Capstone 1 Final Report: Bankruptcy in Poland

1: Background

Poland has historically reported one of the largest number of companies that have filed for bankruptcy on an annual basis in Europe. Between 2016 and 2017, Polish bankruptcy cases have averaged well over 4000, placing them in line for second place with the United Kingdom but well behind Switzerland with over 10,000 bankruptcy filings. Bankruptcies not only hurt the company itself – employees needing to find new jobs, often unexpectedly with short amounts of advanced notice, but also the community at large, such as companies who rely on the goods and services the bankrupt company provide and residents who consume these products and services. Governments and companies must develop a way to use a company's data to predict whether that company will file for bankruptcy to allow the imposition of preventative strategies and intervention before filing occurs.

2: General Information About Collected Data

The dataset that will be used is a dataset representing 64 financial metrics from the 2012 financial statements of 5910 Polish companies. It is a clean dataset that was accessed through the UCI Machine Learning Repository in the form of a text file with no categorical variables. Based on those statements, 410 companies filed for bankruptcy in 2013 representing almost 7% of the companies in the study. Some of the 64 financial metrics that will be used as attributes in this study include the net profit/total assets, earnings before interest and taxes (EBIT)/total assets, working capital, net profit/sales, total assets/total liabilities, equity/fixed assets, gross margin, and profit margin. It is not determined at this time if all the attributes are needed to make a prediction but optimizing the tradeoff between a low bias and a low variance will be considered.

3: Data Preprocessing

The data file for this project is "5year.arff". The file extension, .arff, stands for Attribute Relation File Format in which each column of data is an attribute that may be a numeric value, a nominal specification, a string, or a date. In this file, the 64 features are assigned as numeric values, and the target is classified as a nominal value of either 0 or 1 specifying that a Polish company did not file (0) or filed (1) for bankruptcy.

To successfully convert the .arff file to a dataframe, a collection of Python libraries is utilized. Specifically, the requests, zipfile, io, and scipy.io.arff.loadarff libraries allow the "5year.arff" file to be converted into a dataframe. Before handling the missing values, the data was examined for the presence of outliers, in which, several were present. The outliers that fell outside the 95% confidence interval (outside $\mu \pm 2\sigma$) were removed. These values were transformed into Nan values that are addressed below.

There are several missing values in the data. Since the features are numeric financial ratios, a simple way to handle the missing data is to first, take the mean value of the observations per feature. Then, assign those means to the missing values to preserve the full set of observations.

4: Exploratory Data Analysis

The next step of exploring the data was useful to help answer some questions about the observations. Some of the questions that were considered were 1.) has the bankruptcy rate in Poland statistically increased over the five-year period from 2008-2012, 2.) if so, what factors may be causing the increase, and 3.) could accurate reporting of financial records provide significant information to help businesses continue operations?

Taking the ratios of bankrupt companies to companies still in business for each of these five years point to an increase in bankruptcy of Polish companies for that time frame (Figure 1). Looking at the correlation distribution by year of how a company's bankruptcy filing decision is correlated with the reported financial attributes, it is evident that there is relatively low correlation for the first four years (Figure 2). The tails of these distributions (showing the highest magnitude of correlation) are below 0.15. However, the 2012 correlation distribution shows increased correlation magnitudes (around 0.30), albeit, still relatively low. Thus, what features are directly related to this increase in correlation magnitude?

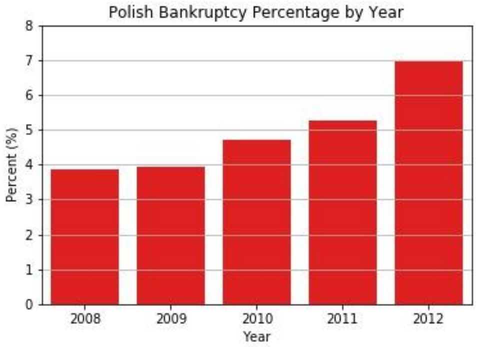


Figure 1. Polish bankruptcy percentage by year from 2008-2012.

Figure 3 shows the attributes of the ten largest correlation magnitudes. Interestingly, 80% of those attributes are inversely proportional to total assets, while the other 20% are inversely proportional to sales. This result highlights the importance of those two metrics in predicting if a company will file for bankruptcy. The attribute names of these financial ratios are shown in Figure 4.

5: Further Statistical Analysis

After taking an initial look at the summary statistics, some questions still existed about the statistical validity of some of the features, such as their correlations and means. Further analysis and discussion

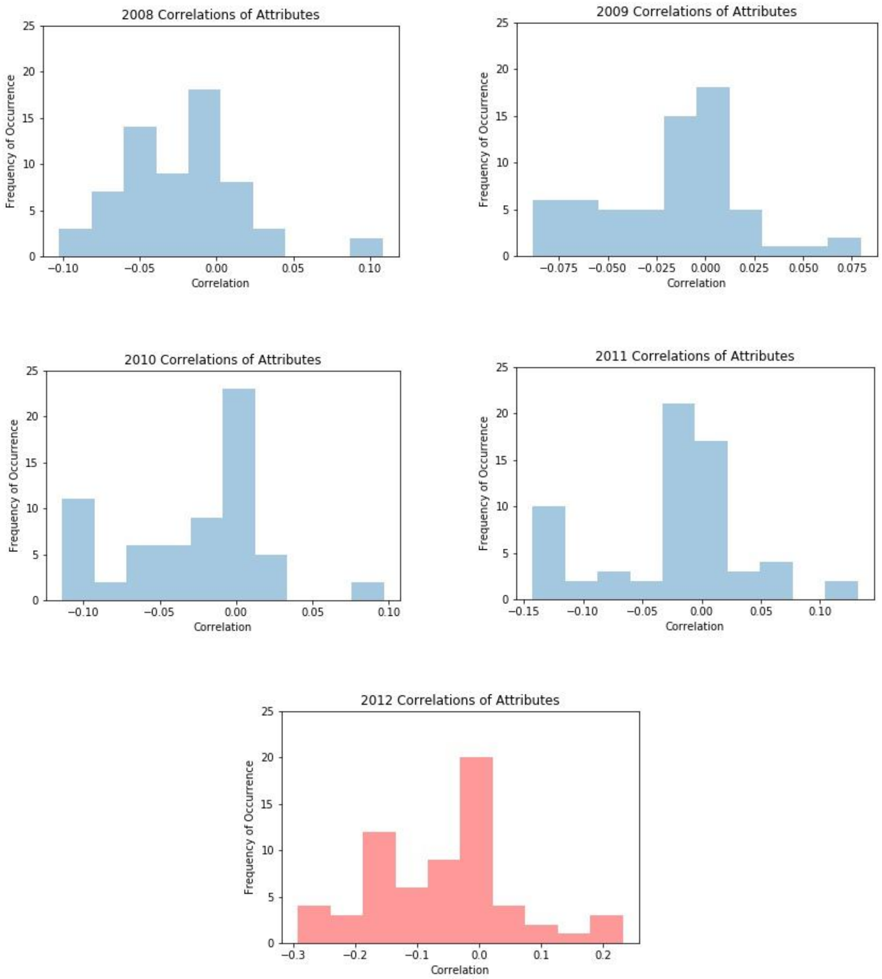


Figure 2. Correlation distributions for each year between 2008-2012.

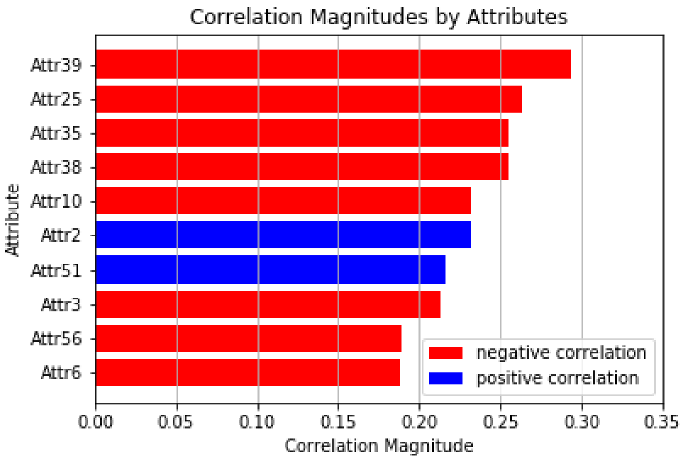
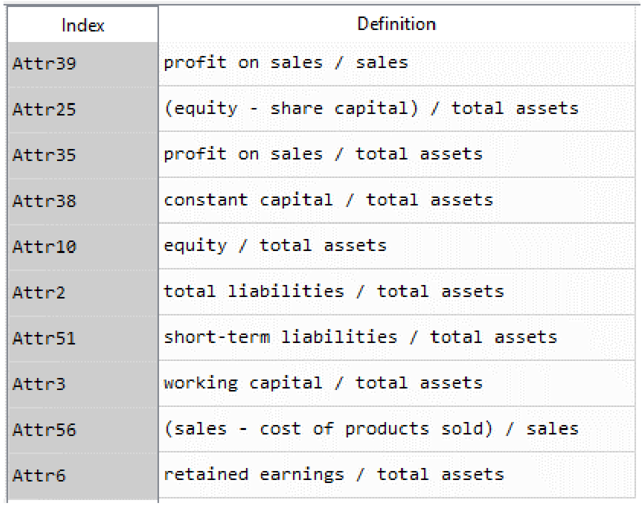


Figure 3. Ten financial attributes with highest correlation magnitude.



Index	Definition
Attr39	profit on sales / sales
Attr25	(equity - share capital) / total assets
Attr35	profit on sales / total assets
Attr38	constant capital / total assets
Attr10	equity / total assets
Attr2	total liabilities / total assets
Attr51	short-term liabilities / total assets
Attr3	working capital / total assets
Attr56	(sales - cost of products sold) / sales
Attr6	retained earnings / total assets

Figure 4. Attribute name of ten financial attributes with highest correlation magnitude.

about the correlations of the features will provide some thoughts on the potential of multicollinearity in the data set. Afterwards, a hypothesis is tested with the bootstrap inference method. This test looks to determine if the mean of the Attr39 feature (profit on sales divided by sales) is the same when this feature is split between companies who filed and did not file for bankruptcy. Early indications point to the mean of the feature when the company filed for bankruptcy being much different than the feature's mean when the company did not file. However, the sample size for the feature when the company filed for bankruptcy is much less than the other subgroup.

The correlation matrix of features (independent variables) was generated, and a data frame of strong correlations (correlation magnitude ≥ 0.9) was developed. Of the 2016 correlation pairs of the 64 features (64 choose 2), there were only 26 pairs of these strong correlations. This could present a problem in future models due to the existence of collinearity. The presence of collinearity can cause overfitting of models in regression analysis and type II errors. Some of pairs had a correlation magnitude of 1, which was later determined to be pairs of the same data. One of the features of each corresponding pair will be removed in the prediction model as feature duplicity is highly discouraged.

However, some of the pairs had correlation magnitudes above 0.95. These will probably be removed in the future as well for the prediction model, but further analysis of the feature and its significance as a financial metric will be conducted before this decision is finalized.

In the last section, the top ten features with the largest correlation magnitudes with respect to the target were identified. Attr39 (profit on sales divided by sales) had the largest correlation magnitude of 0.29. Although these correlation magnitudes are relatively small, it is important to understand if the means of this feature are the same or different regardless of whether a company filed (Attr39yes) or did not file (Attr39no) for bankruptcy. The difference of the subgroup empirical means (0.053 for Attr39yes and -0.111) was 0.164, suggesting that the means are different. Thus, the null hypothesis is that the means of the two subgroups are different, while the alternative hypothesis is that the means of the two subgroups are the same. After the means of the subgroups were shifted to the concatenated mean, they were replicated 10000 times through a bootstrap inference method, and a distribution of the difference of the bootstrapped means was developed. The corresponding p-value was 0.40, which means that the probability of seeing a mean greater than 0.164 is 40%, and therefore, the null hypothesis is accepted, and the means of these two groups are, in fact, different.

6: Machine Learning Model Analysis

In this section, a few algorithms are explored to determine which is the most suitable for predicting if a bank will file for bankruptcy. These algorithms are the standard logistic regression, the popular extreme gradient boosting (XGBoost), and LightGBM. Logistic regression is one of the most basic (but effective) machine learning approaches that describes data and attempts to explain the relationship between one dependent variable and one or more nominal, ordinal, or ratio variables that are independent. XGBoost uses efficient gradient boosting decision trees to train a simple model on the data before using the first error of the model as a feature to build successive models. This process reduces the model's error because each successive model's error improves from the previous model's weaknesses. LightGBM starts with the modeling approach of XGBoost but expands on its effectiveness by implementing a gradient-based one-side sampling to ascertain as much information as possible from the data in the shortest amount of time. LightGBM achieves this by including alpha, a hyper parameter that retains the most informative samples while duplicating the least informative samples to help preserve the original distribution.

Before using the algorithms to predict if a company would file for bankruptcy, the dataset has a total of 64 predictors, of which, 17.5% of them are highly correlated with at least one other predictor at a correlation factor above 0.95. These highly correlated predictors cause multicollinearity that can result in inaccurate or poorly estimated coefficients and inflated standard errors of these coefficients, which must be handled before fitting the data to a model. One of the most widely used methods of determining and eliminating multicollinearity is the computation of the variable inflation factor (VIF). VIFs that are greater than 5 may signal the presence of multicollinearity. For this dataset, the VIF was calculated for each predictor by regressing that predictor on the other 63 predictors to obtain a R-squared for each model. After calculating the VIFs, the largest VIF was removed from the dataset, and each predictor was regressed on the other 62 predictors to obtain an updated R-squared for each model. This iterative process was repeated until the R-squared of the condensed model was less than 0.8, corresponding to a VIF less than 5. From this model, 33 predictors were removed from the model leaving 31 attributes for predicting if a bank would file for bankruptcy.

The data was split into a training set and a test set where the test set comprised 30% of the data and the training set contained the other 70%. 50 random samples of the training and test sets were created from the data. The first algorithm implemented for classification was a standard logistic regression. The confusion matrix and the receiver operating characteristic (ROC) for the model are presented in Figures 5 and 6, respectively.

		Predicted	
		0	1
Actual	0	1631	14
	1	116	12

Figure 5. Confusion matrix of test set data for standard logistic regression model.

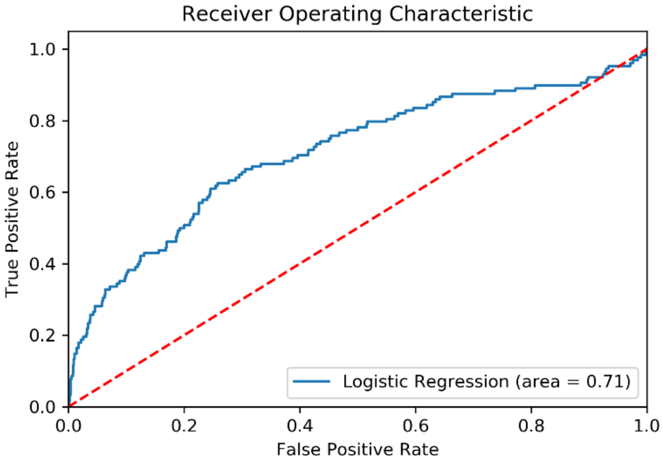


Figure 6. Receiver operating characteristic of test set data for standard logistic regression model.

Next, the data was applied to the XGBoost algorithm to generate a model for prediction. The XGBoost algorithm contains hyper parameters that should be defined before the data fitting takes place. In particular, the learning rate, the maximum depth of the trees, and the number of estimators were optimized by 10-fold cross-validation to determine the values of the parameters that produce the highest prediction accuracy on the training set before the model is created to predict the test set. The GridSearchCV function is an automated means of performing this optimization. For the cross-validation of the hyper parameters, the following ranges of the input variables were fed to GridSearchCV: learning rate = [0.0001, 0.001, 0.01, 0.1], max depth of the trees = [4, 5, 6, 7, 8, 9, 10], and the number of estimators = [10, 20, 30, 40, 50]. Using 10-fold cross-validation, the average optimized parameters are the learning rate = 0.1, the max depth = 7, and the number of estimators = 50. The confusion matrix and the receiver operating characteristic (ROC) for this model are presented in Figures 7 and 8.

		Predicted	
		0	1
Actual	0	1626	19
	1	86	42

Figure 7. Confusion matrix of test set data for XGBoost classification model.

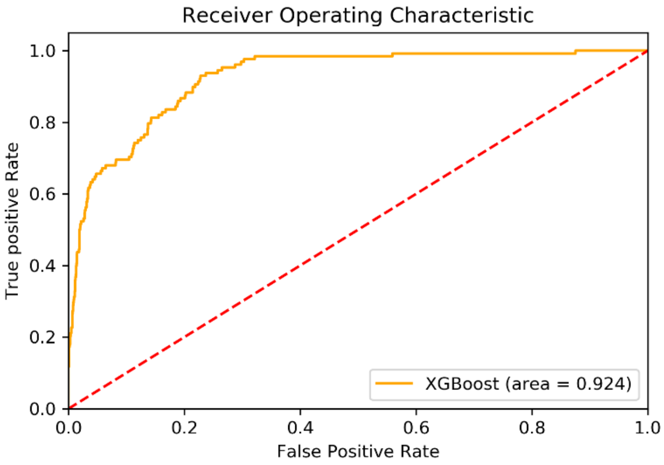


Figure 8. Receiver operating characteristic of test set data for XGBoost classification model.

Finally, the data was applied to the LightGBM algorithm to generate a model for prediction. Like XGBoost, the LightGBM algorithm contains hyper parameters that should be defined before the data fitting takes place. In particular, the learning rate, the maximum depth of the trees, and the number of estimators were optimized by 10-fold cross-validation to determine the values of the parameters that produce the highest prediction accuracy on the training set before the model is created to predict the test set. The GridSearchCV function was used to perform this optimization. For the cross-validation of the hyper parameters, the following ranges of the input variables were fed to GridSearchCV: learning rate = [0.0001, 0.001, 0.01, 0.1], max depth of the trees = [4, 5, 6, 7, 8, 9, 10], and the number of estimators = [10, 20, 30, 40, 50]. Using 10-fold cross-validation, the average optimized parameters are the learning rate = 0.1, the max depth = 4, and the number of estimators = 50. The confusion matrix and the receiver operating characteristic (ROC) for this model are presented in Figures 9 and 10.

		Predicted	
		0	1
Actual	0	1627	18
	1	75	53

Figure 9. Confusion matrix of test set data for LightGBM classification model.

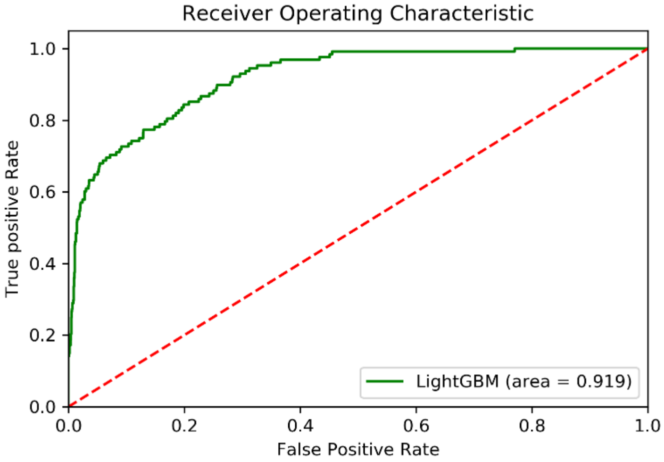


Figure 10. Receiver operating characteristic of test set data for LightGBM classification model.

Table 1 shows the average summary statistics from the three machine learning algorithms considered in this study. From these results, the average accuracy of prediction, precision score and recall score are superior for the LightGBM model, but the result of the XGBoost model very close to those of the LightGBM model. One point to note is that a model evaluator may merely look at the standard logistic regression accuracy and determine that it is just as good as the XGBoost and LightGBM models, but this is under the assumption of a 50% threshold. The ROC curves show that at different thresholds between 0 and 100%, the XGBoost and LightGBM models perform much better than the standard logistic regression model. A review of the definitions of the features that were maintained in the reduced dataset (after multicollinearity was removed) show that the attributes with financial ratios that are inversely proportional to total assets and sales represent 13 of the 31 features. Earlier, in the exploratory data analysis section, it was shown that the financial ratios that are inversely proportional to total assets and sales represented the top ten ratios with the highest correlation to the target variable. This shows the importance of these features in predicting a bank's decision to file for bankruptcy.

Table 1. Average summary statistics of test set predictions for all three machine learning algorithms (assuming a 50% threshold except for AUC)

	Standard Regression	XGBoost	LightGBM
Average Accuracy of Prediction	0.93	0.94	0.95
Precision (Bankruptcy = 0 or False)	0.93	0.95	0.96
Precision (Bankruptcy = 1 or True)	0.46	0.69	0.75
Recall (Bankruptcy = 0 or False)	0.99	0.99	0.99
Recall (Bankruptcy = 1 or True)	0.09	0.33	0.41
Misclassification Rate	0.07	0.06	0.05
Area Under the Curve (AUC)	0.71	0.92	0.92

The machine learning analysis was performed again using the three algorithms on scaled predictors. The observations for each predictor were scaled using a standard scaling process of subtracting the mean of the feature followed by scaling the feature to a unit variance. Table 2 compares the average summary statistics of the three algorithms using feature scaling with the unscaled features of Table 1. The scaled dataset shows a great improvement in the AUC score for the standard regression model, while the performance metrics for the XGBoost and LightGBM models do not show significant changes. The AUC score shows that LightGBM with scaled features has the best performance when the thresholds are varied.

Table 2. Average summary statistics of test set predictions (unscaled vs scaled features) for all three machine learning algorithms (assuming a 50% threshold except for AUC)

	Unscaled			Scaled		
	Standard Regression	XGBoost	LightGBM	Standard Regression	XGBoost	LightGBM
Average Accuracy of Prediction	0.93	0.94	0.95	0.92	0.94	0.95
Precision (Bankruptcy = 0 or False)	0.93	0.95	0.96	0.94	0.95	0.95
Precision (Bankruptcy = 1 or True)	0.46	0.69	0.75	0.44	0.75	0.77
Recall (Bankruptcy = 0 or False)	0.99	0.99	0.99	0.99	0.99	0.99
Recall (Bankruptcy = 1 or True)	0.09	0.33	0.41	0.14	0.34	0.39
Misclassification Rate	0.07	0.06	0.05	0.08	0.06	0.05
Area Under the Curve (AUC)	0.71	0.92	0.92	0.85	0.92	0.93

Finally, the standard logistic regression summary statistics are more closely examined by looking at the coefficients of each predictor and their statistical significance. Using a 95% confidence interval, nine of the 31 predictors in the reduced dataset are significantly different from zero (p-value greater than 0.05). Models were generated again, this time, using the nine-feature dataset and the three machine learning algorithms. Table 3 shows the average summary statistics for the models using a 31-feature data and those using a nine-feature dataset. (Note that the 31-feature results are the same as those in Table 1 and the unscaled portion of Table 2.) The results show that the accuracy, precision, and recall decrease with the smaller number of attributes, but the AUC score improves for the standard logistic model with reduced features. The definitions of the features in the nine-feature dataset) show that the attributes with financial ratios that are inversely proportional to total assets and sales represent eight of the nine features (working capital was the remaining attribute), thus highlighting the importance of the inversely proportional relationship of the ratio of total assets and sales in determining whether a company will file for bankruptcy.

Table 3. Average summary statistics of test set predictions (31-feature vs 9-feature dataset) for all three machine learning algorithms (assuming a 50% threshold except for AUC)

	31-feature			9-feature		
	Standard Regression	XGBoost	LightGBM	Standard Regression	XGBoost	LightGBM
Average Accuracy of Prediction	0.93	0.94	0.95	0.92	0.94	0.94
Precision (Bankruptcy = 0 or False)	0.93	0.95	0.96	0.93	0.94	0.95
Precision (Bankruptcy = 1 or True)	0.46	0.69	0.75	0.38	0.69	0.67
Recall (Bankruptcy = 0 or False)	0.99	0.99	0.99	0.99	0.99	0.99
Recall (Bankruptcy = 1 or True)	0.09	0.33	0.41	0.09	0.24	0.27
Misclassification Rate	0.07	0.06	0.05	0.08	0.06	0.06
Area Under the Curve (AUC)	0.71	0.92	0.92	0.81	0.88	0.89

7: Conclusion

Three machine learning models were explored to predict if a bank would file for bankruptcy using a dataset that represented financial ratios from the 2012 financial statements of 5910 companies. The ratios that are inversely proportional to total assets and sales have the most significance to the prediction. Using a standard logistic regression with a 50% threshold, the model correctly predicts between 92-93% of the bankruptcy decisions of 1773 companies that were used to test the model. An AUC score of 0.85 shows that the model performs very well for changing thresholds. The XGBoost and LightGBM models perform better than the standard logistic regression model with accuracies between 94-95% and AUC scores greater than 0.92. Using these machine learning models coupled with the statistical analysis of the financial ratios has provided valuable information that can allow government bodies to create preventative strategies and intervention steps before a bankruptcy filing occurs.