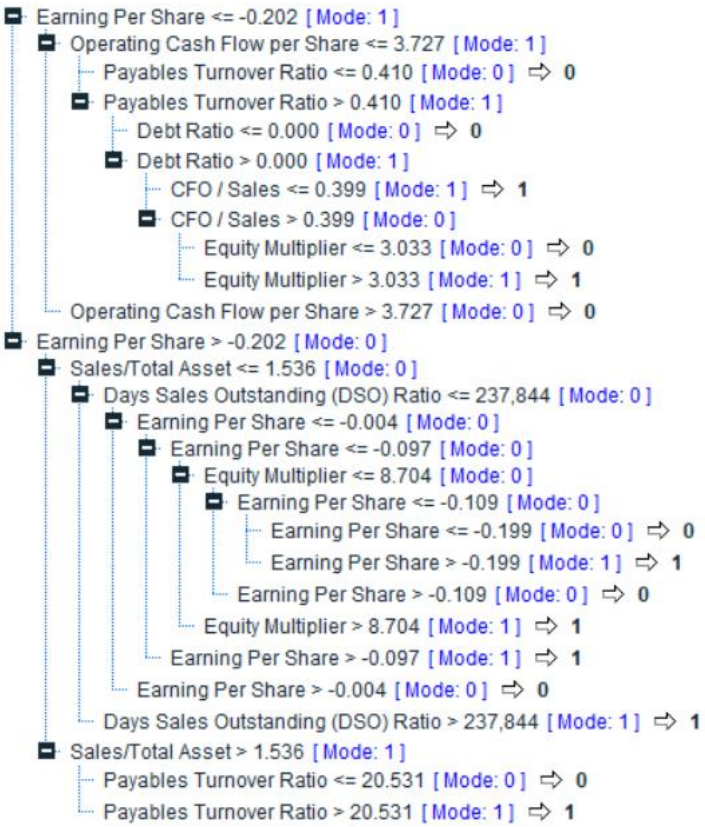

Figure 14: Decision rules generated

Figure 14 shows the decision rules generated. From the decision rules stated in Figure 14, firms who are more likely bankrupt generally has the following traits:

1	Lower Earnings per Share
2	Lower Operating Cash Flow per Share
3	Higher Debt Ratio
4	Lower Cash Flow from Operations over Sales

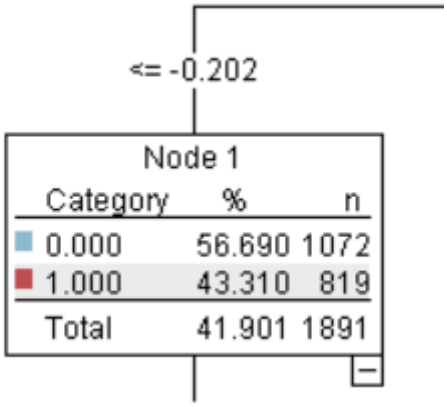

Figure 15: Node 1 of DT model

Node 1 reflects the importance of earning per share as a predictor. 819 out of 911 of the bankrupt cases were separated out from the dataset for cases with lesser than -0.202 earning per share, which accounts for 89.9% of the cases.

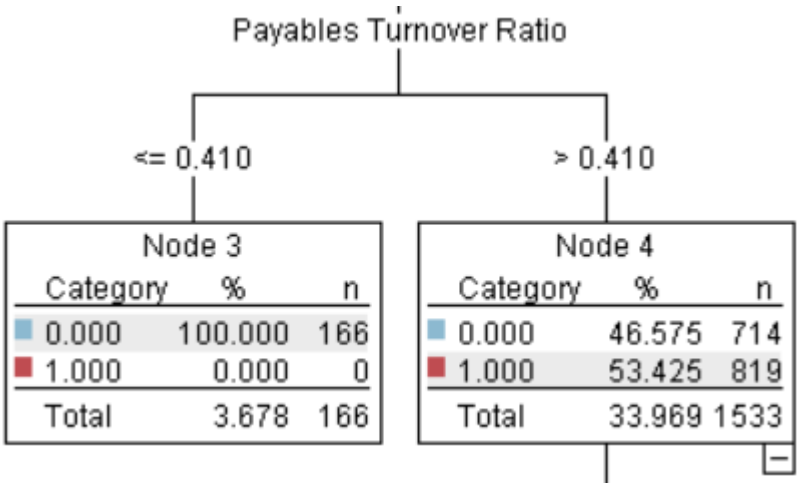

Figure 16: Node 3 of DT model

Node 3 suggests that firms with a payables turnover ratio of less than 0.410 are more likely to be non-bankrupt. However, it is known so far that payables turnover ratio are preferred to be

higher which reflects efficiency on the firm's part in clearing its short-term debt (Netsuite, 2020). Thus, it is illogical to have a firm to have higher chance of bankruptcy despite having better efficiency. This suggests that the payables turnover ratio is ineffective in the prediction of firm bankruptcy and thus can be disregarded in the final model.

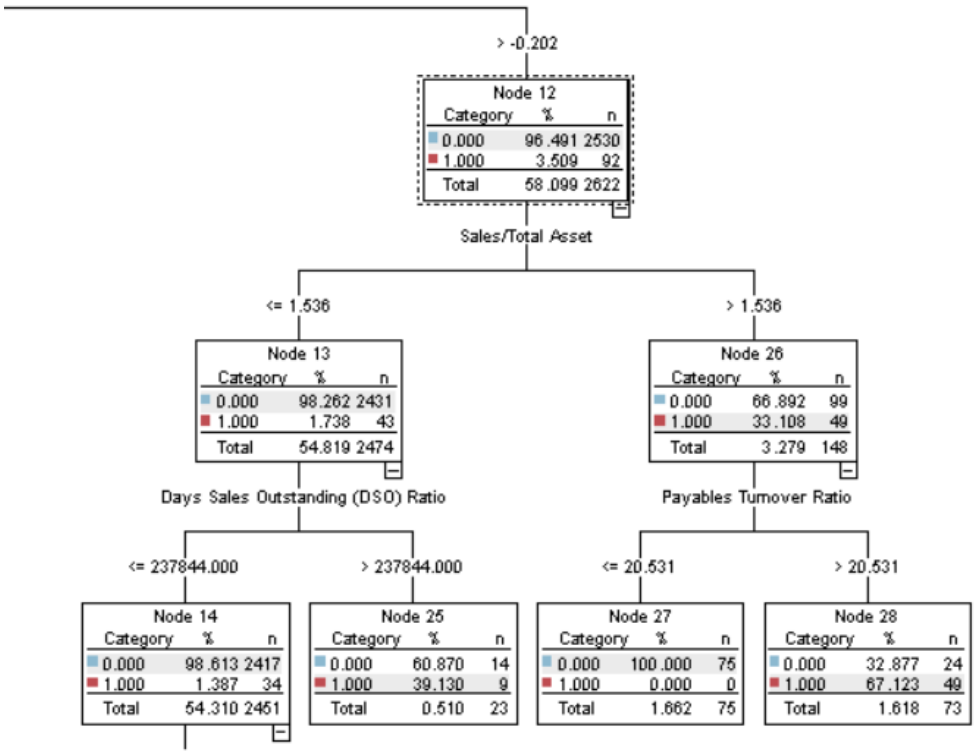

Figure 17: Node 28 of DT model

Node 28 highlights an interesting scenario, where the firm has higher than -0.202 earning per share, higher than sales over total asset of 1.536, and a great payable turnover ratio of over 20.531. This shows that, despite having great financial ratios for the year prior to bankruptcy, there might be other factors, such as macroeconomic factor or other legal troubles, that may have caused the company to go bankrupt. However, these cases make up for 49 out of 911 bankruptcy cases which is around 5%, which is quite rare. More studies can be done regarding this aspect.

4.5.3 Results of Champion DT model against Logit and NN Models

The following figures shows the results of the comparison between the champion DT model against the Logit and NN models.

		Recall	AUC	Accuracy (for reference)
Logistic Regression	Training	7.46%	0.657	80.23%
	Testing	4.94%	0.555	**91.56%**
NN	Training	67.95%	0.934	89.79%
	Testing	53.09%	0.882	91.14%
C5.0	Training	99%	0.944	90.41%
	Testing	**93.83%**	**0.933**	89.29%

Figure 18: Result of DT against Logit and NN models

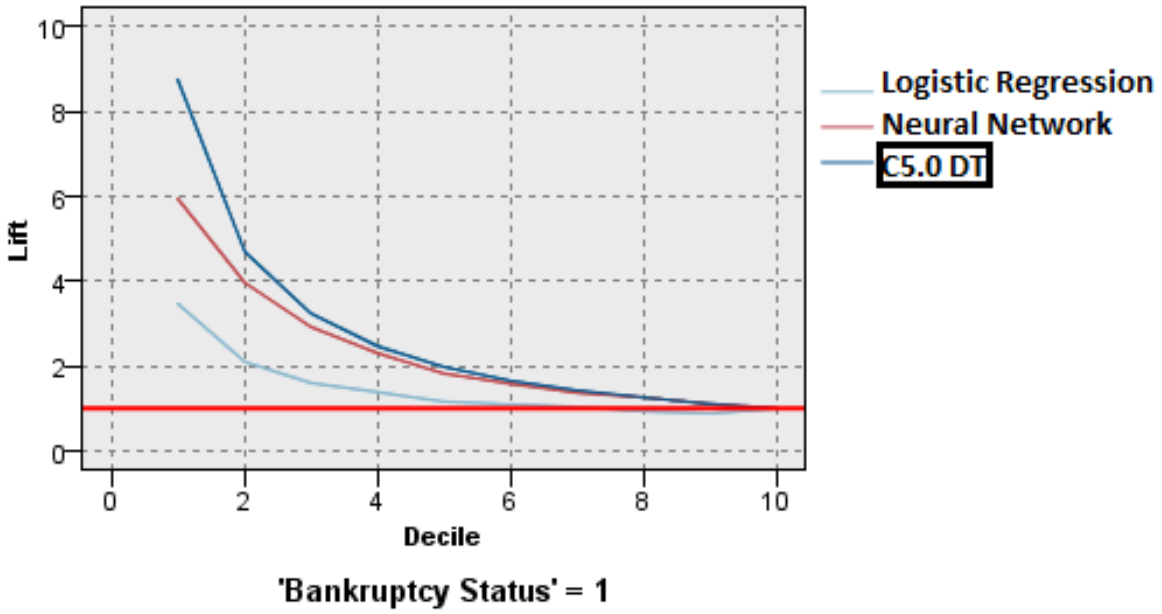

Figure 19: Lift chart of DT against Logit and NN models

C5.0 DT, which is the champion DT model is used to compare against the Logit and NN models. Despite the Logit model producing the highest test accuracy of 91.56%, it can be noted that the model has abysmal recall rates of just 4.94% and an AUC of 0.555. As the non-bankrupt cases

holds majority in the test dataset, the overall accuracy rate increases as the model correctly predicts more non-bankrupt cases.

Logistic Regression

Comparing $L-Bankruptcy Status with Bankruptcy Status

Correct	889	91.56%
Wrong	82	8.44%
Total	971	

Coincidence Matrix for $L-Bankruptcy Status (rows show actuals)

	0	1
0	885	5
1	77	4

Figure 20: Confusion matrix of Logit Model for test dataset

From the confusion matrix of the Logit model for the test dataset shown in Figure 20, it is understood that the Logit model can produce a model that is able to accurately predict non-bankrupt cases rather than bankrupt cases. The recall rate and AUC highlight the true nature of the model being unable to predict bankrupt cases, which is what this study requires. Thus, this further reinforces the idea using recall rates and AUC as the main evaluation method for bankruptcy over accuracy rates. Since the focus of study is to predict firm's bankruptcy, the Logit model proves ineffective in predicting bankrupt cases despite having higher overall accuracy.

Neural Network

Comparing $N-Bankruptcy Status with Bankruptcy Status

Correct	885	91.14%
Wrong	86	8.86%
Total	971	

Coincidence Matrix for $N-Bankruptcy Status (rows show actuals)

	0	1
0	842	48
1	38	43

Figure 21: Confusion matrix of NN Model for test dataset

This is similar in the case of NN model as the model produces higher accuracy rates but unfavourable recall rates. One possible reason for Logit and NN's bad performance might be due to both models being unable to assign misclassification costs, which was set for DT models to optimise the performance by reducing false negative cases.

Decision Tree (C5.0)

Comparing $C-Bankruptcy Status with Bankruptcy Status

Correct	867	89.29%
Wrong	104	10.71%
Total	971	

Coincidence Matrix for $C-Bankruptcy Status (rows show actuals)

	0	1
0	791	99
1	5	76

Figure 22: Confusion matrix of C5.0 DT Model for test dataset

The C5.0 DT emerges the champion model with a higher recall rate, AUC, and best performance in the lift chart. This highlights DT's performance over NN and Logit in the prediction of bankruptcy prediction.

In terms of explainability of model, the Logit model produces an equation with the final predictors as shown in Table 5. The model is explainable as it reflects how much exactly would a change in predictor value influence the result. For example, if the debt to equity ratio increases, the chances of bankruptcy increases (approaches 1).

Parameter Estimates

Bankruptcy Status[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
0	Intercept	1.629	.045	1298.772	1	.000			
	Debt to Equity Ratio	.061	.009	49.517	1	.000	1.062	1.045	1.080
	Equity Multiplier	-.054	.006	84.577	1	.000	.947	.937	.958
	Fixed Asset Turnover	-.050	.023	4.839	1	.028	.951	.910	.995

a. The reference category is: 1.

Table 5: Resulting equation of Logistic Regression

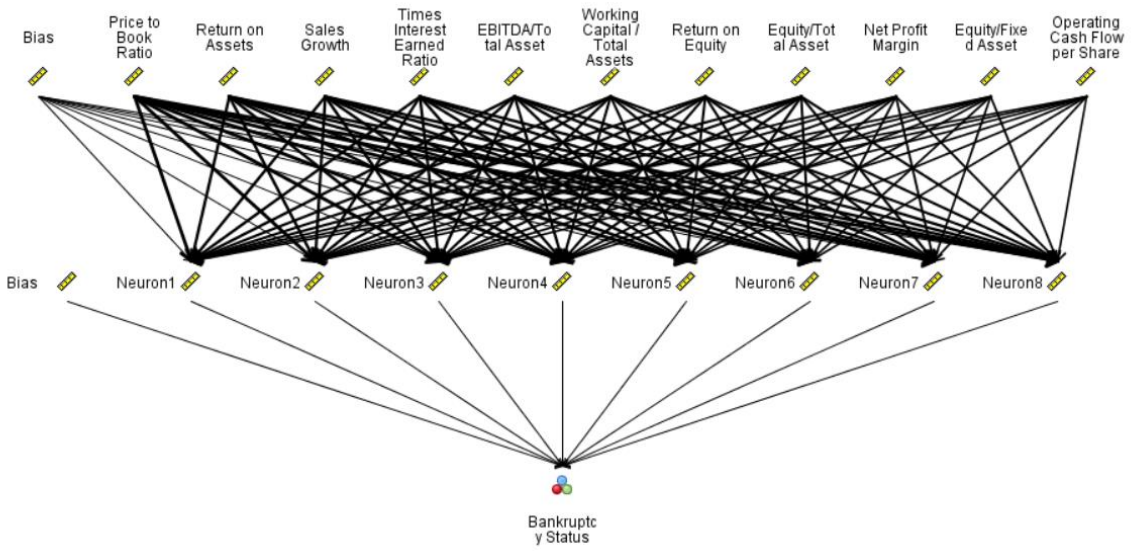

Figure 23: Resulting neural network created

From the neural network model shown in Figure 23, there are eight hidden neurons within the only layer of network. However, it does not reflect how much exactly each input affects the probability of outcome unlike the Logit model. Thus, the NN model is unable to determine how much exactly each input influences the outcome and not explainable.

The DT model produces a decision tree as shown in Figure 12 above. Although not as values are not represented explicitly in an equation like Logit model, the decision tree is more intuitive as it reflects a threshold which separates bankruptcy and non-bankruptcy and the probability of the occurrence based on that threshold. For example, in Node 1, 819 out of 911 cases of bankruptcy can be determined by having an earnings per share of lower than -0.202. This

highlight the input's influence on the model's outcome and is thus considered explainable. However, it is unfair to compare the logit model and the DT model's explainability as both models have different ways of highlighting the input's influence on the model. Thus, it can be concluded that DT is considered as an explainable model.

The strength of DT falls on the intuitiveness of the model. For DT, each input is represented as a threshold while logit is represented as a coefficient. For a layman's application, the DT model easily distinguishes the specific threshold of an input which will affect the outcome rather than the application of an equation given in the logit model. Thus, the model is easily understood, and users will know the specific implications of certain inputs which will affect the outcome. This allows DT to be easily deployed in bankruptcy prediction system, which will highlight the risk based on the threshold provided in the DT model, as compared to a logit or NN model.

5. Conclusion

This paper aimed to review and address issues with existing methods of corporate bankruptcy prediction and proposed the use of decision tree models as a viable alternative. This paper also included the use of ADASYN data sampling technique instead of conventional random oversampling and undersampling, which improved model training performance without overfitting. An alternative evaluation method, a mixture of recall, AUC, and Lift Chart, was also proposed to better determine the predictive model's performance in bankruptcy prediction as opposed to solely overall accuracy rates. The C5.0 decision tree model performed the best amongst the decision tree algorithms when tested on the test dataset based on financial data of energy sector from 2011 to 2018. With a recall of 93.83%, AUC of 0.933 and generally higher lift chart values, the C5.0 decision tree model outperformed the Logistic Regression and Neural Network model. The C5.0 decision tree has also showcased its explainability against both models and intuitiveness of use. Thus, the C5.0 decision tree model is proven to be a viable alternative to existing corporate bankruptcy prediction model.

A limitation of this study is that the model is unable the determine the state of economy of the firm as only solely financial ratios were used. The dataset provided significant knowledge of the firm's performance but there was no resolution on the state of economy surrounding the firm. This might account for the few rare cases where the firms did well but still end up bankrupt. To alleviate this problem, further studies can include the use of macroeconomic variables such as GDP or inflation rates, alongside financial ratio to better understand the effects of state of economy of the firm on bankruptcy prediction.

With single decision trees being proven viable for bankruptcy prediction, further studies should include the use of ensemble models such as random forest and boosted tree models to test their viability on bankruptcy prediction. These models are made up of multiple decision trees, either in series or parallel, which increases the model performance.

Lastly, the model has been tested viable and outperformed existing methods for the prediction of bankruptcy within the energy sector. The model should also be employed in other industries such as retail or shipping for bankruptcy prediction.

References

Agrawal, K., & Maheshwari, Y. (2018). Efficacy of industry factors for corporate default prediction. *IIMB Management Review 31(1).*

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance, Vol. 23, No. 4 (Sep., 1968),* 589-609.

Anandarajan, M., Lee, P., & Anandarajan., A. (2004). Bankruptcy predication using neural networks. *Business Intelligence Techniques: A Perspective from Accounting and Finance, M. Anandarajan, A. Anandarajan and C. Srinivasan (eds.).*

Bank for International Settlements. (2006). *Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework.* Basel, Switzerland: Bank for International Settlements, Press & Communications .

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications, Volume 83,* 405-417.

Bathaee, Y. (2018). The Artificial Intelligence Black Box and the Failure of Intent and Causation. *Harvard Journal of Law & Technology, Volume 31, Number 2 Spring 2018,* 890 - 938.

Beaver, W. H. (1966). Financial Ratios As Predictors of Failure. *Vol. 4, Empirical Research in Accounting: Selected Studies 1966 (1966),* 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, Vol. 33 (Winter 2007),* 1-42.

Benítez, J. M., Castro, J. L., & Requena, I. (1997). Are Artificial Neural Networks Black Boxes? *IEEE Transactions on Neural Networks 8(5):,* 1156-1164.

Chan, S.-P., Teo, C. C., Ng, S. A., Goh, N., Tan, C., & Yap , M. (2006). Validation of various osteoporosis risk indices in elderly Chinese females in Singapore. *Osteoporosis International volume 17*, 1182–1188.

Chawla, N. V. (2005). Data Mining for Imbalanced Datasets: An Overview. *Data Mining and Knowledge Discovery Handbook*, 853-867.

Devi, R. A., & Nirmala, K. (2013). Construction of Decision Tree : Attribute Selection. *International Journal of Advancements in Research & Technology, Volume 2, Issue 4, April-2013*, 343-347.

Eisenbeis, R. A. (1977). Pitfalls in the Application of Discriminant Analysis in Business, Finance, and Economics. *Journal of Finance, American Finance Association, vol. 32(3)*, 875-900.

Ellis, C. (2019). Are Corporate Bond Defaults Contagious. *Int. J. Financial Stud. 2020, 8, 1*.

El-temtamy, O. (1995). Bankruptcy prediction : a comparative study on logit and neural networks. *Middle Tennessee State University*.

Engelmann, B., Hayden, E., & Tasche, D. (2003). Measuring the Discriminative Power. *Discussion paper Series 2: Banking and Financial Supervision, No 01/2003*.

Financial Times. (21 March, 2020). *Corporate borrowing costs soar amid default fears*. Retrieved from Financial Times: https://www.ft.com/content/2602d57c-6ad4-11ea-800d-da70cff6e4d3

FitzPatrick, P. J. (1932). *A comparison of the ratios of successful industrial enterprises with those of failed companies.* Washington: The Ohio State University.

Forbes. (9 April, 2020). *Gasoline Demand Collapses To A 50-Year Low*. Retrieved from Forbes: https://www.forbes.com/sites/rrapier/2020/04/09/gasoline-demand-collapses-to-a-50-year-low/?sh=12d5b7ab196e

Forbes. (26 March, 2020). *Rising Unemployment And Imminent Corporate Defaults Will Hurt Banks' Profitability And Capital*. Retrieved from Forbes: https://www.forbes.com/sites/mayrarodriguezvalladares/2020/03/26/rising-unemployment-and-imminent-corporate-defaults-will-hurt-banks-profitability-and-capital/?sh=1e2dfa2e654b

Forbes. (4 August, 2020). *The Increase In Corporate Bankruptcies Is Bad News For Workers And Job Seekers*. Retrieved from Forbes: https://www.forbes.com/sites/jackkelly/2020/08/04/the-increase-in-corporate-bankruptcies-is-bad-news-for-workers-and-job-seekers/?sh=26bd033169de

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 80-89.

Golbayani, P., Florescu, I., & Chatterjee, R. (2020). A comparative study of forecasting Corporate Credit Ratings using Neural Networks, Support Vector Machines, and Decision Trees. *North Am J Med Sci. Econ. Finance.*, 54.

Granström, D., & Abrahamsson, J. (2019). Loan Default Prediction using Supervised Machine Learning Algorithms. *Computer Science*.

Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med. 2013 Spring; 4(2)*, 627–635.

Hamel, L. H. (2009). *Knowledge Discovery with Support Vector Machines.* Wiley-Interscience; 1st edition (September 21, 2011).

Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Seliya, N. (2019). Examining characteristics of predictive models with imbalanced big data. *Journal of Big Data volume 6, Article number: 69.*

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, (pp. 1322-1328). Hong Kong.

Horvat, M., Jovic, A., & Ivošević, D. (2020). Lift Charts-Based Binary Classification in Unsupervised Setting for Concept-Based Retrieval of Emotionally Annotated Images from Affective Multimedia Databases. *Information (Switzerland) 11(9)*, 429.

Huang, Z., Chen, H.-c., Hsu, C.-J., Chen, W.-H., & Wu, S. (2004). Credit Rating Analysis With Support Vector Machines and Neural Networks: A Market Comparative Study. *Decision Support Systems 37(4)*, 543-558.

Kaur, P., & Gosain, A. (2018). Comparing the Behavior of Oversampling and Undersampling Approach of Class Imbalance Learning by Combining Class Imbalance Problem with Noise. *ICT Based Innovations*, 23-30.

Kim, H., & Sohn, S. (2010). Support vector machines for default prediction of SMEs based on technology credit. *European Journal of Operational Research 201(3)*, 838-846.

Kim, H., Cho, H., & Ryu, D. (2020). Corporate Default Predictions Using Machine Learning: Literature Review. *Sustainability 12(16)*, 6325.

Koh, H. (2005). *Data mining applications for small and medium enterprises.* Singapore: Centre for Research on Small Enterprise Development.

Le, T., Lee, M., Park, J., & Baik, S. (2018). Oversampling Techniques for Bankruptcy Prediction: Novel Features from a Transaction Dataset. *Symmetry 2018, 10(4)*, 79.

Levratto, N. (2013). From failure to corporate bankruptcy: a review. *Journal of Innovation and Entrepreneurship volume 2, Article number: 20 (2013).*

Liu, A. Y.-c. (2004). *The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets.* Austin: The University of Texas at Austin.

Martin, D. (1977). Early warning of bank failure: A logit regression approach. *Journal of Banking & Finance, Volume 1, Issue 3*, 249-276.

Mihalovič, M. (2016). Performance Comparison of Multiple Discriminant Analysis and Logit Models in Bankruptcy Prediction. *Economics and Sociology 9(4)*, 101-118.

Netsuite. (20 August, 2020). *Accounts Payable Turnover Ratio Defined: Formula & Examples*. Retrieved from Netsuite: https://www.netsuite.com/portal/resource/articles/accounting/accounts-payable-turnover-ratio.shtml#:~:text=The%20accounts%20payable%20turnover%20ratio,during%20a%20specified%20time%20period.

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research, Vol. 18, No. 1. (Spring, 1980)*, 109-131.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems, Volume 52, Issue 2*, 464-473.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., & Brunk, C. (1994). Reducing Misclassification Costs. *Machine Learning Proceedings 1994* (pp. 217-225). Irvine: Department of Information and Computer Science, University of California.

S&P Global Ratings. (2020). *Default, Transition, and Recovery: 2019 Annual Global Corporate Default And Rating Transition Study.* New York, United States: S&P Global.

Saad, S., Udin, Z. M., & Hasnan, N. (2014). Dynamic Supply Chain Capabilities: A Case Study in Oil and Gas Industry. *Int. J Sup. Chain. Mgt, Vol. 3, No. 2*, 70-76.

Ting, K. M. (2011). *Encyclopedia of Machine Learning.* Boston: Springer.

Towards Data Science. (11 October, 2018). *Taking the Confusion Out of Confusion Matrices*. Retrieved from Towards Data Science: https://towardsdatascience.com/taking-the-confusion-out-of-confusion-matrices-c1ce054b3d3e

U.S. Securities and Exchange Commision. (3 Febuary, 2009). *Bankruptcy: What Happens When Public Companies Go Bankrupt*. Retrieved from U.S. Securities and Exchange Commision: https://www.sec.gov/reportspubs/investor-publications/investorpubsbankrupthtm.html

Wiggins, R., Piontek, T., & Metrick, A. (2014). The Lehman Brothers Bankruptcy A: Overview. *Yale Program on Financial Stability Case Study 2014-3A-V1*.

Winakor, A. H., & Smith, R. F. (1935). Changes in the financial structure of unsuccessful industrial corporations. *University of Illinois bulletin, vol. XXXII, no. 46.*

Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *ACM SIGMOD Record, Volume 31, Issue 1*, 76-77.

Wood Mackenzie. (10 Febuary, 2020). *How will global gas and LNG markets respond to oversupply in 2020? Foresight 2020*. Retrieved from Wood Mackenzie:

https://www.woodmac.com/news/opinion/how-will-global-gas-and-lng-markets-respond-to-oversupply-in-2020/

Appendix A

Description of Financial Ratios

Field	Type of Ratio	Description	Equation
Current Ratio	Liquidity Ratios	Measure ability to pay short-term debts	$\frac{Current\ Assets}{Current\ Liability}$
Quick Ratio	Liquidity Ratios	Measures ability to pay short-term liabilities using liquid assets	$\frac{Current\ Assets - Inventory}{Current\ Liability}$
Cash Ratio	Liquidity Ratios	Measures ability to pay short-term debts using cash and cash equivalents	$\frac{Cash + Cash\ Equivalents}{Current\ Liability}$
Working Capital/Total Assets	Liquidity Ratios	Compares net liquid assets to total assets of firm	$\frac{Current\ Assets - Current\ Liabilities}{Total\ Assets}$
Cash Flow from Operations (CFO) / Sales	Cash Flow Ratios	Measures ability of firm to generate cash from sales	$\frac{Operating\ Cash\ Flow}{Sales\ Revenue}$
CFO / Total Assets	Cash Flow Ratios	Measures ability of firm to generate cash from assets	$\frac{Operating\ Cash\ Flow}{Total\ Assets}$
FCF / Current Liabilities	Cash Flow Ratios	Measures the amount of cash to repay creditors or pay dividends over short-term debt	$\frac{Free\ Cash\ Flow}{Current\ Liabilities}$
Debt Ratio	Leverage Ratios	Measures the extent of firm's leverage	$\frac{Total\ Liabilities}{Total\ Asset}$
Debt to Equity Ratio	Leverage Ratios	Measures the degree of firm	$\frac{Total\ Liabilities}{Total\ Shareholders'Equity}$
Times Interest Earned Ratio	Leverage Ratios	Measures ability to pay off short-term debts based on current income	$\frac{EBITDA}{Interest\ Expense}$
Equity Multiplier	Leverage Ratios	Measures the extent of firm's asset being financed by shareholders	$\frac{Total\ Assets}{Total\ Shareholders'Equity}$
Retained Earnings / Current Liabilities	Leverage Ratios	Measures the amount of net income left as a percentage of firm's short-term debt	$\frac{Retained\ Earnings}{Current\ Liabilities}$
Equity / Total Assets	Leverage Ratios	Measures amount of equity compared to all asset owned	$\frac{Net\ Worth}{Total\ Assets}$
Equity / Fixed Assets	Leverage Ratios	Measures amount of equity compared to long-term tangible assets owned	$\frac{Net\ Worth}{Total\ Fixed\ Assets}$
Receivables Turnover	Activity Ratios	Measures a firm's effectiveness in collecting accounts receivables	$\frac{Net\ Credit\ Sales}{Average\ Accounts\ Receivable}$

Days' Sales Outstanding (DSO) Ratio	Activity Ratios	Measures the average number of days for a firm to collect payment	$\frac{Accounts\ Receivable}{Annual\ Revenue} \times 365$
Payables Turnover Ratio	Activity Ratios	Measures the average number of times a company pays creditors	$\frac{Net\ Credit\ Purchases}{Average\ Accounts\ Payable}$
Days Payable Outstanding	Activity Ratios	Measures the average number of days that a firm takes to repay its bills	$\frac{Accounts\ Payable}{Cost\ of\ Goods\ Sold} \times 365$
Sales / Total Asset	Activity Ratios	Measures the value of firm's sales as compared to its assets	$\frac{Total\ Sales}{(\frac{Beginning\ Assets + Ending\ Assets}{2})}$
Sales Growth	Activity Ratios	Measures the rate of firm increasing revenue from sales over a fixed prior of time	$\frac{Current\ Net\ Sales - Prior\ Net\ Sales}{Prior\ net\ sales} \times 100$
Net Profit Margin	Profitability Ratios	Measures the amount of income generated from the percentage of revenue	$\frac{Revenue - Cost}{Revenue}$
Gross Profit Margin	Profitability Ratios	Measures the amount of profit before deducting costs	$\frac{Revenue - Cost\ of\ Goods\ Sold}{Revenue}$
Operating Margin Ratio	Profitability Ratios	Measures the percentage of total revenue that is made up by operating income	$\frac{Operating\ Income}{Net\ Sales}$
Return on Assets	Profitability Ratios	Measures profitability of a firm as compared to its total assets	$\frac{Net\ Income}{Total\ Assets}$
Return on Equity	Profitability Ratios	Measures profitability of a firm as compared to its total shareholder's equity	$\frac{Net\ Income}{Total\ Shareholder's\ Equity}$
Earnings per Share	Profitability Ratios	Measures the company's net income that is attributed to each outstanding share	$\frac{Net\ Income - Preferred\ Dividends}{End\ of\ Period\ Common}$
EBITDA / Total Assets	Profitability Ratios	Measures firm's efficiency in generation EBITDA	$\frac{Net\ EBITDA}{Total\ Asets}$
Equity Growth	Profitability Ratios	Measures the rate of the growth of a firm's equity	$\frac{Net\ Earning - Stock\ Dividends}{Total\ Shareholders\ Equity}$
Price to Book Ratio	Valuation Ratios	Compares the firm's market value to the book value	$\frac{Market\ Price\ per\ Share}{Book\ Value\ per\ Share}$
Price / Sales	Valuation Ratios	Measures the value of a firm for each dollar of revenue earned	$\frac{Market\ Capitalisation}{Sales}$

Appendix B

Codes used for Data Preparation/Sampling

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('cleaneddata.txt', sep='\t', thousands=",")

cols = [0, 1, 2]
df1 = df.drop(df.columns[cols], axis= 1)
#drop name, year, and index

replace = df1["Bankruptcy Status"].map({"BANKRUPT":1, "Non-bankrupt":0})
df1["Bankruptcy Status"] = replace
#recode target variable with 1 and 0

df1 = df1.dropna(thresh= int(0.7*len(df)), axis=1)
#drop features which have less than 70% completeness

df2 = df1.fillna(df.mean())
#fill missing values with mean value

features = df2.columns.tolist()[:-1]
#set features as all but "bankruptcy status"

target = df2.columns.tolist()[-1]
#set target as "bankruptcy status"
y = df2[[target]]
x = df2[features]

x.dtypes
#check data types
```

```
# Imports
import numpy as np
from imblearn.over_sampling import ADASYN
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(
    x, y, test_size=0.20, random_state=42)

# ADASYN sampling technique (on training data set only)
ads = ADASYN(sampling_strategy=0.25,random_state=44)
x_ads, y_ads = ads.fit_resample(x_train, y_train)

# Convert smote output into dataframe
df_train = pd.DataFrame(x_ads, columns=x.columns)
df_train['Bankruptcy Status'] = y_sm

df_test = pd.DataFrame(x_test,columns=x.columns)
df_test['Bankruptcy Status'] = y_test

df_test.fillna(df_test.mean(), inplace=True)

# df_test.isnull().sum()
df_train.isnull().sum()
# Check data types
df.dtypes

#exports out in csv file as tsv format
df_train.to_csv('adasyn_train.csv', encoding='utf-8', sep='\t')
df_test.to_csv('adasyn_test.csv', encoding='utf-8', sep='\t')
```