Goals/Objectives: Can we predict what credit limit someone should be assigned? If not, can we predict who should/shouldn’t be given credit? How do you ensure customers can/will pay their loans? Can we approve customers with high certainty? What is statistically significant to this problem? What real information can we derive?

Data Description: The data contains 30,000 transactions. This data set consists of 30 features including limit balance, age, marital status, payment amounts over 6 months, bill amounts over 6 months, customer gender, customer education level, and whether or not the customer is in default status. The data did have some errors, including duplicates so I cleaned it and preprocessed it. Additionally, I used one-hot encoding to change non-numeric features to numeric features. To look for patterns, I conducted an exploratory data analysis using Python and other libraries.

Analysis

For the first attempt at using machine learning for modeling and predictions, I used credit limit balance as the dependent variable to see if we could predict what credit limit someone should be assigned. Regression modeling was most appropriate for this first attempt. I used four different algorithms from Scikit-learn: Random Forest Regressor, Gradient Boosting Regressor, Linear Regression, and Support Vector Regression. The best model, Gradient Boosting Regressor, yielded an R squared of 0.478 so it is approximately 48% better than the mean. The RMSE yielded a high error rate of 93552.822. The cross validation score of Gradient Boosting Classifier was 0.47, so this initial model was not effective at determining if any other features were good predictors of who should and shouldn’t be given credit.

Next, I discretized credit limit balance into buckets, first into quartiles and next into 10 buckets. Using this feature as a dependent variable required classification modeling and predictions. I used the following classification algorithms: Random Forest Classifier, Gradient Boosting Classifier, Decision Tree Classifier, KNeighbors Classifier. The quartiles modeling attempt had an accuracy score of 0.59 and a cross validation score of .59 with Gradient Boosting Classifier. The models with limit balance discretized into 10 buckets also had an accuracy score of 0.59 and a cross validation score of 0.57, where Gradient Boosting Classifier was the best model. However, these accuracy scores and cross validation scores are only slightly better than a coin flip, which is not helpful to Credit One’s business goals.

Finally, I sought to address the question of how Credit One can determine whether or not customers should be given credit at all. I used classification modeling with customers in default as the dependent variable. This modeling again showed Gradient Boosting Classifier to be the best model. It returned an accuracy score of 0.82 and a cross validation score of 0.82. With this model, we can determine that customers who are in default should not be given credit since they are likely to default on future loans.

Conclusions (observations, recommendations) From the data provided, it does not appear that we can predict what credit limit someone should be assigned. We can predict with 82% accuracy if a customer is already in default that they are likely to default on future loans. I recommend that based on the model, those customers should not be given credit.