Supervised Learning Project
Introduction to Machine Learning: Supervised Learning
Steve van Schouwen

Detecting Credit Card Fraud

Credit card fraud is an issue that impacts every business that takes credit card payments as well as every credit card holder. In 2023, credit card fraud losses totaled 34 billion dollars. Credit card processing companies and retailers who process high volumes of transactions require the ability to quickly identify fraudulent credit cards to cancel the card to reduce losses.

Data
I will be analyzing a dataset of 284K credit card transactions with 30 features. The dataset was downloaded from kaggle.com and was compiled through a collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles). (1)

The goal of this project will be to successfully identify the features having the strongest impact on fraud. This will be a non-parametric, supervised, binary-class classification project. The documentation states that this data is imbalanced, which makes it a good candidate for random forest and boosting classification methods. The feature labels are mostly non-descriptive (V1 through V28 plus time and amount). This project will be particularly interesting to me as a software engineer working for a credit card processing company; it will also be immediately useful.

Data cleaning/ munging:
The first challenge is to analyze and clean the data as needed. First, searching for missing class values and deleting those records, then identifying any highly correlated features and deciding how to modify or drop. Correlation will be identified using a seaborn correlation heatmap with correlation values and plotting the scatter_matrix for the features.

Steps Taken
Search for records missing class values: none found, no action taken
Measure imbalance: class values were highly imbalance with less than 2% positives
Drop useless Time feature: the Time value was a sequential value and was not useful for analysis, dropped
Convert Amount feature from $##.## to $##. value. Rounded value to nearest dollar and converted to int. This reduced the variation of the data by 100x, basically a histogram at the full dollar level

Exploratory Data Analysis:
Goal, identify problems with collinear features which could over-state the impact and reduce the impact of other features.

Steps Taken
Correlation Matrix to pull high-level correlation information about to predict features with highest impact on fraud and to look for features that were too tightly correlated. Identified a set of features predicted to have the highest influence. feature to feature correlation was not significant.

Pair Plot features to identify and correct issues of collinearity. Ran pair plot, identified and dropped several features. Re-ran pair plot and identifed two more feature pairs to trim.

Models
Because the data in my dataset was highly imbalanced (<2%), I selected models that minimize the impact of the imbalance through ensembling and boosting. The models I selected are Random Forest and ADAboost. I will be testing multiple hyperparameters to tune the model for the best results. I will compare the results to a simple Decision Tree model as a benchmark value. I will set the class_weight to balanced for each model to reduce the impact of the imbalanced data. I will use Recall/Sensitivity to measure the success of each model; Recall is better than accuracy as a measurement of success for imbalanced data since the models could simply return 0 for every test and have a >98% accuracy for this data.

Recall = True Positives / True Positives + False Negatives

Steps Taken:
Run Decision Tree for benchmark value.
Run Random Forests for multiple n_estimators hyperparameters, capture an array of recall and time values
Run ADABoost model for multiple of weak learners, capture an array of recall and time values
Check Cross-Validation Accuracy Scores: This was not useful data since it relies on accuracy and reported 99.9%+ for all cases

Results and Analysis
Both the Random Forest model and the ADABoost model had better results than the standard Decision Tree. The Random Forest model showed optimal results at n_estimators = 7. The ADABoost model showed increasing results as the number of weak learners increased; however, the improvement flattened out at 600 weak learners. The Random Forest tests were much faster than the ADABoost. Both models identified the same top 5 features impacting credit card fraud, though in a slightly different order.

Steps Taken:
Graph RF and ADA Boost recall values for each hyperparameter test compared to DT benchmark value: both RF and ADA had significantly higher recall values vs. DT
Plot ROC curve for each model, both RF and ADA Boost had significant AUC values indicating a good tests
Plot Time for RF and ADA Boost for each hyperparameter test: RF time grew linearly as the hyperparameter grew, but all tests were fast < 5 seconds. ADA Boost tests
grew linearly as the hyperparameter grew, but the tests were much slower, topping at 500 seconds.
Identify the top features by importance for each model

Discussion and Conclusion

<u>Learning and takeaways:</u>
This project walked through each step required to analyze data using supervised ML tools. This session has been challenging, but completing this project proves that I have acquired the knowledge needed to run ML analyzation projects from raw data through to analyzing the results.

I will be able to easily apply the model I have built on live production data for my current position as a software engineer for a credit card processing company. I'm positive the results will be of great interest to my client.

The specific takeaways I got from this project are that the longest-running/ most complex model is not necessarily the best one, and the fastest way to find out is to test several methods. I was surprised that the dollar value feature was not a top-five predictor of fraud.

what worked:
Nearly everything worked really well for this project. The course really provided a road map from start to finish for this project. I would have liked to test more hyperparameters for ADA Boost, but each additional test was increasing the time longer than I could afford to wait.

how to improve:
I only tweaked one hyperparameter for each model. I would like to test each possible hyperparameter for Random Forest as well as testing some other models. I could also go deeper on the hyperparameter I did choose.

CITATIONS:

(1). Machine Learning Group & Université Libre de Bruxelles (2018, March) Credit Card Fraud Detection: Anonymized credit card transactions labeled as fraudulent or genuine. Retrieved March 2, 2025, from https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data