Ellie Blaschko & Cadence Stewart
COSC 316
Data Science Final Writeup
11 December 2024

In order to predict what factors would make a prospective customer more likely to default on their credit card, we ran the data through three different machine learning models to understand what factors were most important in predicting each outcome. All these models had relatively high accuracy and we have developed a robust and reliable final model that we believe can be used to predict credit card defaults, which can be helpful to many aspects of the business including adjusting the budget, assigning credit lines, and deploying retention strategies.

Before running the data through our models, we cleaned up many aspects of the dataset so that the models could accurately read the data. There were a lot of missing values in the dataset due to incomplete customer information. We resolved this issue by dropping out the missing values so the algorithms wouldn't consider them in their analyses. We were faced with the option of replacing these null values with the mean of the data or dropping them out entirely. We decided that dropping out the null values would be more accurate to the dataset rather than replacing them with different values because only a small proportion of our data contained null values. The next step we took to clean the data was to convert all the date columns to a readable date format. After that, we subtracted all the date values from the value in the Current Date column using the pandas function to_datetime. Current Date was assumed to be when the data was extracted from the company database. Some of these subtracted values were negative because some of the dates occur after the date in the current date column. This may be a data extraction error, or some sort of future date for the customers such as a payment due date, but it won't affect the model since all of the Current Date values are the same day. We also dropped out the Unique ID column and the Current Date column because they aren't relevant variables to

consider. The last data cleaning step we took was to convert the categorical columns to binary variables. Each category of variable was added as a column and filled out with a 0 or a 1. After the data was cleaned, we moved onto model selection and training.

The three models that we selected for a large dataset with a balanced target variable and many feature types were a Logistic Regression, a Random Forest Classifier, and a Gradient Boosting Classifier. We hypothesized that all of these models would handle different aspects of the dataset well. A Logistic Regression is fast and easily interpretable, and can identify linear relationships well. Random Forest Classifiers are robust to overfitting and handle mixed data types and a large number of features particularly well. Lastly, Gradient Boosting Classifiers can handle complex, nonlinear relationships. For all of these models we varied over a different parameter and ran the test set data through the parameter that yielded the highest accuracy for the training set data. We iterated over the cost parameter for the Logistic Regression, the number of estimators and max depth for the Random Forest, and the learning rate and the number of estimators for the Gradient Booster. We decided to put the data through a standard scaler for all three models, as a standard scaler is applicable for all three of the models. We applied a column transformer to the numeric and categorical features to preprocess the data.

After scaling, preprocessing, and running a grid search on the data, we selected the parameters that produced the best results on the training set. We then cleaned the test set data using the same method we used on the training set. We cleaned the training and testing set separately to avoid data leakage. After cleaning the test set, we ran all three of the models with the best parameters on the test data and got the AUROC score for each model. We also plotted the ROC curve, which plots the false positive rate against the true positive rate. All of the models yielded around an 80% accuracy score. The Random Forest Classifier had the highest accuracy

score of 0.8275. We believe that this is an fitting model for this dataset, given the high number of features and the complex interactions between features. While we wouldn't recommend this model to be used as the only criteria for accepting customers, adjusting budgets, or employing certain retention strategies, it will be a helpful tool to use in tandem with existing methods.

LogisticRegression AUROC	0.807780942581072
RandomForest AUROC	0.8274658557562248
GradientBoosting AUROC	0.8248837994540101

Fig 1: AUROC Results

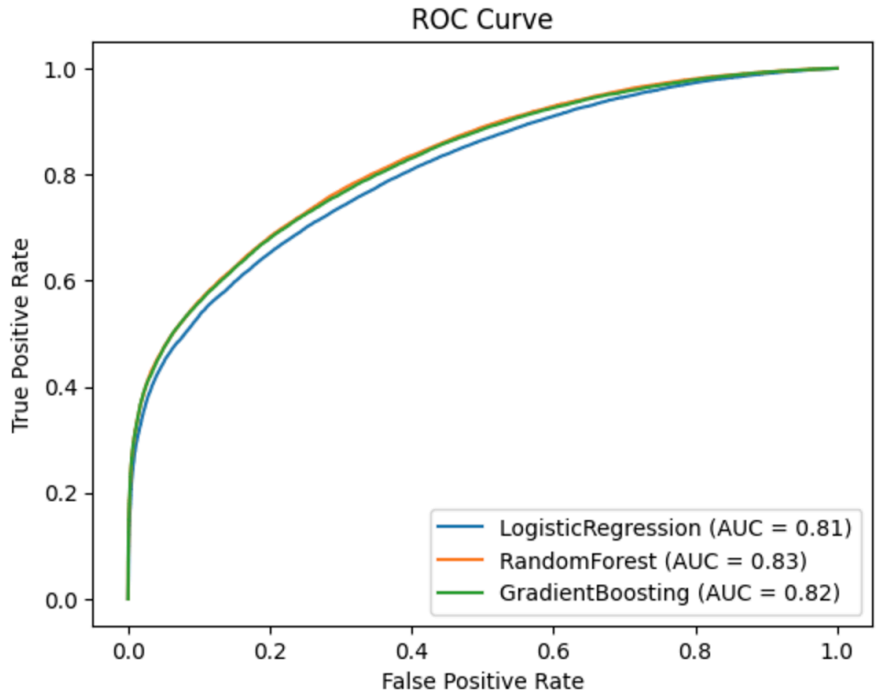

Fig 2: ROC Curve Plot