PROJECT REPORT
On

PREDICTION OF CREDIT CARD FRAUD(PROJECT 2)

Submitted
by
Akshara Anil

ABSTRACT

The main aim of this project is the detection of credit card fraudulent transactions, as it's important to figure out the fraudulent transactions so that customers don't get charged for the purchase of products that they didn't buy. The detection of the credit card fraudulent transactions will be performed with multiple ML techniques then a comparison will be made between the outcomes and results of each technique to find the best and most suited model in the detection of the credit card transactions that are fraudulent.

INTRODUCTION

Fraud detection focuses on identifying transactions that deviate significantly from normal patterns. In credit card fraud, fraudulent transactions often exhibit these anomalies. The key challenge is distinguishing between unusual but legitimate behavior and genuine fraud.With the increase in people using credit cards in their daily lives, credit card companies should take special care in the security and safety of the customers. According to (Credit card statistics 2021) the number of people using credit cards around the world was 2.8 billion in 2019, in addition 70% of those users own a single card at least.

OBJECTIVE

As stated before credit card fraud is increasing drastically every year, many people are facing the problem of having their credits breached by those fraudulent people, which is impacting their daily lives, as payments using a credit card is similar to taking a loan. If the problem is not solved many people will have large amounts of loans that they cannot pay back which will make them face a hard life, and they won't be able to afford necessary products, in the long run not being able to pay back the amount might lead to them going to jail. Basically, the problem proposed is the detection of the credit card fraudulent transactions made by fraudsters to stop those breaches and to ensure customers security.

My analysis regarding the dataset:

1. The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions.(These were already given in the question).
2. The dataset contains only numeric columns which had already undergone PCA transformation to ensure further exploitation of the dataset , since the data was confidential and sensitive.
3. There were no null values present in the dataset so null values were not to be managed.
4. The only independent columns which were as direct values were “Time” and “Amount” columns.These were the only values which made sense to be examined during EDA.Also the values under other columns seemed to have little or no correlation in the correlation matrix.
5. First the data was taken as such and was trained by the model.After that **Undersampling** was performed to check whether the model would work better.Th

PROJECT METHODOLOGY

Phase 1:Bussiness understanding

As stated before credit card fraud is increasing drastically every year, many people are facing the problem of having their credits breached by those fraudulent people, which is impacting their daily lives, as payments using a credit card is so frequent on a daily basis.the problem proposed is the detection of the credit card fraudulent transactions made by fraudsters to stop those breaches and to ensure customers security.

Phase 2: Data Understanding

In the Data understanding phase, it was critical to obtain a high-quality dataset as the model is based on it, the dataset was explored by taking a closer look into it which gave the knowledge needed to confirm the quality of the dataset, additionally to reading the description of the whole dataset and each attribute. It's also important to have a dataset that contains several mixed transaction types , Fraudulent and Real, and a class to clarify the type of transaction.

EDA was performed to:

1. To know the distribution of values 0 and 1 in 'Class' column using a bar chart.
2. Correlation matrix was implemented to find the correlation between each of the independent features.

   Using this it was found that the value v17 had some positive correlation with the “Class”.Using this it was found that the value v17 had some positive correlation with the “Class”.
3. A scatter plot was used to xheck the relation between “Amount” and “v17” column

Phase 3: Data Preparation

After choosing the most suited dataset the preparation phase begins, the preparation of the dataset includes selecting the wanted attributes or variables, cleaning it by excluding Null rows, deleting duplicated variables, treating outlier if necessary, in addition to transforming data types to the wanted type, data merging can be performed as well where two or more attributes get merged.

This data has not gone through much of cleaning since the data,

1.Had no null values

2.The dataset was numerical in nature.

3.Most all the columns had already gone through PCA to protect the data from further exploitation and since the data was sensitive and confidential.

4.There were duplicate values and they were dropped.

Rest the data was clean.

Phase4:Model Evaluation:

Models used for prediction:

Two machine learning models were created in the modeling phase-

- Logistic Regression
- Decision Tree Classifier

A comparison of the results will be presented later in the paper to know which technique is most suited in the credit card fraudulent transactions detection.

Logistic Regression and **Decision Tree Classifier** were first trained on a dataset after splitting in the ratio 80:20 then these models were applied on Undersampled data.

Uppersampling wasnt used since it can make the dataset erroneous

The metric score was greater in Undersampled data .

Metric evaluation-Logistic regression before undersampling

```
Accuracy : 0.999048391076023

Precision : 0.75

Recall : 0.6

F1 score : 0.6666666666666666

 Classification_Report:              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56656
           1       0.75      0.60      0.67        90

    accuracy                           1.00     56746
   macro avg       0.87      0.80      0.83     56746
weighted avg       1.00      1.00      1.00     56746
```

Metric evaluation-Decision Tree Classifier before undersampling

```
Accuracy : 0.9990307686885419

Precision : 0.6804123711340206

Recall : 0.7333333333333333

F1 score : 0.7058823529411765

 Classification_Report:              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56656
           1       0.68      0.73      0.71        90

    accuracy                           1.00     56746
   macro avg       0.84      0.87      0.85     56746
weighted avg       1.00      1.00      1.00     56746
```

Metric evaluation-Logistic regression after undersampling

```
===========Logistic Regression=========

Accuracy : 0.9368421052631579

Precision : 0.96875

Recall : 0.9117647058823529

F1 score : 0.9393939393939394

Classification Report :              precision    recall  f1-score   support

           0       0.90      0.97      0.93        88
           1       0.97      0.91      0.94       102

    accuracy                           0.94       190
   macro avg       0.94      0.94      0.94       190
weighted avg       0.94      0.94      0.94       190
```

Metric evaluation-Decision Tree Classifier after undersampling

```
==========Decision Tree Classifier=========

Accuracy : 0.9

Precision : 0.8952380952380953

Recall : 0.9215686274509803

F1 score : 0.9082125603864735

Classification Report :              precision    recall  f1-score   support

           0       0.91      0.88      0.89        88
           1       0.90      0.92      0.91       102

    accuracy                           0.90       190
   macro avg       0.90      0.90      0.90       190
weighted avg       0.90      0.90      0.90       190
```

Which metric is more important in this project:

1.Recall is crucial in fraud detection because the cost of missing a fraudulent transaction (false negative) is typically higher than the cost of investigating a legitimate transaction flagged as fraudulent (false positive).

2.However, if **precision is too low**, there will be many false alarms, which can lead to customer dissatisfaction and inefficiencies in reviewing flagged transactions.

3.F1 Score is commonly used to balance recall and precision, especially in highly imbalanced datasets like fraud detection, where the majority of transactions are legitimate.