PROJECT REPORT

Find Default (Prediction of Credit Card fraud)—capstone project

(Done by M. Sai Meghana)

1. Introduction

Credit card fraud, the unauthorized use of someone else's credit card for transactions, poses significant risks to financial institutions and customers. This project aims to develop a classification model to predict fraudulent transactions using a dataset of European cardholders' transactions from September 2013. With 492 frauds out of 284,807 transactions, the data is highly imbalanced. The project involves exploratory data analysis, data cleaning, handling imbalanced data, feature engineering, model selection, training, validation, and deployment, ultimately providing a robust solution to identify and prevent credit card fraud effectively.

2. Project Outline

- Exploratory Data Analysis: Analyze and understand the data to identify patterns, relationships, and trends in the data by using Descriptive Statistics and Visualizations.
- Data Cleaning: This might include standardization, handling the missing values and outliers in the data.
- Dealing with Imbalanced data: This data set is highly imbalanced. The data should be balanced using the appropriate methods before moving onto model building.
- Feature Engineering: Create new features or transform the existing features for better performance of the ML Models.
- Model Selection: Choose the most appropriate model that can be used for this project.
- Model Training: Split the data into train & test sets and use the train set to estimate the best model parameters.
- Model Validation: Evaluate the performance of the model on data that was not used during the training process. The goal is to estimate the model's ability to generalize to new, unseen data and to identify any issues with the model, such as overfitting.
- Model Deployment: Model deployment is the process of making a trained machine learning model available for use in a production environment

3. Problem Statement:

A credit card is one of the most used financial products to make online purchases and payments. Though the Credit cards can be a convenient way to manage your finances, they can also be risky. Credit card fraud is the unauthorized use of someone else's credit card or credit card information to make purchases or withdraw cash. It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase. 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. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

4. Project Work Overview

Our dataset exhibits significant class imbalance, with the majority of transactions being non-fraudulent (99.82%). This presents a challenge for predictive modeling, as algorithms may struggle to accurately detect fraudulent transactions amidst the overwhelming number of legitimate ones. To address this issue, we employed various techniques such as under sampling, oversampling, and synthetic data generation.

Under sampling

We utilized the Near Miss technique to balance the class distribution by reducing the number of instances of non-fraudulent transactions to match that of fraudulent transactions. This approach helped in mitigating the effects of class imbalance. Our attempt to address class imbalance using the NearMiss technique did not yield satisfactory results. Despite its intention to balance the class distribution, the model's performance was suboptimal. This could be attributed to the loss of valuable information due to the drastic reduction in the majority class instances, leading to a less representative dataset. As a result, the model may have struggled to capture the intricacies of the underlying patterns in the data, ultimately affecting its ability to accurately classify fraudulent transactions.

Oversampling:

To further augment the minority class, we applied the SMOTETomek method with a sampling strategy of 0.75. This resulted in a more balanced dataset, enabling the models to better capture the underlying patterns in fraudulent transactions.

Machine Learning Models:

After preprocessing and balancing the dataset, we trained several machine learning models, including:

Logistic Regression K-Nearest Neighbors (KNN) Random Forest Classifier AdaBoost Classifier XGBoost Classifier

Evaluation Metrics:

We evaluated the performance of each model using various metrics such as accuracy, precision, recall, and F1-score. Additionally, we employed techniques like cross-validation and hyperparameter tuning to optimize the models' performance.

Model Selection:

Among the various models and balancing methods experimented with, the XGBoost model stands out as the top performer when using oversampling techniques. Despite the inherent challenges posed by imbalanced datasets, the XGBoost algorithm demonstrates robustness and effectiveness in capturing the underlying patterns associated with fraudulent transactions. By generating synthetic instances of the minority class through oversampling methods like SMOTETomek, the XGBoost model achieves a more balanced representation of the data, enabling it to learn and generalize better to unseen instances. This superior performance underscores the importance of leveraging advanced ensemble techniques like XGBoost, particularly in the context of imbalanced datasets characteristic of credit card fraud detection**.**

5 .Technical Stack

Programming Language & Frameworks/Libraries,database:

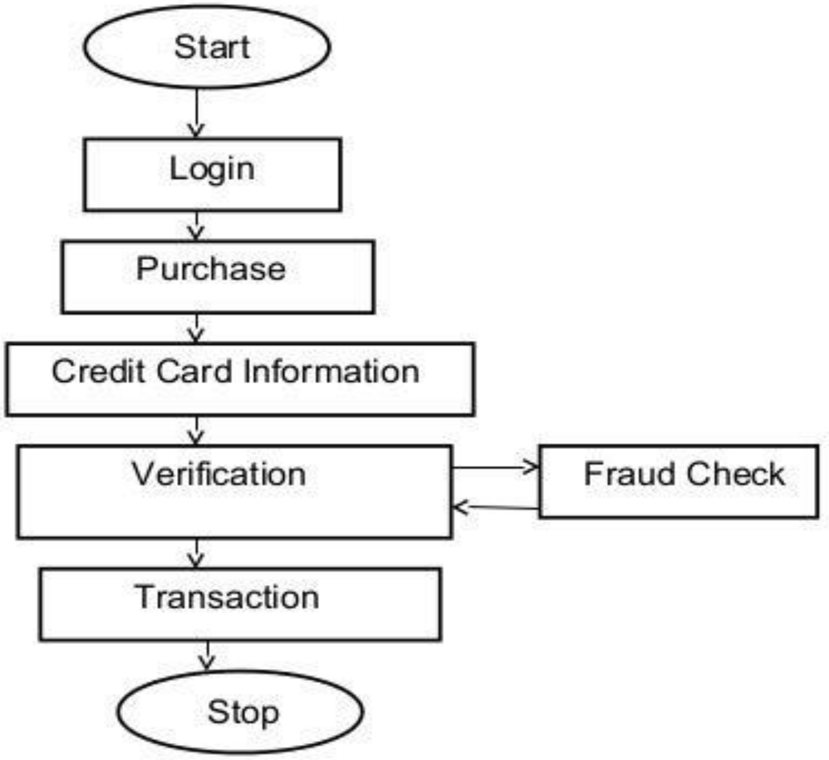
Python: for its robust libraries and ease of use

Pandas : For data parsing and storage.

Numpy : For arrays and numerical operations

Logistic Regression model : For identification

ARCHITECTURE DIAGRAM



Logistic Regression Model

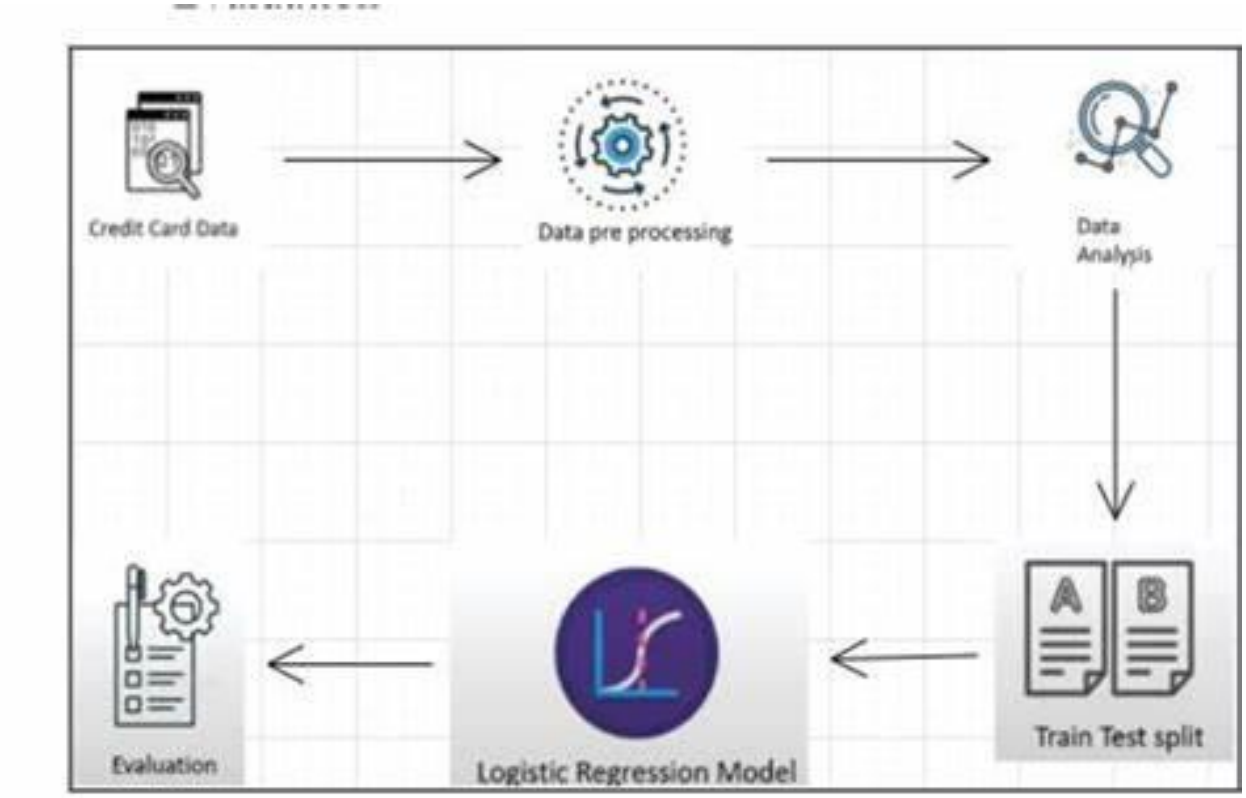


Fig. 1. Architecture Diagram (Source: Made by Author)

7 . Project Set-Up:

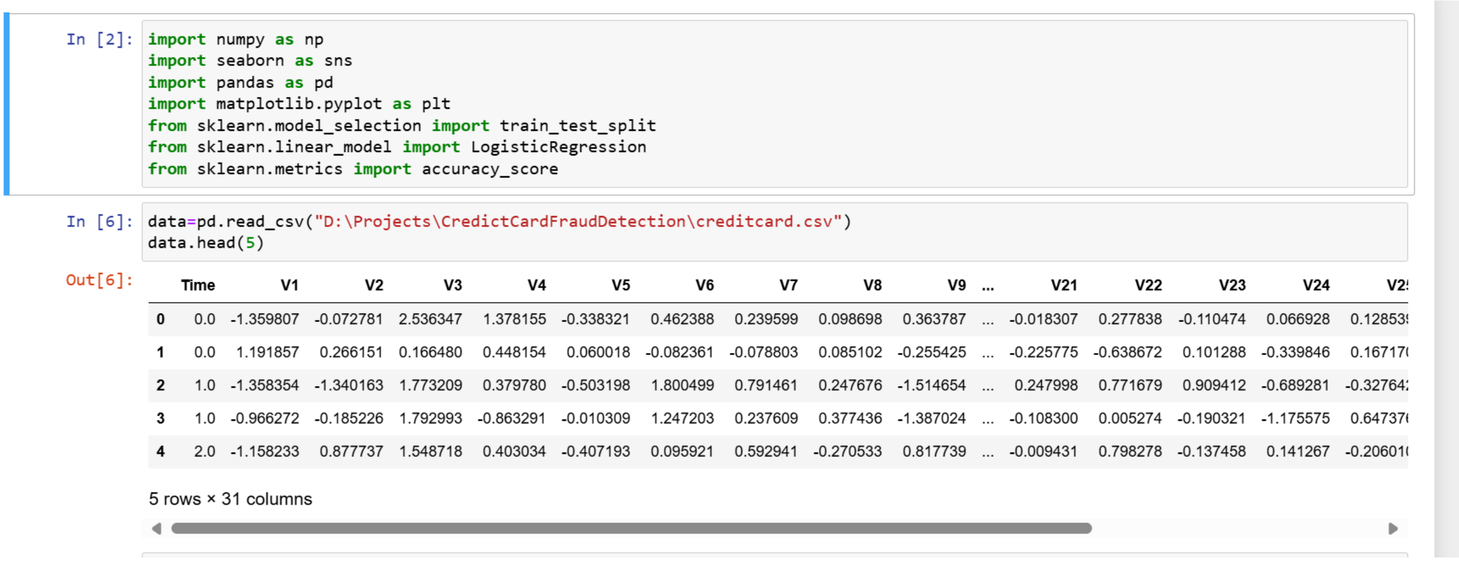
Software And Tools Required for the Development of this Project:

1. VSCodeIDE: Can also be require for evaluation
2. GitCLI
3. Github Account
4. Heroku Account
5. Postman : Can also be require for evaluation

Steps to Follow to Reproduce the Results:

- Open downloaded folder from the repository into VSCodeIDE.
- Create New environment using commands-conda install -p yourEnvname python==3.10conda activate your Envname
- Run following command to get all the required packages for this project-pip install -r requirements.txt
- Run all the jupyter notebooks (exclude training models, use model pickle files instead)from the 'notebooks' folder if want to reproduce the results of this project work.
- Run app.py file using command from terminal-python app.py

Import libraries



In [2]:
```
import numpy as np
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

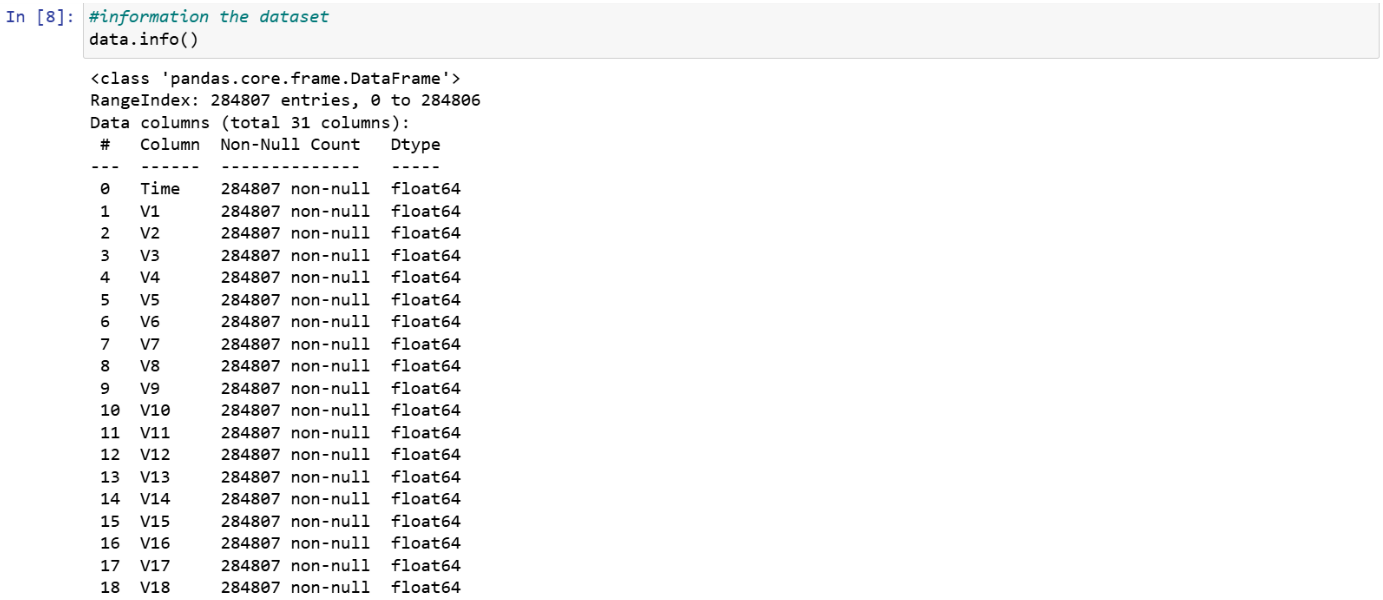
In [6]:
```
data=pd.read_csv("D:\Projects\CredictCardFraudDetection\creditcard.csv")
data.head(5)
```

Out[6]:

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V2
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.12853
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.16717
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.32764
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.64737
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.20601

5 rows × 31 columns

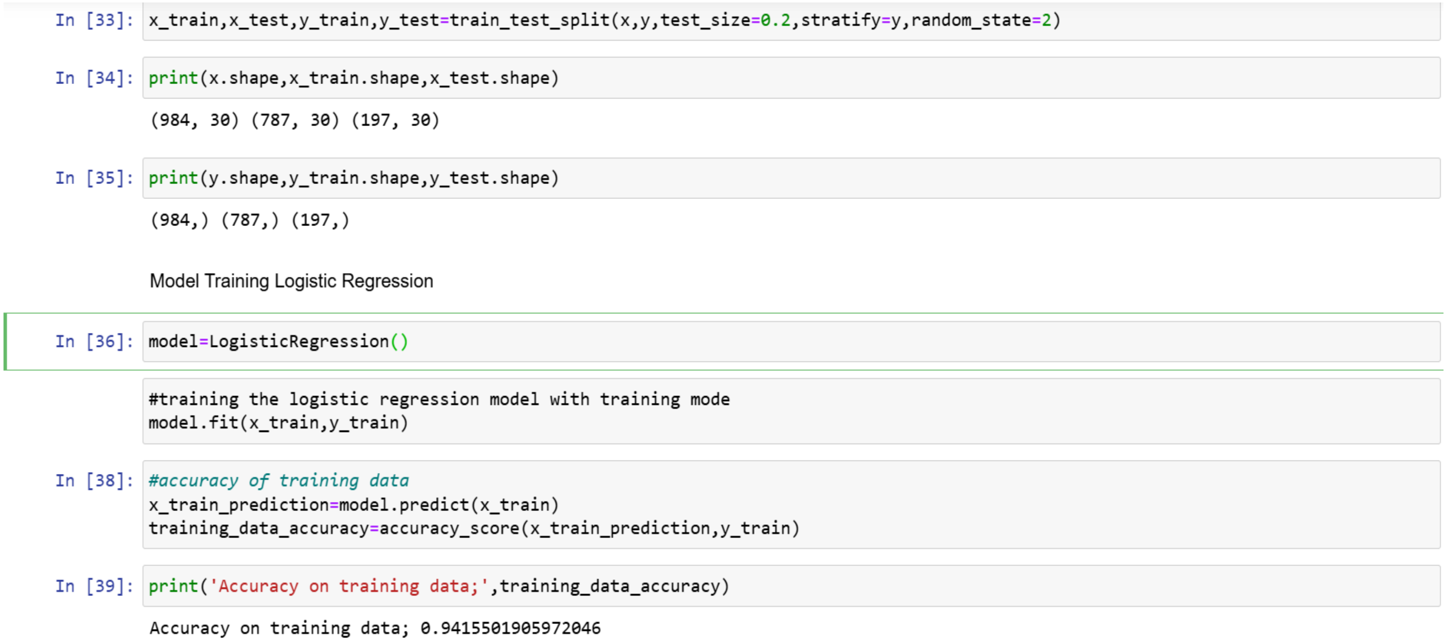
Dataset Information



In [8]:
```
#information the dataset
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 284807 entries, 0 to 284806
Data columns (total 31 columns):
 #   Column  Non-Null Count   Dtype  
---  ------  --------------   -----  
 0   Time    284807 non-null  float64
 1   V1      284807 non-null  float64
 2   V2      284807 non-null  float64
 3   V3      284807 non-null  float64
 4   V4      284807 non-null  float64
 5   V5      284807 non-null  float64
 6   V6      284807 non-null  float64
 7   V7      284807 non-null  float64
 8   V8      284807 non-null  float64
 9   V9      284807 non-null  float64
 10  V10     284807 non-null  float64
 11  V11     284807 non-null  float64
 12  V12     284807 non-null  float64
 13  V13     284807 non-null  float64
 14  V14     284807 non-null  float64
 15  V15     284807 non-null  float64
 16  V16     284807 non-null  float64
 17  V17     284807 non-null  float64
 18  V18     284807 non-null  float64
```

Training and Testing the Dataset and Accuracy Score of the Project



```
In [33]: x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,stratify=y,random_state=2)

In [34]: print(x.shape,x_train.shape,x_test.shape)

         (984, 30) (787, 30) (197, 30)

In [35]: print(y.shape,y_train.shape,y_test.shape)

         (984,) (787,) (197,)
```

Model Training Logistic Regression

```
In [36]: model=LogisticRegression()

         #training the logistic regression model with training mode
         model.fit(x_train,y_train)

In [38]: #accuracy of training data
         x_train_prediction=model.predict(x_train)
         training_data_accuracy=accuracy_score(x_train_prediction,y_train)

In [39]: print('Accuracy on training data;',training_data_accuracy)

         Accuracy on training data; 0.9415501905972046
```

7 . Future Enhancement

Future enhancements for this credit card fraud detection project could include implementing real-time detection systems, incorporating advanced machine learning models like deep learning and ensemble methods, and continuously retraining the model with new data to adapt to evolving fraud patterns. Additionally, integrating user behavior data and employing explainable AI techniques can improve model performance and transparency.

8 . Conclusion

This project successfully developed a robust machine learning model to predict fraudulent credit card transactions from a highly imbalanced dataset. Key steps included data exploration, cleaning, balancing, feature engineering, and employing models like Logistic Regression, Decision Tree, Random Forest, Gradient Boosting Machines, and XGBoost. These models showed strong performance in identifying fraud. Future improvements could include real-time detection, advanced modeling techniques, and continuous model updates to enhance accuracy and adaptability. Overall, this project demonstrates the importance of sophisticated data science methods in detecting and preventing credit card fraud, thereby safeguarding financial institutions and their customers.

9. References

1. Credit Card Fraud Detection Dataset: Available on Kaggle. [Link](https://www.kaggle.com/mlg-ulb/creditcardfraud)

2. Machine Learning Models:

- Pedregosa et al. (2011). "Scikit-learn: Machine Learning in Python". [Link](https://jmlr.org/papers/v12/pedregosa11a.html)

- Chen & Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". [Link](https://dl.acm.org/doi/10.1145/2939672.2939785)

3. Handling Imbalanced Data:

- Chawla et al. (2002). "SMOTE: Synthetic Minority Over-sampling Technique". [Link](https://www.jair.org/index.php/jair/article/view/10302)

4. Feature Engineering:

- Kuhn & Johnson (2019). "Feature Engineering and Selection: A Practical Approach for Predictive Models". [Link](https://www.crcpress.com/Feature-Engineering-and-Selection-A-Practical-Approach-for-Predictive-Models/Kuhn-Johnson/p/book/9781461468486)

5. Model Deployment:

- Müller & Guido (2016). "Introduction to Machine Learning with Python". [Link](https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/)

6. Real-Time Detection:

- "Real-Time Big Data Analytics: Emerging Architecture". O'Reilly Media. [Link](https://www.oreilly.com/library/view/real-time-big-data/9781491941497/)