Credit Card Fraud Detection

Outline

- Problem
- Dataset
- Fraud Detection
    - Machine Learning approaches in Fraud Detection
    - Anomaly Detection
- Modeling
    - Supervised Learning
    - Unsupervised Learning
    - Anomaly vs Supervised Learning
- Testing and Tuning
- Deployment and Pipeline
- Sum up

Problem Statement

Fraud detection is a set of activities that are taken to prevent money or property from being obtained through false pretenses. Fraud can be committed in different ways and in many industries. Most detection methods combine a variety of fraud detection datasets to form a connected overview of both valid and non-valid payment data to make a decision. The Credit Card Fraud Detection Problem includes modeling past credit card transactions with the knowledge of the ones that turned out to be a fraud. This model is then used to identify whether a new transaction is fraudulent or not. My aim here is to detect 100% of the fraudulent transactions while minimizing the incorrect fraud predictions.

Machine Learning techniques in Fraud Detection

One of the common techniques to detect fraud in credit card payment is Anomaly Detection that used to

identify unusual patterns that do not conform to expected behavior, called outliers.

Some Machine Learning Detection: approaches for Anomaly

- Clustering-Based Anomaly Detection
- Support Vector Machine
- Isolation Forest
- XGBoost

Dataset

https://www.kaggle.com/mlg-ulb/creditcardfraud/download
Dataset is transformed Principal Component Analysis (PCA) which is commonly used:

- Dimensionality reduction algorithm
- Speed-up Machine Learning algorithms

Observation

- some features (V1, V2, V3, ... ,V28) transformed to PCA and Time, Amount features not transformed.
- target is Class (1-Fraud, 0-Valid)

Metrics
In a binary classification problem such as this, a model classifies examples as either positive (fraudulent) or negative (genuine). The decision made by the model, either positive or negative can be represented in a structure known as confusion matrix. This confusion matrix has four elements that define it, contextually they are:

- True Positive (TP) – An example where a transaction is fraudulent and is classified correctly as fraudulent.

- False Positive (FP) – An example where a transaction is valid and is classified as incorrectly as fraudulent.
- True Negative (TN) – An example where a transaction is fraudulent but is classified incorrectly as valid.

- False Negative (FN) – An example that is valid and is classified correctly as fraudulent.

Anomaly Detection algorithms I used:

- Random Forest
- XGBoost
- Based on this initial EDA, this dataset does not have any null values and highly imbalance. Anomaly Detection is best for unbalanced dataset and supervised learning is better if dataset balanced.

Benchmark

For this project, considering the imbalance class ratio, accuracy would not be used to judge the model since it can be misleading. Instead, as a benchmark the model should have an Area Under the Precision-Recall Curve (AUPRC) score of % 97 or greater.
This score was gotten from a XGBoost classifier I built; this serves as benchmark towards building a better model.

Conclusion

After the Logistic Regression algorithm was chosen, hyper parameter tuning was performed to optimize the model. Grid Search was used to find the optimal parameters, this implies that the Area Under Precision Recall Curve of models was 0.97 which is a very good score. A recall score of 0.78 implies that the final model predicted 78% of the fraudulent transactions in the test dataset correctly while the precision measures that fraction of cases predicted to be fraudulent that are truly fraudulent.