Fraud Detection in Credit Card Transactions

...

Mentor - Dr. Bhaskar Biswas
Vivek Kumar Yadav 416
Kunal Gupta 347
Utsav Raj 411

Problem statement

The aim of the project is to predict fraudulent credit card transactions using machine learning models.
Also choose best suited models by comparisons.

This is crucial from the bank's as well as customer's perspective. The banks cannot afford to lose their customers' money to fraudsters. Every fraud is a loss to the bank as the bank is responsible for the fraud transactions.

The dataset contains transactions made over a period of two days by credit cardholders.

Prerequisites

Most of the libraries like sklearn, pandas, seaborn, matplotlib, numpy, scipy.

We will be Training the model with various algorithm such as Logistic regression, SVM, Decision Tree, Random forest also their comparison.

We are working on jupyter Notebook

Using Supervised ML Models for fraud detection

We know KNN, Decision Trees, Random Forest, XGboost etc.

K-Nearest Neighbors is a Classification algorithm

- It counts similarities based on the distance
- The data point, will be assigned the class that the nearest neighbors
- This method is not vulnerable to noise and missing data points, which means composing larger datasets in less time.
- It is quite accurate and requires less work from a developer in order to tune the model.

Using Supervised ML Models for fraud detection

Random Forest is a classification algorithm

- It is comprised of many Decision Trees.
- Each tree has nodes with conditions,which define the final decision based on the highest value.
- For fraud detection and prevention it has two cardinal factors that make it good at predicting things.
- The first one is randomness, means the rows and columns of data are chosen randomly from the dataset and fit into different Decision Trees.
- The second factor is diversity, means there’s a forest of trees that contribute to the final decision instead of just one decision tree, diversity decreases the chance of model overfitting, while the bias remains the same.

Using Supervised ML Models for fraud detection

XGboost algorithm

It a single type of gradient-boosted Decision Trees algorithm,

which was created for speed as well as maximizing the efficiency of computing time and memory resources.

This algorithm is a blending technique where new models are added to fix the errors caused by existing models.

Steps to be taken

Reading, understanding and visualising the data (EDA)

- It is a CSV file, contains 31 features, the last feature is used to classify the transaction whether it is a fraud or not.

Preparing the data for modelling

Building the model (with different model)

Evaluate the model

DataSet

It is a CSV file, contains 31 features, the last feature is used to classify the transaction whether it is a fraud or not

This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. Which is 0.17%.

It contains only numeric input variables. Due to secrecy issues original features are not provided Features V1, V2, … V28 are the principal components obtained is the only features

link :https://www.kaggle.com/mlg-ulb/creditcardfraud

Using ROC-AUC Receiver Operating characteristic curve

As we have seen that the data is heavily imbalanced, where only 0.17% transactions are fraudulent, we should not consider accuracy as a good measure for evaluating the model. Because in the case of all the data points return a particular class(1/0) irrespective of any prediction, still the model will result very high Accuracy.

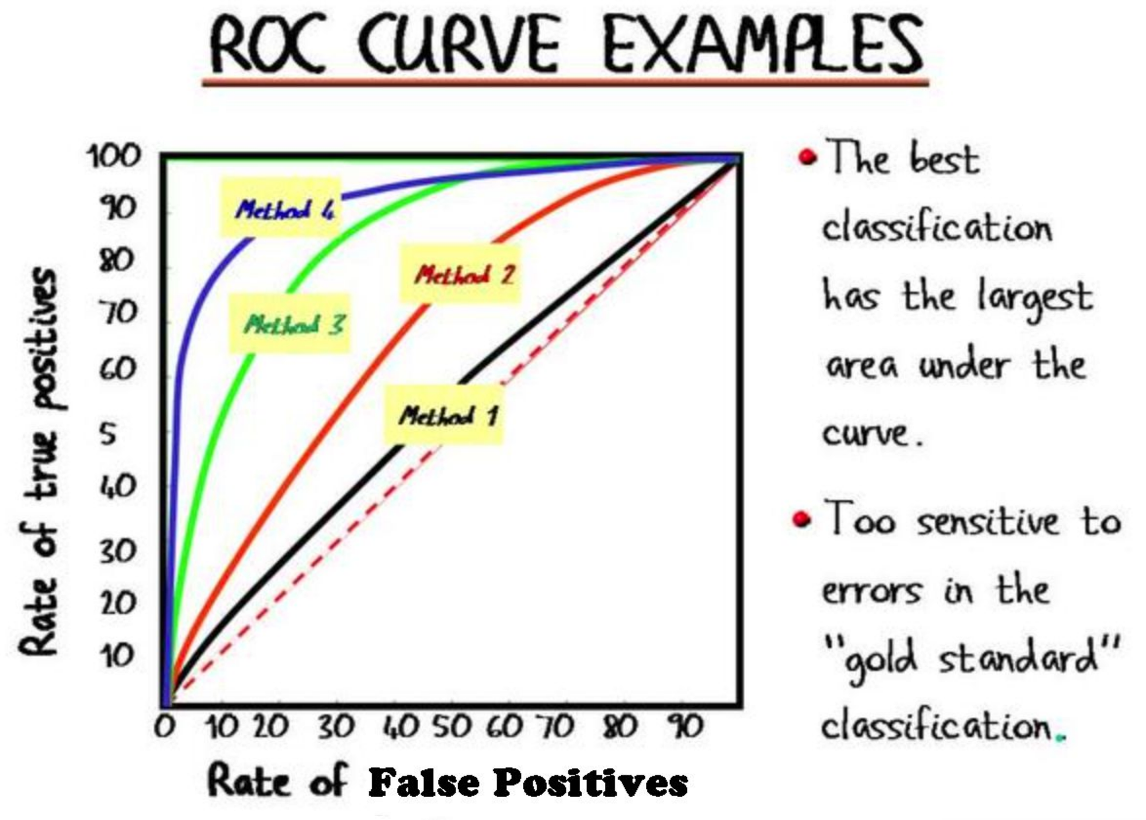
Hence, we have to measure the ROC-AUC score for fair evaluation of the model.

The ROC curve is used to understand the strength of the model by evaluating the performance of the model at all the classification thresholds.

Overall accuracy is based on one specific threshold, while ROC tries all of the threshold and plots the sensitivity and specificity. So when we compare the overall accuracy, we are comparing the accuracy based on some threshold.

ROC curve

fig.



ROC CURVE EXAMPLES

- The best classification has the largest area under the curve.
- Too sensitive to errors in the "gold standard" classification.

Using ROC-AUC

The ROC curve is measured at all thresholds, the best threshold would be one at which the TPR(True Positive Rate) is high and FPR(False Positive Rate) is low, i.e., misclassifications are low. After determining the optimal threshold, we can calculate it.

F1 Score(is a measure of a model's accuracy on a dataset) of the classifier to measure the precision and recall at the selected threshold.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Accuracy , Precision , Recall

Accuracy - it is simply a ratio of correctly predicted observation to the total observations.

Accuracy = TP+TN/TP+FP+FN+TN

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

Precision = TP/TP+FP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class

Recall = TP/TP+FN

What we are Learning.

ALL different models discussed above,

Along with implementation.

References:

https://spd.group/machine-learning/credit-card-fraud-detection/

https://www.kaggle.com/mlg-ulb/creditcardfraud

https://towardsdatascience.com/credit-card-fraud-detection-using-machine-learning-python-5b098d4a8edc

AFTER EDA MODEL SELECTION AND PREDICTION.

CONTINUE FROM HERE

MODEL Demonstration

Classification Models we are using over imbalanced data set.

1 .LOGISTIC Regression.

2 .DECISION TREE

3 .XGBOOST

If needed we can check with some other classification models.

But our goal is to find best suited model that consumes less computational resources and build on infrastructure that takes less deploying cost.

Aim of the Project

Choosing best model on the balanced data.

Here we are trying to balance the data with various approach such as

- Undersampling,
- Oversampling,
- SMOTE ->Synthetic minority oversampling technique
- Adasy.

With every data balancing technique we built several models such as Logistic, XGBoost, Decision Tree, and Random Forest.

WHY NEED TO BALANCE DATA

We have seen that the data is heavily imbalanced, where only **0.17% transactions are fraudulent**, we should not consider simply a accuracy as a good measure for evaluating the model. Because in the case of all the data points return a particular class(1/0) irrespective of any prediction, still the model will result very high Accuracy.

Hence, we have to measure the ROC-AUC score for fair evaluation of the model,along with Balancing techniques given below.

Aim of the Project to choose best balancing technique.

Undersampling :- Here for balancing the class distribution, the non-fraudulent transactions count will be reduced to 396 (similar count of fraudulent transactions)

Oversampling :- Here we will make the same count of non-fraudulent transactions as fraudulent transactions.

SMOTE :- Synthetic minority oversampling technique. It is another oversampling technique, which uses nearest neighbor algorithm to create synthetic data.

Adasyn:- This is similar to SMOTE with minor changes that the new synthetic data is generated on the region of low density of imbalanced data points.

Our Foremost Task, Predict accuracy of model and Feasibility

Finally our task is to predict accuracy using ROC-AUC curve after choosing best balancing technique mentioned above.

Like we can perform all three classification models in **Undersampling, Oversampling , SMOTE, and in Adasyn respectively.**

Only to get best suited models over tried classification model.

Reason : *We also have to consider that for little change of the ROC score how much monetary loss of gain the bank incur..*

Thank You

Thank You