Finance

DESCRIPTION

Problem Statement

- The Finance Industry is the biggest consumer of Data Scientists. It faces constant attack by fraudsters, who try to trick the system. Correctly identifying fraudulent transactions is often compared with finding a needle in a haystack because of the low event rate.
- It is important that credit card companies are able to recognize fraudulent credit card transactions so that the customers are not charged for items that they did not purchase. You are required to try various techniques such as supervised models with oversampling, unsupervised anomaly detection, and heuristics to get good accuracy at fraud detection.

Dataset Snapshot

The datasets contain transactions made by credit cards in September 2013 by European cardholders. This dataset represents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

Project Task: Week 1

Exploratory Data Analysis (EDA):

Perform an EDA on the Dataset.

- Check all the latent features and parameters with their mean and standard deviation. Value are close to 0 centered (mean) with unit standard deviation

- - Find if there is any connection between Time, Amount, and the transaction being fraudulent.
- Check the class count for each class. It’s a class Imbalance problem.

Use techniques like undersampling or oversampling before running Naïve Bayes, Logistic Regression or SVM.

- Oversampling or undersampling can be used to tackle the class imbalance problem

- - Oversampling increases the prior probability of an imbalanced class and in case of other classifiers, error gets multiplied as the low-proportionate class is mimicked multiple times.
- Following are the matrices for evaluating the model performance: Precision, Recall, F1-Score, AUC-ROC curve. Use F1-Score as the evaluation criteria for this project.

Observations: Week 1

- Training set contains two days transaction details, the rate transaction low at night times
- The amount of transaction is highly skewed distribution, only low number count detected for high amount transaction
- There is no connection between Time, Amount, and the transaction being fraudulent
- All the fraudulent transactions are low amount
- This is highly class Imbalance problem, fraudulent transaction are very rare less than 0.17 %

Project Task: Week 2

Modeling Techniques:

- Try out models like Naive Bayes, Logistic Regression or SVM. Find out which one performs the best

Use different Tree-based classifiers like Random Forest and XGBoost.
a. Remember Tree-based classifiers work on two ideologies: Bagging or Boosting

- b. Tree-based classifiers have fine-tuning parameters which take care of the imbalanced class. Random-Forest and XGBboost.
- Compare the results of 1 with 2 and check if there is any incremental gain.

Observations: Week 2

- Compare Logistic, SVM the ensembles method had more incremental gain.
- RandomForestClassifier Bagging classifire outperfome compare to other modals
- Computation wise EtraTreeClassifire is the better one and a little downwards

Project Task: Week 3

Applying ANN:

Use ANN (Artificial Neural Network) to identify fraudulent and non-fraudulent.
a) Fine-tune number of layers

b) Number of Neurons in each layers

c) Experiment in batch-size

d) Experiment with number of epochs. Check the observations in loss and accuracy

e) Play with different Learning Rate variants of Gradient Descent like Adam, SGD, RMS-prop

f) Find out which activation performs best for this use case and why?

- g) Check Confusion Matrix, Precision, Recall and F1-Score
- Try out Dropout for ANN. How is it performed? Compare model performance with the traditional ML based prediction models from above.

- Find the best setting of neural net that can be best classified as fraudulent and non-fraudulent transactions. Use techniques like Grid Search, Cross-Validation and Random search.

Anomaly Detection:

Implement anomaly detection algorithms.

a) Assume that the data is coming from a single or a combination of multivariate Gaussian

- b) Formalize a scoring criterion, which gives a scoring probability for the given data point whether it belongs to the multivariate Gaussian or Normal Distribution fitted in a)

Observations: Week 3

- After fine tuning this is the best combination of hyperparsm learn_rate: 0.2, epochs: 50, dropout_rate: 0.4, batch_size: 40, optimization: SGD
- ANN oupperfome most, with more than 99% accuracy and 80% of f1 score
- In terms of Anomaly Detection Isolation Forest is better compare to Local Outlier Factor

Project Task: Week 4

- Visualize the scores for Fraudulent and Non-Fraudulent transactions.
- Find out the threshold value for marking or reporting a transaction as fraudulent in your anomaly detection system.
- Can this score be used as an engineered feature in the models developed previously? Are there any incremental gains in F1-Score? Why or Why not?
- Be as creative as possible in finding other interesting insights.

Find out the threshold value for marking or reporting a transaction as fraudulent in your anomaly detection system.

- **the AUC score of ANN training its more than 91 %,which can use as threshold for Anomaly detection.**

Find out the threshold value for marking or reporting a transaction as fraudulent in your anomaly detection system.

- **F1 takes both recall and precision, which make a more accurate way to measure performance of different modals.**
- **F1 score for the neural network model is better than the traditional Machine Learning model.**

Be as creative as possible in finding other interesting insights.

- **The training data contains two days of data, transaction rate day time much more than compare to nights**
- **The fraudulent transaction had less amount**
- **The time and amount had no correlation in fraudulent transaction**