Capstone Project 1: In-Depth Analysis

In this section, a few algorithms will be 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 1 and 2, respectively.

		Predicted	
		0	1
Actual	0	1631	14
	1	116	12

Figure 1. Confusion matrix of test set data for standard logistic regression model.

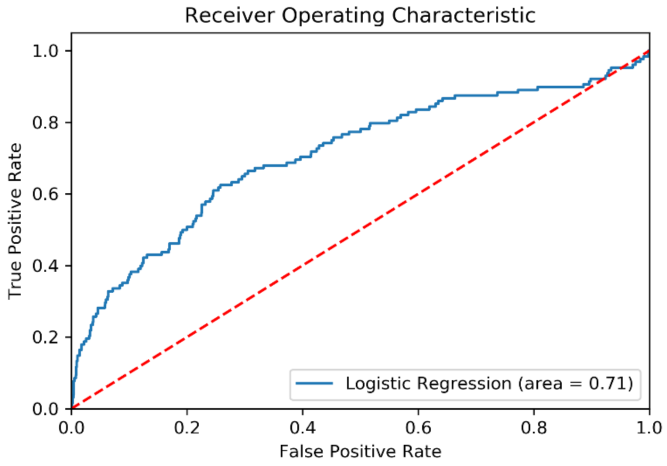

Figure 2. Receiver operating characteristic of test set data for standard logistic regression model.

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 3 and 4, respectively.

		Predicted	
		0	1
Actual	0	1626	19
	1	86	42

Figure 3. Confusion matrix of test set data for XGBoost classification model.

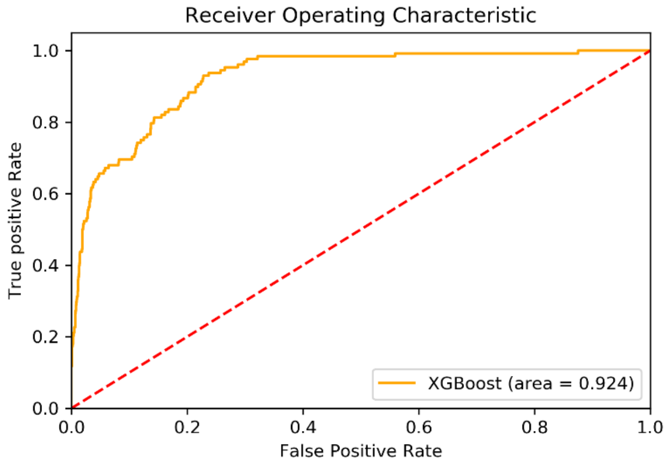

Figure 4. Receiver operating characteristic of test set data for XGBoost classification model.

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 5 and 6, respectively.

		Predicted	
		0	1
Actual	0	1627	18
	1	75	53

Figure 5. Confusion matrix of test set data for LightGBM classification model.

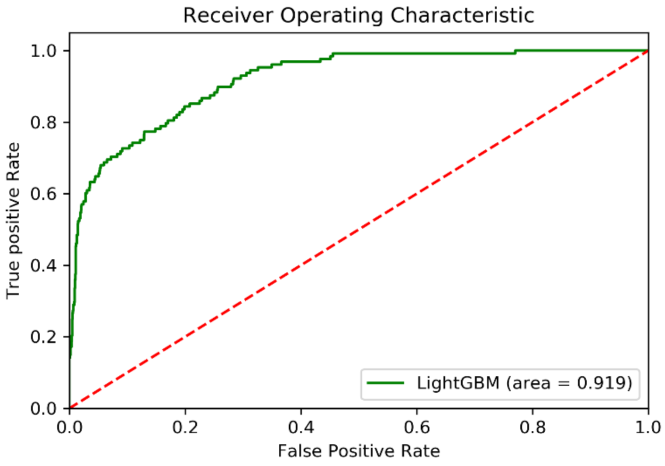

Figure 6. Receiver operating characteristic of test set data for LightGBM classification model.

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