

# **ANL 488 PROJECT PROPOSAL**

**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**

## **Table of Contents**

<b>1. Introduction.....</b>	<b>2</b>
<b>2. Literature Review .....</b>	<b>5</b>
<b>2.1 Usage of Financial Ratios as Predictors of Bankruptcy .....</b>	<b>5</b>
<b>2.2 Statistical Models for Corporate Bankruptcy Prediction .....</b>	<b>5</b>
<b>2.3 Machine Learning Models for Corporate Bankruptcy Prediction .....</b>	<b>6</b>
2.3.1 Neural Network.....	6
2.3.2 Support Vector Machines .....	6
2.3.3 Decision Trees as Viable Alternative .....	7
<b>3. Data Understanding and Preparation.....</b>	<b>9</b>
<b>3.1 Dependent and Independent Variables.....</b>	<b>9</b>
<b>3.2 Issues faced with Dataset.....</b>	<b>12</b>
<b>3.3 Data Preparation.....</b>	<b>14</b>
<b>4. Proposed Modelling and Evaluation.....</b>	<b>16</b>
<b>4.1 Proposed Methodology .....</b>	<b>16</b>
<b>4.2 Proposed Evaluation.....</b>	<b>18</b>
4.2.1 Evaluation of DT models .....	18
4.2.2 Evaluating the Performance of DT against Logit Regression and NN.....	21
<b>5. Proposed Schedule .....</b>	<b>25</b>
<b>References .....</b>	<b>29</b>

## **1. Introduction**

The significant spike in corporate bankruptcy has become increasingly 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 Commission, 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, 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 (Bathae, 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 proccession 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 and fourth section provides more insights on the data to be used and the data preparation stage. The fifth section then proposes the methodology to be used for the report and the sixth section will be proposed evaluation method. Lastly, the proposed schedule of the project will be provided 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 Mihalović (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. Eltemtamy (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 machine learning models such as NN and SVM have 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, thus becoming a black box 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.

<b>Year</b>	<b>Number of Companies</b>
2011	781
2012	731
2013	683
2014	638
2015	566
2016	523
2017	499
2018	464
<b>Total</b>	<b><u>4885</u></b>

*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.

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 4,885 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.

Field	Measurement	% Complete	Valid Records	Null Value
Year	Continuous	100	4885	0
Bankruptcy S...	Flag	100	4885	0
Current Ratio	Continuous	87.963	4297	22
Quick Ratio	Continuous	84.278	4117	203
Cash Ratio	Continuous	69.969	3418	68
Cash Conv C...	Continuous	29.314	1432	3452
Working Cap...	Continuous	69.376	3389	46
Operating Ca...	Continuous	61.535	3006	118
Free Cash FI...	Continuous	62.579	3057	239
CFO / Sales	Continuous	68.925	3367	1271
Free Cash FI...	Continuous	63.071	3081	1311
CFO / Total A...	Continuous	76.52	3738	74
CFO / Short T...	Continuous	45.384	2217	2461
FCF / Current...	Continuous	87.963	4297	157
FCF / Current...	Continuous	44.524	2175	2520
Debt Ratio	Continuous	65.814	3215	48
Debt to Equit...	Continuous	67.206	3283	5
Debt to Capit...	Continuous	60.512	2956	760
Times Intere...	Continuous	75.067	3667	1181
Equity Multipl...	Continuous	77.032	3763	1122
Retained Ear...	Continuous	93.081	4547	176
Equity/Total A...	Continuous	97.216	4749	46
Equity/Fixed ...	Continuous	85.486	4176	654
Interest Expe...	Continuous	45.773	2236	1455
Inventory Tur...	Continuous	30.399	1485	3386
Days Invento...	Continuous	30.399	1485	3398
Receivables ...	Continuous	70.154	3427	547
Days Sales ...	Continuous	72.753	3554	1330
Payables Tur...	Continuous	68.557	3349	1084
Days Payabl...	Continuous	72.753	3554	1330
Fixed Asset T...	Continuous	64.463	3149	203
Working Cap...	Continuous	33.122	1618	2638
Sales/Total A...	Continuous	66.223	3235	46
Sales Growth	Continuous	71.648	3500	1380
Net Income ...	Continuous	60.86	2973	1912
Asset Turnov...	Continuous	64.463	3149	203
Asset Impair...	Continuous	1.269	62	4568

Field	Measurement	% Complete	Valid Records	Null Value
Net Profit Mar...	Continuous	74.289	3629	1249
Gross Profit ...	Continuous	72.303	3532	1350
Operating Ma...	Continuous	74.35	3632	1251
Return on As...	Continuous	95.415	4661	204
Return on Ca...	Continuous	21.331	1042	3840
Return on Eq...	Continuous	76.847	3754	1116
Earning Per ...	Continuous	72.201	3527	65
EBITDA/Total...	Continuous	78.69	3844	179
Equity Growth	Continuous	92.323	4510	359
Price to Earni...	Continuous	25.302	1236	3644
Price to Book...	Continuous	69.703	3405	1441
Dividend Pay...	Continuous	10.604	518	3608
Dividend Yiel...	Continuous	13.142	642	550
Retention Ra...	Continuous	25.937	1267	3608
Price to Cas...	Continuous	46.244	2259	2615
Price /Sales	Continuous	66.428	3245	1593

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 65% 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. Undersampling, a sampling method, will be used to reduce the size of non-bankrupt data while keeping the bankrupt data intact. This sampling method is preferred over oversampling due to overfitting concerns (Liu, 2004). As the bankrupt data is very small compared to the non-bankrupt data, oversampling of the bankrupt data will cause the model to constantly train over the same few entries, potentially causing overfitting of the model. Furthermore, as the study focuses on the prediction of bankrupt data rather than non-bankrupt data, the amount of bankrupt data should remain intact. Thus, despite the data loss in the form of non-bankrupt data, the undersampling method is chosen to increase the proportion of bankrupt data.

### 3.3 Data Preparation



The data preparation process will be done entirely on the IBM SPSS Modeller. Only fields with more than 70% complete data will be used for the model. Table 3 below shows the remaining 30 data fields.

<b>Liquidity Ratios</b>	<b>Cash Flow Ratios</b>	<b>Leverage Ratios</b>	<b>Activity Ratios</b>	<b>Profitability Ratios</b>	<b>Valuation Ratios</b>
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 data types as flag for the dependent variable while the rest of the independent variables (financial ratios) are set as continuous.

The sampling step is done using a Balance Node in IBM SPSS, reducing the size of the non-bankrupt entries to 20% of its original size. The initial proportion of bankrupt cases are as shown in Figure 2. The proportion of bankruptcy cases in the final dataset is shown in Figure 3 below.

Value /	Proportion	%	Count
BANKRUPT		0.98	48
Non-bankrupt		99.02	4837

*Figure 2: Proportion of Bankrupt Cases prior to Sampling*

Value /	Proportion	%	Count
BANKRUPT		4.6	48
Non-bankrupt		95.4	995

*Figure 3: Proportion of Bankrupt Cases after Sampling*

The cleaned and sampled data will then be partitioned 80:20, with 80% being used as the training dataset and 20% used as testing dataset.



## **4. Proposed Modelling and Evaluation**

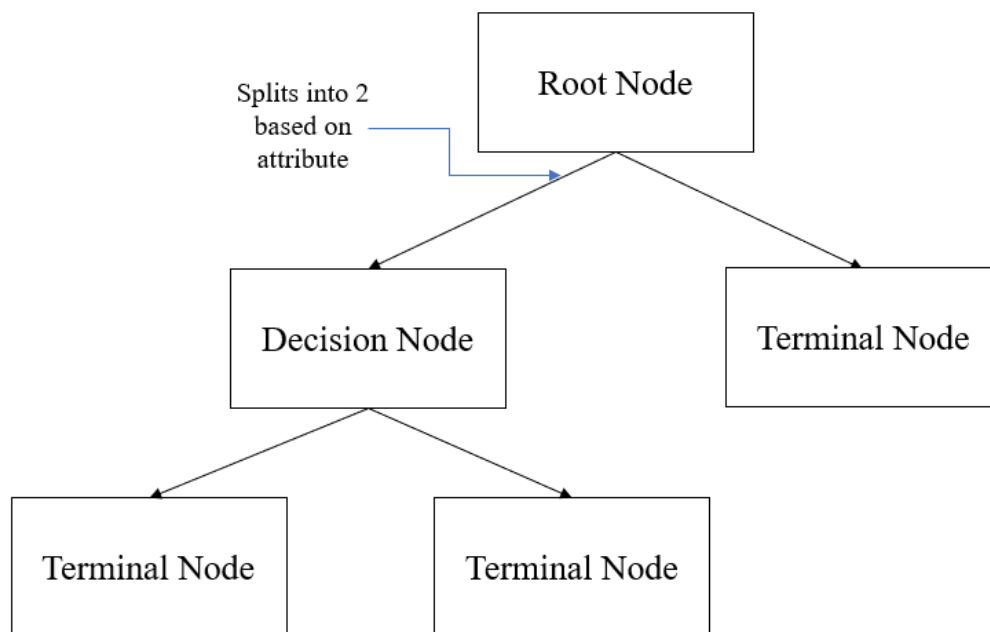
### **4.1 Proposed Methodology**

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.

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 4 below.



*Figure 4: 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 will be 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.

## **4.2 Proposed Evaluation**

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 Accuracy Ratio and the Receiver Operating Characteristic (ROC) curve.

### **4.2.1 Evaluation of DT models**

Two main methods are proposed to evaluate the performance of the DT models created. They are the confusion matrix and the lift charts.

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.

	<b>Actual (True or False)</b>	
<b>Predicted (Positive or Negative)</b>	True Positive	False Positive (Type I Error)
	False Negative (Type II error)	True Negative

*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, sensitivity, 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). Sensitivity measures solely the extent of correct positive prediction (TP) as compared to all actual positive prediction (TP + FN), while specificity measures the extent of correct 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.

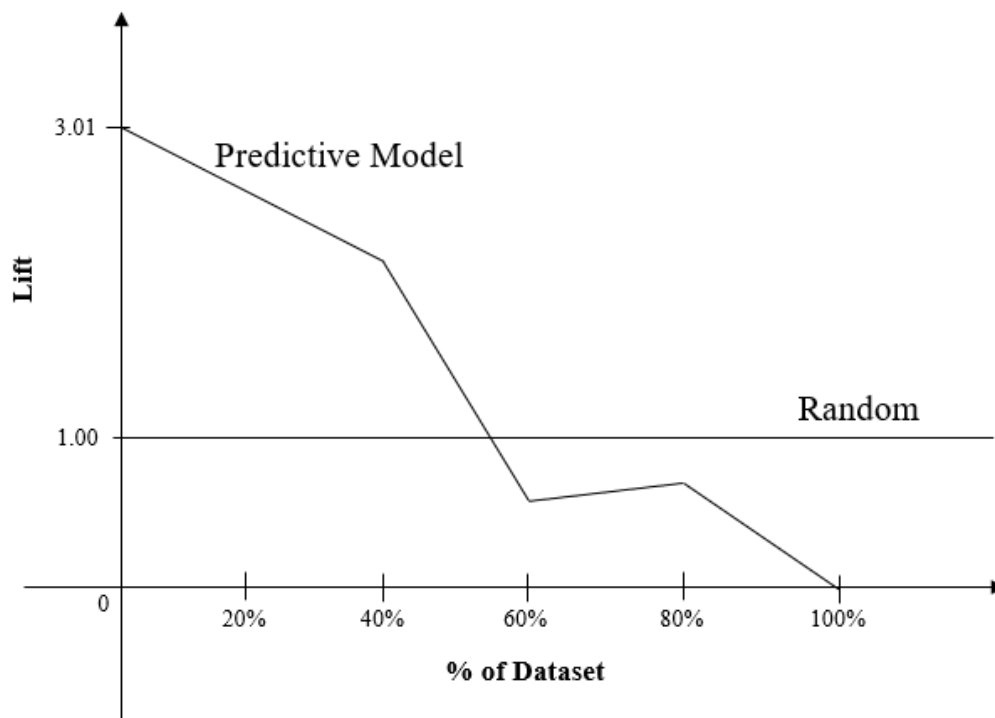
These metrics are summarised in Figure 6 below.

	<b>Formula</b>
<b>Accuracy</b>	$\frac{TP + TN}{TP + TN + FP + FN}$
<b>Misclassification</b>	$\frac{FP + FN}{TP + TN + FP + FN}$
<b>Sensitivity</b>	$\frac{TP}{TP + FN}$
<b>Specificity</b>	$\frac{TN}{TN + FP}$
<b>Precision (Positive)</b>	$\frac{TP}{TP + FP}$
<b>Precision (Negative)</b>	$\frac{TN}{TN + FN}$

*Figure 6: Confusion Metrics*

Since this paper focuses on the prediction of corporate bankruptcy and that the positive event is set as “Bankrupt”, the evaluation of proposed models should focus heavily on the accuracy, sensitivity, and the positive precision. There should be more emphasis on the accurate prediction of bankrupt cases as compared to non-bankrupt as bankrupt cases are rarer. 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 these confusion metrics to evaluate the performance of the model.

The second 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 7 below shows an example of the lift chart.



*Figure 7: 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.

#### **4.2.2 Evaluating the Performance of DT against Logit Regression and NN**

Two evaluation methods will be used to evaluate the performance of DT against the Logit model and NN model. They are the Accuracy Ratio (AR) and the Area Under the ROC curve (AUC)

The Accuracy Ratio is one of the most popular approach used for predictive model evaluation in the credit rating industry. For predictive model, the AR is used as the summary statistic of the model's Cumulative Accuracy Profile (CAP). The CAP curve serves as a visual representation of the discriminative power of a model (Engelmann, Hayden, & Tasche, 2003).

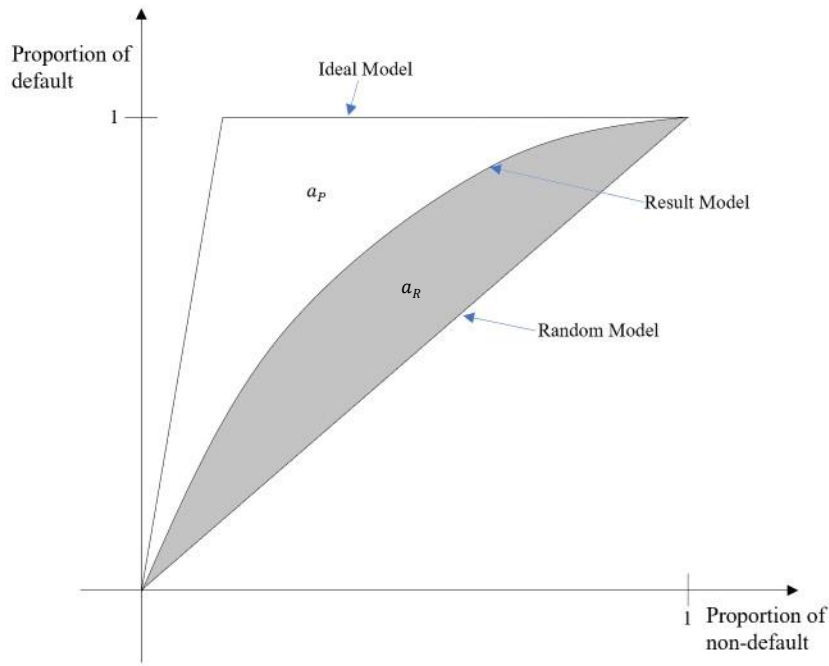


Figure 8: Example of CAP Curve

The concept of CAP is illustrated in Figure 8 above. An ideal model will be able to assign the lowest scores to the firms with “bankrupt”, thus the proportion of “bankrupt” will rapidly increase to 1 and stay there. A random model with no discriminatory power will provide a linear function that do not actively separate the firms based on their chance of bankruptcy as shown in the CAP model. The result model’s function will fall in between the perfect model and the random model as shown as the curve in between. The closer the curve is to the ideal model’s line, the better it is. In contrast, the closer it is the random model, the weaker the result model is. The area between the ideal model line and the result model line is defined as  $a_P$  while the area between the result model and the random model is defined by  $a_R$ . The AR is defined as the ratio of  $a_R$  to  $a_P$ , which is representative of the model’s discriminating power.

$$AR = \frac{a_R}{a_P}$$

An AR will have values from 0 to 1, with 1 being the strongest model and 0 being the weakest in discriminating power.

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 9 below shows an example of the ROC curve.

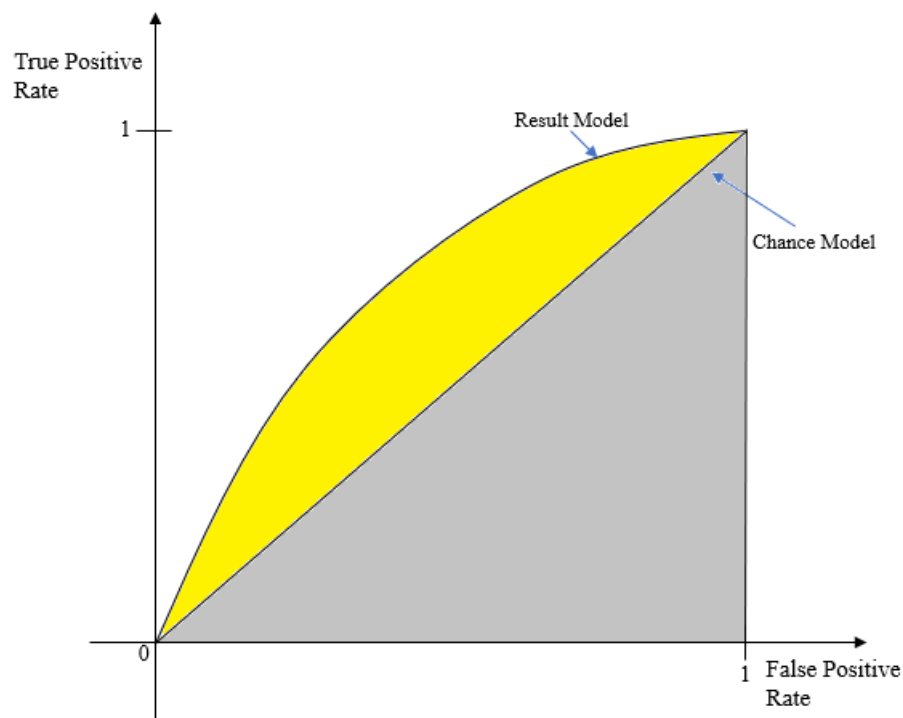


Figure 9: Example of ROC Curve

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 9 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.

## 5. Proposed Schedule

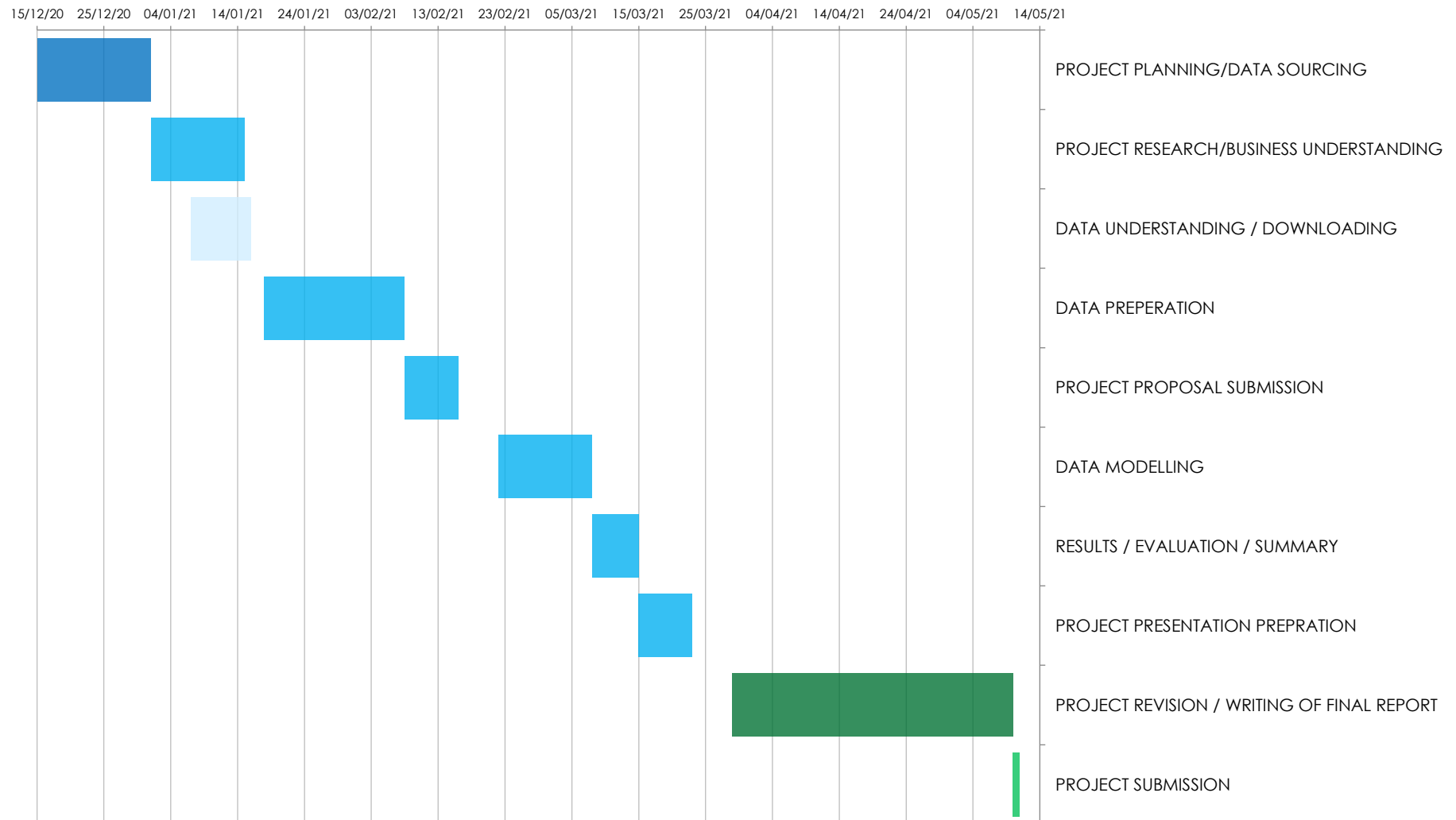
### FYP Project Schedule

PROJECT TITLE	START DATE	PROJECT DURATION
Prediction of Corporate Default with Decision Trees	21/12/20	in days
PROJECT MANAGER	END DATE	141
Chua Poh Chai	10/05/21	

WBS NO.	TASK NAME	STATUS	START DATE	END DATE	DURATION in days	Completed on Date	COMMENTS	STATUS
<b>1</b>	<b>PROJECT PLANNING/DATA SOURCING</b>	In Progress	15/12/20	31/12/20	17	04/01/21		Not Started
1.1	Drafting of Project Plan / Schedule	Complete	15/12/20	31/12/20	17	27/01/21		In Progress
1.2	Data Sourcing (If able to source from 3rd Party)	On Hold	15/12/20	30/12/20	16			Complete
1.2.1	— Requesting for Data	On Hold	15/12/20	25/12/20	11		Requested from MAS but failed, moving to option 2	On Hold
1.2.2	— Submission of Data Request Form	On Hold	25/12/20	27/12/20	3			
1.2.3	— Preparation of NDA	On Hold	27/12/20	31/12/20	5			
1.3	Data Sourcing (If preparing from FactSet/Bloomberg)	Complete	25/12/20	31/12/20	7	04/01/21	Data will be sourced from FactSet	Complete
1.3.1	— Source for Data on FactSet	Complete	25/12/20	30/12/20	6	04/01/21		On Hold
1.3.2	— Determine Duration of Data Needed	Complete	30/12/20	02/01/21	4	25/01/21	2011 - 2018	
<b>2</b>	<b>PROJECT RESEARCH/BUSINESS UNDERSTANDING</b>	Complete	01/01/21	14/01/21	14	20/01/21		
2.1	Topic Research	Complete	15/12/20	30/12/20	16	25/12/20	Researching on credit risk models, corporate distress, financial ratios, etc.	
2.2	Sector Research	Complete	15/12/20	25/12/20	11	23/12/20	Determine sectors which are most affected	
2.3	Identification of Business Problem	Complete	25/12/20	28/12/20	4	28/12/20		
2.4	Identification of Data Mining Problem	Complete	28/12/20	31/12/20	4	31/12/20		
2.5	Literature Review	Complete	01/01/21	07/01/21	7	16/01/21		
2.5.1	— Technique Research	Complete	01/01/21	03/01/21	3	14/01/21		

2.5.2	— Determining Model	Complete	04/01/21	07/01/21	4	16/01/21	
2.6	Drafting of Project Proposal	Complete	07/01/21	15/01/21	9	20/01/21	
2.6.1	— Drafting of Introduction Paragraph	Complete	07/01/21	10/01/21	4	18/01/21	
2.6.2	— Drafting of Literature Review	Complete	11/01/21	14/01/21	4	20/01/21	
<b>3</b>	<b>DATA UNDERSTANDING / DOWNLOADING</b>	<b>Complete</b>	<b>07/01/21</b>	<b>15/01/21</b>	<b>9</b>	<b>23/01/21</b>	BEFORE WEEK 1
3.1	Determine Data to be Downloaded	Complete	07/01/21	13/01/21	7	15/01/21	
3.1.1	— Determine Data Fields Needed	Complete	07/01/21	09/01/21	3	10/01/21	
3.1.2	— Determine Data Types of Fields	Complete	09/01/21	10/01/21	2	10/01/21	Continuous for independent/flag for depend
3.1.3	— Determine Quantity of Data Needed	Complete	10/01/21	10/01/21	1	10/01/21	
3.1.4	— Finalise Dependent and Independent Variables	Complete	11/01/21	11/01/21	1	11/01/21	Financial Ratios
3.1.5	— Determine Duration of Data Needed	Complete	12/01/21	13/01/21	2	13/01/21	15 years data
3.1.6	— Determine Companies to be Chosen for Analysis	Complete	12/01/21	13/01/21	2	13/01/21	
3.2	Downloading of Data	Complete	14/01/21	15/01/21	2	23/01/21	Did in tandem to Project Research, took more than 2 days
<b>4</b>	<b>DATA EXPLORATION</b>	<b>Complete</b>	<b>15/01/21</b>	<b>17/01/21</b>	<b>3</b>	<b>24/01/21</b>	BEFORE WEEK 1
4.1	Descriptive Statistics of Variables	Complete	15/01/21	16/01/21	2	24/01/21	Delayed due to late completion of download
4.2	Review of Data Quality	Complete	16/01/21	17/01/21	2	24/01/21	Delayed due to late completion of download
<b>5</b>	<b>DATA PREPERATION</b>	<b>Complete</b>	<b>18/01/21</b>	<b>07/02/21</b>	<b>21</b>	<b>04/02/21</b>	BEFORE WEEK 1 (until week 3)
5.1	Data Cleaning	Complete	18/01/21	31/01/21	14	27/01/21	Finished earlier than expected
5.2	Data Transformation	Complete	01/02/21	03/02/21	3	29/01/21	*to review steps upon receiving data source
5.2.1	— Data Reduction/Sampling	Complete	01/02/21	03/02/21	3	29/01/21	
5.3	Preparation of Test/Training Set	Complete	04/02/21	05/02/21	2	01/02/21	To determine % of test/train set: 80/20
5.4	Drafting of Data Understanding & Preparation Chapter in Project Proposal	Complete	06/02/21	07/02/21	2	04/02/21	

<b>6</b>	<b>PROJECT PROPOSAL SUBMISSION</b>	<b>In Progress</b>	<b>08/02/21</b>	<b>15/02/21</b>	<b>8</b>		Due on WEEK 4 (15/02/21)
6.1	Finalise Introduction/Literature Review	In Progress	08/02/21	09/02/21	2		
6.2	Finalise Proposed Modeling/Evaluation	In Progress	09/02/21	11/02/21	3		
6.3	Finalise Data Understanding & Prepration Chapter in Project Proposal	In Progress	12/02/21	13/02/21	2		
6.4	Finalise Proposed Schedule*	In Progress	13/02/21	14/02/21	2		
6.5	Submit Proposal	In Progress	14/02/21	15/02/21	2		
<b>7</b>	<b>DATA MODELLING</b>	<b>Not Started</b>	<b>22/02/21</b>	<b>07/03/21</b>	<b>14</b>		WEEK 5 AND 6
7.1	Model Selection	Not Started	25/02/21	28/02/21	4		
7.2	Model Creation	Not Started	26/02/21	03/03/21	6		Coding of Model
7.3	Model Optimisation	Not Started	25/02/21	06/03/21	10		
<b>8</b>	<b>RESULTS / EVALUATION / SUMMARY</b>	<b>Not Started</b>	<b>08/03/21</b>	<b>14/03/21</b>	<b>7</b>		WEEK 7
8.1	Evaluation of Model Efficiency	Not Started	08/03/21	09/03/21	2		
8.2	Selection of Model	Not Started	09/03/21	10/03/21	2		
8.3	Deployment of Model	Not Started	10/03/21	11/03/21	2		
8.4	Drafting Results Section	Not Started	11/03/21	12/03/21	2		
8.5	Drafting Summary	Not Started	13/03/21	14/03/21	2		
<b>9</b>	<b>PROJECT PRESENTATION PREPRATION</b>	<b>Not Started</b>	<b>15/03/21</b>	<b>22/03/21</b>	<b>8</b>		WEEK 8 - 9, WEEK 9 PRESENTATION
9.1	Preperation of Presentation Slides	Not Started	15/03/21	16/03/21	2		To be used in conjunction with presentation
9.2	Preperation of Presentation Speech	Not Started	16/03/21	18/03/21	3		Content to be covered / flow
9.3	Q&A Preparation	Not Started	18/03/21	18/03/21	1		Determine possible questions to be asked and find answers
9.4	Rehearsal	Not Started	19/03/21	20/03/21	2		**Exact date to be confirmed
9.5	Final Revision of Slides/Speech	Not Started	21/03/21	22/03/21	2		
<b>10</b>	<b>PROJECT REVISION / WRITING OF FINAL REPORT</b>	<b>Not Started</b>	<b>29/03/21</b>	<b>09/05/21</b>	<b>42</b>		WEEK 10 - 15
10.1	Finalising Literature Review	Not Started	29/03/21	04/04/21	7		
10.2	Finalising Data Understanding / Preparation	Not Started	05/04/21	11/04/21	7		
10.3	Finalising Modelling / Evaluation	Not Started	12/04/21	18/04/21	7		
10.4	Finalising Conclusion / Recommendation	Not Started	19/04/21	25/04/21	7		
10.5	Finalise Abstract	Not Started	26/04/21	02/05/21	7		
10.6	Finalise Report	Not Started	03/05/21	09/05/21	7		
<b>11</b>	<b>PROJECT SUBMISSION</b>	<b>In Progress</b>	<b>10/05/21</b>	<b>10/05/21</b>	<b>1</b>		WEEK 16



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

- 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/rpapier/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>

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).

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.



- 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.
- 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.
- 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.
- 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/>