WAYNE STATE
College of Engineering

CSC - 5825 Fall 2021

Introduction to Machine Learning and Applications

Term Project

CREDIT CARD FRAUD DETECTION

By

Md Reza

Project Background:

Credit card fraud is an ever-growing threat, impacting the financial domains, corporates, and government organizations with outnumbered consequences. Nowadays, organizations, companies, and government agencies have experienced substantial financial losses due to credit card fraud activities. An exponential increase in credit card fraud has led many researchers to develop advanced machine learning algorithms and techniques for early fraud detection. Nevertheless, leveraging credit card fraud datasets to detect fraudulent activities becoming more and more challenging mainly for the following reasons:

- the profiles & behavior of fraudulent and genuine users constantly changing.
- the credit card fraud dataset is highly imbalanced/skewed because the legit transactions in the sample has outnumbered the fraudulent transactions.

The objective of this project is to leverage supervised and unsupervised machine learning algorithms to evaluate a data set of known outcomes. The focus will be on performing predictive analysis to identify whether a given transaction either falls under the genuine or fraudulent classification. Different modeling techniques include: Logistic Regression, Decision Tree, Naïve Bayes, and Isolation Forest will be tested, where the goal will be selecting the best model(s) that would predict whether a charge is fraudulent or genuine.

Methods:

The research described in this paper is to create the trained model(s) that predicts the fraudulent credit card charge before it happens. To approach this problem the credit card fraud dataset was used from Kaggle to classify if a charge is genuine or fraudulent based on a set of predictors. To reduce the feature space, a PCA was performed on the data which resulted in a highly condensed data set for creating a model. This data was very unbalanced due to the high sample size and the low number of fraud classifications, only ~2% of the data was classified as fraud, see figure - 1.

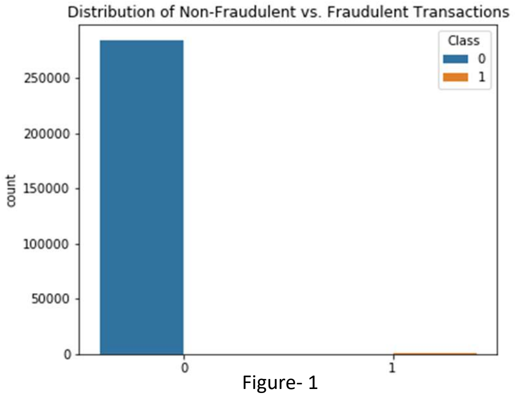


Figure- 1

To overcome the issue with imbalanced data an upsampling or oversampling technique is used to create artificial or duplicate data points or of the minority class sample to balance the class label. Then split data into train, test, and post-upsampling 10-fold cross-validation was performed. Finally, these data sets were used to train and test the following supervised & unsupervised machine learning algorithms:

1. **Logistic Regression:** Logistic regression describes and estimates the relationship between one dependent binary variable and independent variables.
2. **Decision Tree:** A decision tree is a hierarchical data structure implementing the divide-and-conquer strategy. It is an efficient nonparametric method, which can be used for both classification and regression. [1]
3. **Naïve Bayes:** Conditional probability is a measure of the probability of an event given that another event has already occurred.
4. **Isolation Forest:** Belong to ensemble models, which means that the predictions do not rely on a single model, instead, it combines the results of multiple models to distinguish outliers from regular data in an unsupervised fashion.

All models were generated and ran in Python. The models were trained on the training data, then the accuracy was measured with the testing data set. The model with highest accuracy and f1-score was selected as the overall best performance. It should be noted that all the models have more than 90% accuracy. Each model was tested for accuracy using F1-Score to compare each model in terms of confusion matrix.

Experiment, Results and Discussion:

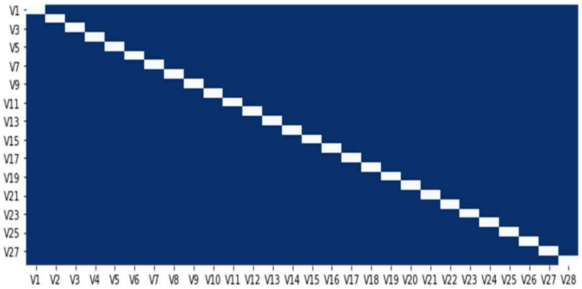
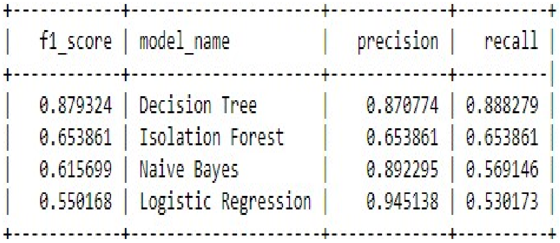


Figure- 2

In the correlation plot (figure - 2), notice that none of the PCA transformed (V1-V28) have any correlation, which justifies the non-linearity and the features to be uncorrelated due to the PCA transform. Because of that though the classification of a transaction based on the predictors becomes possible; however, due to confidentiality citing the exact predictors that cause fraudulent transactions is out of scope. [2]

Comparison of Supervised and Unsupervised Algorithms:

After comparing models' output, it turned out that F1-Score, Precision, & Recall play a key role in evaluating the performance of the models, and the goal was to capture possible fraudulent cases without raising false alarms too frequently. The observations-based models' outputs are as follows:



```
+-----------+-------------------+-----------+---------+
|  f1_score | model_name        | precision |  recall |
+-----------+-------------------+-----------+---------|
|  0.879324 | Decision Tree     |  0.870774 | 0.888279 |
|  0.653861 | Isolation Forest  |  0.653861 | 0.653861 |
|  0.615699 | Naive Bayes       |  0.892295 | 0.569146 |
|  0.550168 | Logistic Regression|  0.945138 | 0.530173 |
+-----------+-------------------+-----------+---------+
```

1. The Decision Tree performs particularly well balanced.
2. The opposite is true for the Isolation Forest unsupervised model. Here, only a few fraud cases are detected, but the model is comparatively correct to identify fraud cases.
3. The Naive Bayes model has a slightly high F1-Score and detects many fraud cases. But it also raises false alarms very frequently.
4. The Logistic Regression model performs slightly worse than the other models.

Conclusion:

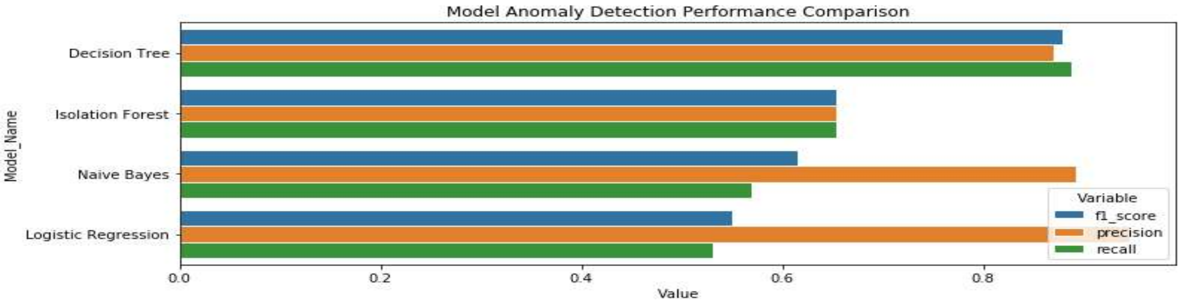


Figure- 3

Finally, in this project, different supervised & unsupervised ML algorithms were developed to detect fraudulent credit card transactions. Different techniques like Upsampling, Z-score, outliers fraction calculation, 10-fold cross-validation were leveraged to prepare the data for training, validating, and testing Logistic Regression, Decision Tree, Naïve Bayes, and Isolation Forest. According to Model Anomaly Detection Performance Comparison (figure – 3), it turned out that the Decision Tree and Isolation Forest perform better for anomaly detection and outperform Regression & Naïve Bayes.

Bibliography:

[1] Ethem Alpaydin “Introduction to Machine Learning”, MIT Press, 2014.

[2] Machine Learning Group – ULB, Credit card Fraud Detection (2018), Kaggle https://www.kaggle.com/mlg-ulb/creditcardfraud

[3] Ian Goodfellow and Yoshua Bengio “Deep Learning”, The MIT Press, 2016.

[4] D. Zhu, personal communication, n.d.