PROBLEM STATEMENT

Credit card companies make money by collecting interest on loans. When a customer is unable to pay, then the loan defaults. Defaults cause credit card companies to lose money by either writing the balance off entirely or selling the balance to a collection agency for pennies on the dollar. Therefore, determining if a customer will default in the next month can allow a bank to implement loss mitigation strategies in advance of the default.

PROPOSAL

This project will investigate whether customer account information can be used to build machine learning models to successfully predict whether or not a customer will default on their next month's payment. This is a supervised learning problem. I will evaluate six machine learning model frameworks' performance to understand if the dataset can be used to accurately predict whether a customer will default on their account in the next month.

The six models used are Logistic Regression, KNN classifier, Bagging classifier, AdaBoost classifier, XGBoost classifier, and Random forest classifier. The hyperparameters will be optimized and model performance evaluated using a confusion matrix, classification report (Precision, Recall, and F1), and Area under the Receiver operating characteristic curve (AUC ROC). The data used for this project is the Default of Credit Card Clients dataset, which can be accessed here: https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/download

THE DATA

The data used for this project is the Default of Credit Card Clients dataset, which can be accessed here: https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/download

From the graph at the right, we can see a moderate class imbalance, but not enough to warrant resampling of the data. There are 25 features, but this number was reduced by using feature selection techniques outlined in the next section.

Customer Defaults Next Month
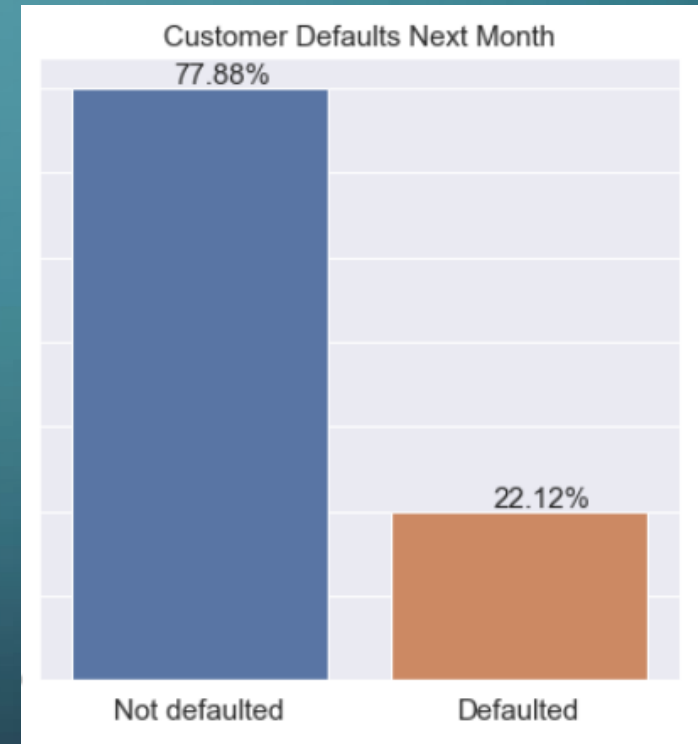
FEATURE SELECTION

After performing exploratory data analysis (EDA) on the training data, multicollinearity was found between several features. Additionally, there were features with missing or incomplete data. Six features were removed as a result of EDA.

Analysis of variance (ANOVA) F-tests were performed to determine the individual relationships between each feature and the response variable. The tests indicated which variables were most likely to contain useful information for predicting the response variable.

A Random Forest classifier was fit to the training data and the most important features were analyzed using the feature importances method. A logistic regression model was also fit and used to determine features with the largest absolute value of coefficients. These features have the greatest influence and are thus considered the most important.

The above methods were used to make the final determination to include 11 of the original 25 features for model building and evaluation.

MODEL EVALUATION RESULTS

Six models were tuned using k-fold cross-validation to optimize their hyperparameters. Below are the results of each optimized model for predicting if a customer will default next month.

Logistic Regression	KNN Classifier	Bagging classifier	AdaBoost classifier	XGBoost classifier	Random Forest
▪ Accuracy = 81.35%	▪ Accuracy = 81.04%	▪ Accuracy = 81.07%	▪ Accuracy = 81.93%	▪ Accuracy = 81.84%	▪ Accuracy = 81.96%
▪ Precision = 0.713	▪ Precision = 0.634	▪ Precision = 0.622	▪ Precision = 0.688	▪ Precision = 0.674	▪ Precision = 0.683
▪ Recall = 0.262	▪ Recall = 0.339	▪ Recall = 0.367	▪ Recall = 0.335	▪ Recall = 0.347	▪ Recall = 0.344
▪ F1 Score = 0.383	▪ F1 Score = 0.441	▪ F1 Score = 0.462	▪ F1 Score = 0.451	▪ F1 Score = 0.458	▪ F1 Score = 0.458
▪ ROC AUC = 0.71	▪ ROC AUC = 0.73	▪ ROC AUC = 0.74	▪ ROC AUC = 0.76	▪ ROC AUC = 0.77	▪ ROC AUC = 0.77

CONCLUSIONS

The logistic regression model accurately predicted the majority class better than the other models but had the lowest recall and F1 scores for the minority class. The minority class is what we are trying to predict, which is if a customer account will default in the next month. For this reason, the Logistic Regression model is the least useful. The Bagging classifier had the highest F1 score (was most skillful) for predicting the minority class but had the lowest F1 score (was least skillful) for predicting the majority class.

The ROC curves showed that the random forest ensemble model had the highest ROC score at 0.77 and the steepest ROC curve. The Random forest ROC curve's steepness indicates that it is the best at maximizing the true positive rate while minimizing the false positive rate. Additionally, the random forest model had the second-highest F1 score for predicting the minority class and the highest F1 score for predicting the majority class. Random forest also had the highest accuracy. For these reasons, random forest appears to be the winner, with XGBoost coming in at a very close second place.