Machine Learning Project

Credit Card Fraud Detection

1. Title Page

- Title - Credit Card Fraud Detection Using Multiple Machine Learning models .
- Author - Amit Kumar Chaudhary

2. Abstract

This project uses machine learning to perform Transaction analysis . We use the **KAGGLE Wesite** dataset containing all fraudlent and normal transactions , applying various machine learning models such as Logistic Regression , Random Forest , Support Vector Machine and XGBoost for classification of transactions being normal or fraudlent .

3. Introduction

- Problem statement - Nowdays Credit Cards usage in all over the world had increased not only beacuse of its one prominent features that it gives people freedom to spend money first then to pay later . Since the more transactions shifted towards using Credit Card , the more fraud/anamoly transactions on it had also increased . It is important that credit card companies are able to recognize fraudulent credit card transactions so that customer's serice is not compromised at any cost ,and ofcourse not charged for items that they did not

purchase.

- Objective - The goal is to build a machine learning model that accurately predicts whether a given transaction is normal or fraudlent . The model is trained on the Kaggle dataset using multiple models .

4. Dataset Description

- Source - The datasets 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. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions .

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-senstive learning .

- Features - . Feature 'Class' is the response variable and it takes value 0 -> Normal transaction and 1 -> for fraudlent transaction .

- Inspiration - Identify fraudulent credit card transactions . Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification.

5 . Methodhology

Model Used - I have used four models namely Logistic Regression , Random Forest , Support Vector Machine and XGBoost .

Dataset manangement - As the dataset consists of 284315 normmal transactions and 432 fraudlent transactions thetefore its very important that we train our model both on normal and fraudlent transactions hence undersampling of the dataset is done .

Model Training -

the models are splitted and trained over 80% training data and tested over 20% testing data .

- In Logistic Regression , The model was trained with maximum iterations set to 1000 to ensure convergence.
- In Random Forest , n_estimators was set to 100 and random no. of state was set to 42.
- In SVM (Support Vector Machine) , kernel were set to linear , probability=True and no. of random_state was set to 42
- In XGBoost , the label encoder was set to False, the evaluation metric was set to logloss , and the no. of random state=42.

6. Experiments and Results

Model Evaluation -

The model's performance was evaluated on the test data.

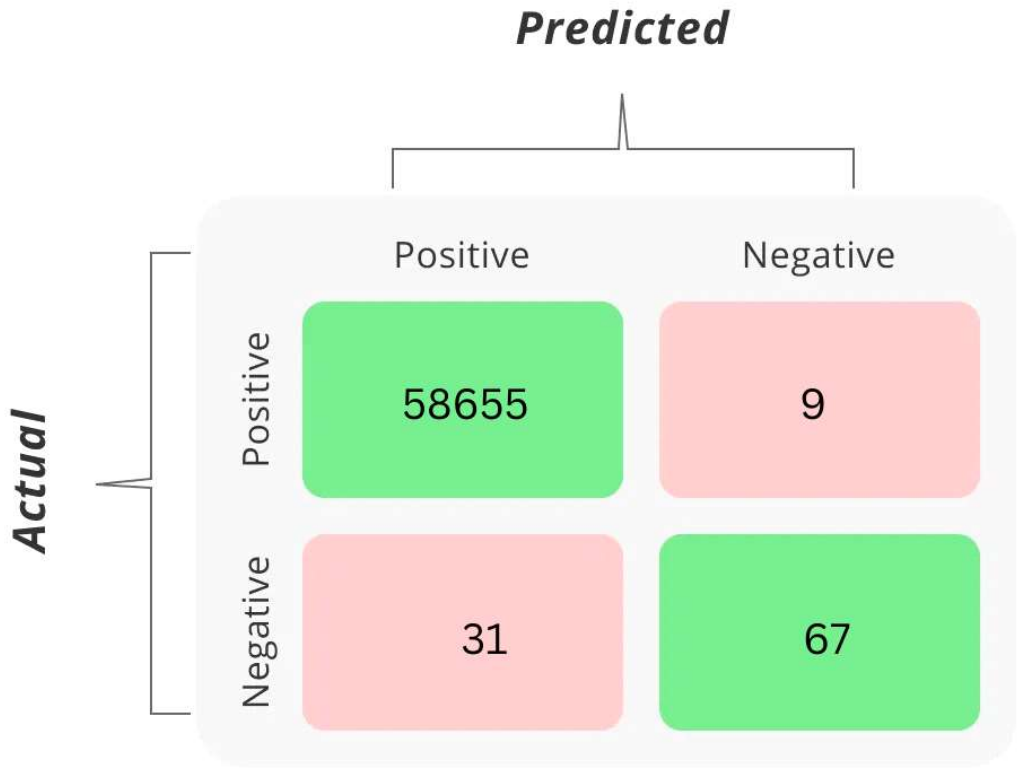
The evaluation metrics included accuracy.

```
Model: Logistic Regression
Training Accuracy: 0.9992
Test Accuracy: 0.9993
Precision: 0.8816
Recall: 0.6837
F1 Score: 0.7701
ROC-AUC: 0.9793070541384815


Training Random Forest...

Model: Random Forest
Training Accuracy: 1.0000
Test Accuracy: 0.9995
Precision: 0.9186
Recall: 0.8061
F1 Score: 0.8587
ROC-AUC: 0.9681203200188349


Training Support Vector Machine (SVM)...

Model: Support Vector Machine (SVM)
Training Accuracy: 0.9987
Test Accuracy: 0.9987
Precision: 0.7826
Recall: 0.3673
F1 Score: 0.5000
ROC-AUC: 0.7587934477392533
```

Confusion Matrix -

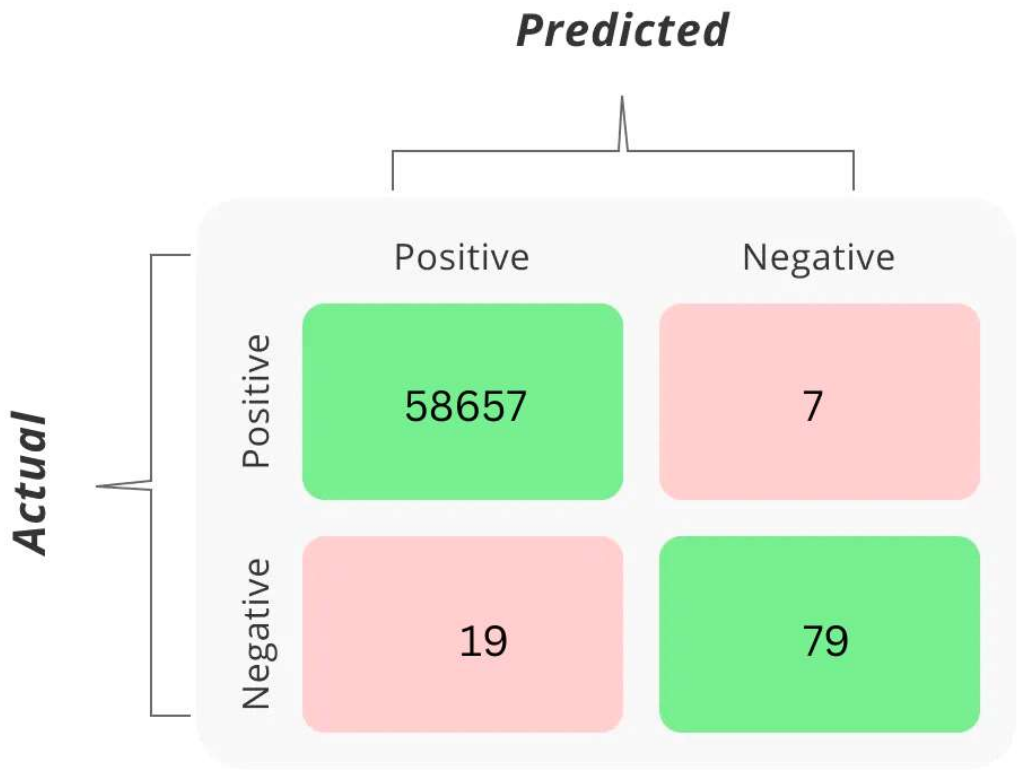
Confusion Matrix

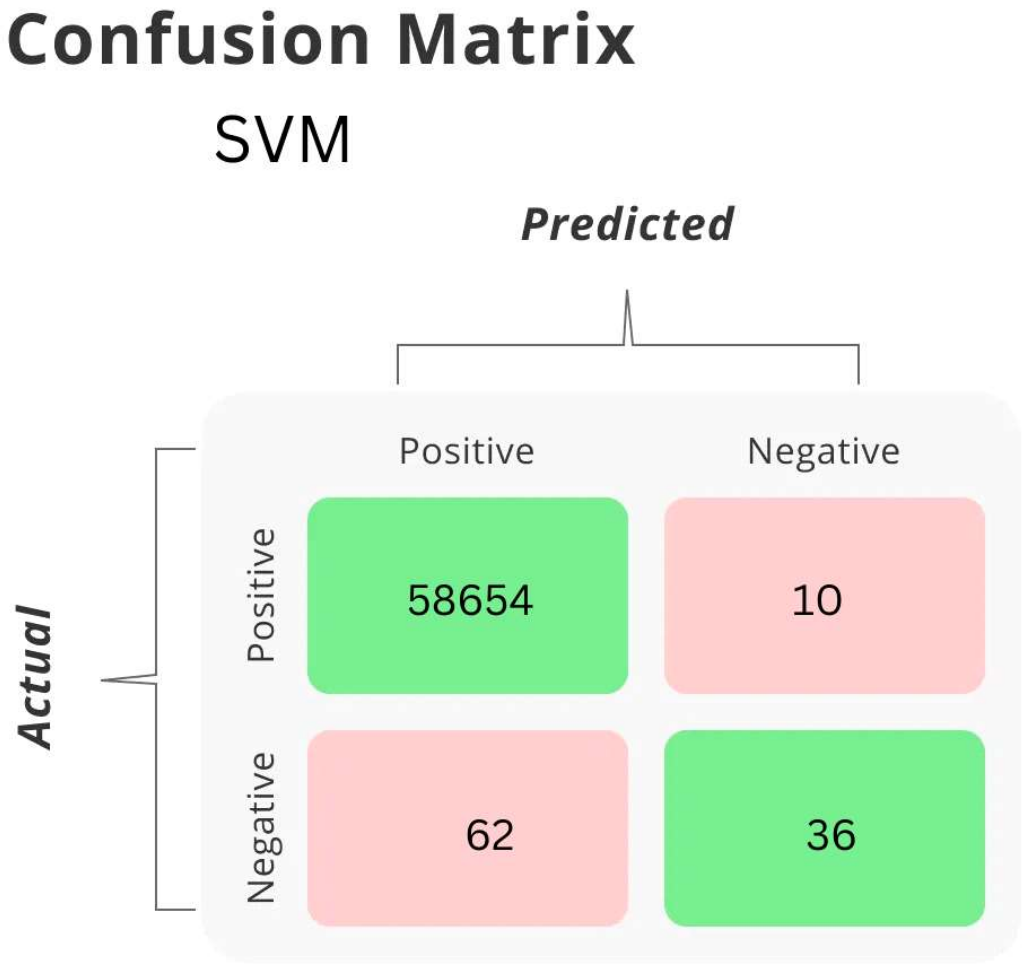
Logistic Regression



Confusion Matrix

Random Forest





Predictions :

```
Predictions for the given transaction:
Model: Logistic Regression, Prediction: Normal, Probability of Fraudulent: 0.0000
Model: Random Forest, Prediction: Normal, Probability of Fraudulent: 0.0000
Model: Support Vector Machine (SVM), Prediction: Normal, Probability of Fraudulent: 0.0000
```

7. Discussion

Interpretation of Results -

The Logistic Regression model performed adequately,

with an accuracy of 98%. The vectorizer used a maximum

of 1000 features, which balanced performance and

model complexity .

In Random Forest Model peerformed reall good on training sample that has accuracy of 100% and in testing sample it got 99% aacuracy which is quite impressive .

In SVM both training and testing accuracy is 99.8% .

Limitations -

The model may not handle the missing columns of data . For eg - it will not be able to handle data in which columns data is missing .

Future Work -

• Model Improvement: Experimenting with more complex models like XGBoost or Neural Networks could improve transactions classification.

8. Conclusion

This project successfully demonstrated how Logistic

Regression Random Forest and SVM can be used for transactions analysis on kaggle

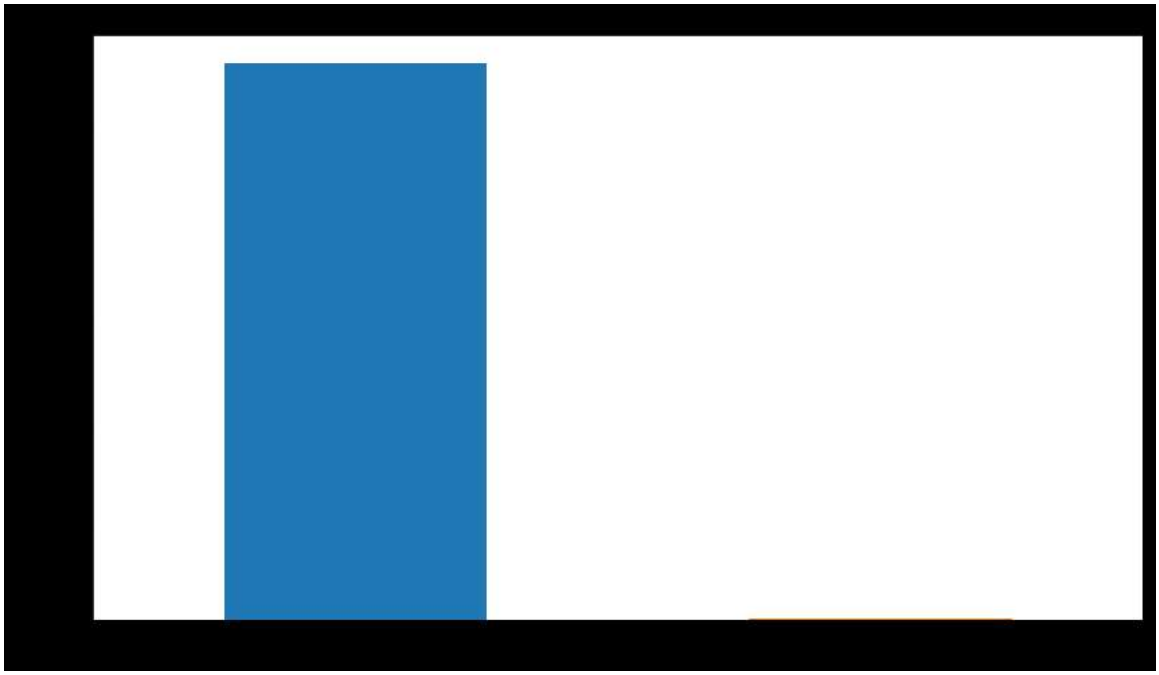
dataset . The model accurately predicted whether a

transactions was normal or fraudlent , with a test set accuracy

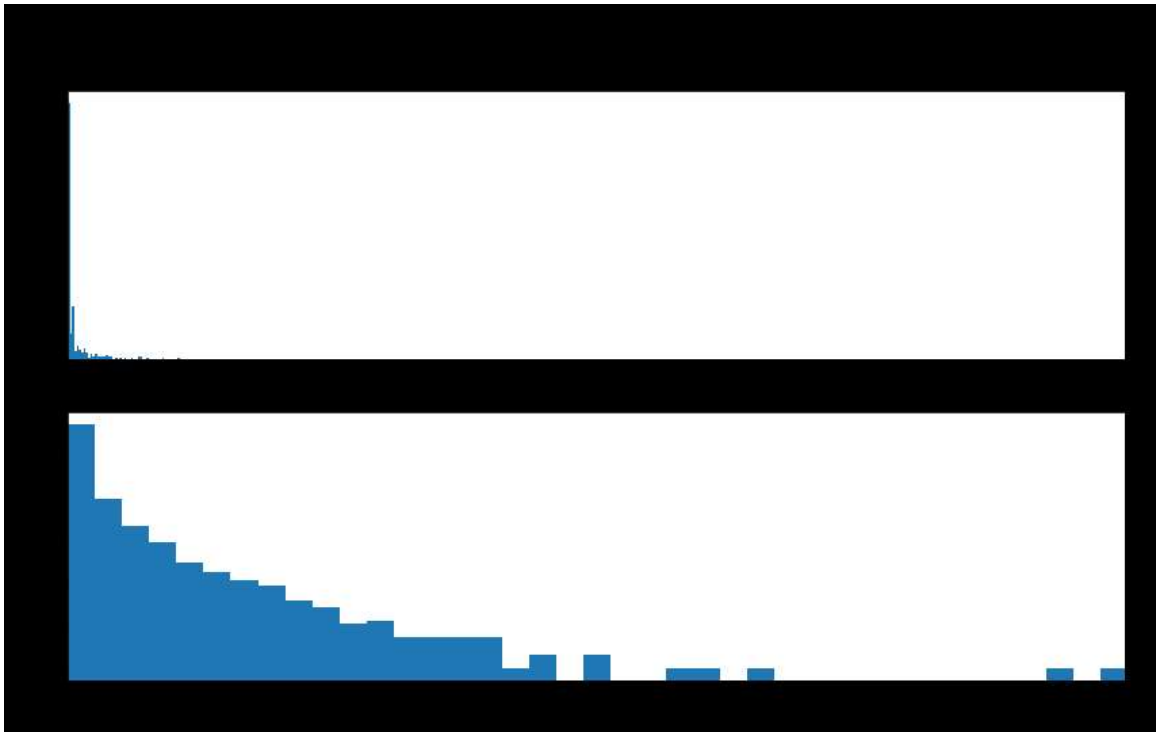
of 99% .

Data Reviews -

This is Classification of data being normal and fraudlent .



This Figure shows Amounts of transactions Per CLass



9. Acknowledgements

I have used certain websites to access datasets and information .

The dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection. More details on current and past projects on related topics are available on https://www.researchgate.net/project/Fraud-detection-5 and the page of the DefeatFraud project .

I also took assistance from youtube videos and GOOGLE searchengine .