FRAUD DETECTION CASE STUDY

Finding the optimal machine learning algorithm to identify and prevent fraudulent transactions

Introduction

Credit card fraud is the most common form of identity theft each year, costing credit card companies hundreds of millions of dollars.

In order to secure customer transactions and save credit card companies from large monetary losses, we built three models to identify fraud transactions and compared model performance and accuracy.

About the Data

This dataset contains transactions made by credit cards.

There were 492 fraudulent transactions out of 284,807 total transactions.

This data is highly imbalanced as the event rate is 0.173%.

The variables in the dataset were time, transaction amount, and class ('0' indicated no fraud and '1' indicated fraud).

The other 28 columns were numerical input variables which were the result of PCA transformation due to confidentiality purposes.

Benefits of Machine Learning Against Fraud

- Modern money laundering prevention and increasing cybersecurity
- Fighting digital fraud with analytics in the banking sector
- Defeating Fraud, Waste, and Abuse that governments spend billions to prevent
- Preventing insurance application fraud

APPROACH

- **Resampling the Data:** In order to combat the imbalanced dataset, the SMOTE oversampling technique was used to randomly increase minority class instances.
- **Data Preprocessing**: In each model, we scaled the data using StandardScaler and used correlation analysis for variable selection to improve model performance.
- **Random Forest:** Ran GridSearchCV to find optimal parameters and applied Random Forest Classifier to build the model.
- **Neural Networks and SVM:** Used MLP Classifier to build classifier and set solver equal to [lbfgs, adam] and activation to [relu, log]. Adjusted hidden layers to see performance.
- **LightGBM:** Used RandomizedSearchCV to find optimal parameters and applied LGBM Classifier to build the model.
- **Tools**: Python (pandas, numpy, matplotlib, imblearn, sklearn, seaborn)

RESULTS

98.1%
Neural Networks:
483/492 frauds detected

91.9%
LightGBM:
452/492 frauds detected

86.0%
Random Forest:
423/492 frauds detected

CONCLUSION

The most optimal machine learning algorithm out of the three models tested was the Neural Networks algorithm. It was able to identify 483 out of the 492 frauds, saving the customer and the company the most amount of money. Although it was the most accurate, it took the longest training time. The second highest performing algorithm was LightGBM which correctly identified 452 out of the 492 frauds. This percentage was still relatively high, however the model training time was significantly less than the neural network. The least accurate algorithm was Random Forest, identifying only 423 out of the 492 frauds. The training time was moderate compared to LightGBM, but still significantly less than the Neural Networks. Our recommendation would be to use the Neural Networks algorithm, as it will correctly identify the most fraud transactions, which is the most important aspect of fraud detection.