CREDIT CARD
FRAUD DETECTION

Where data science becomes the guardian of your wallet.

BML MUNJAL UNIVERSITY™
FROM HERE TO THE WORLD

PRESENTED BY

Group 3

Harshita Rupani

Ashish Saharia

Robin Yadav

Khushi

MENTORED BY

Dr. Hirdesh Kumar Pharasi

INTRODUCTION TO PROJECT

- Credit card fraud is a pervasive problem in today's financial landscape, causing significant losses to both consumers and businesses.
- This data science project focuses on using analytics and machine learning, to safeguard financial transactions fraudulent activities.

PROBLEM STATEMENT

- The challenge at hand is to develop a machine learning model for credit card fraud detection, prioritizing the reduction of false positives(cases where the model predicted that a transaction was fraudulent, but it was actually not fraudulent.).
- This project seeks to enhance transaction security by rapidly detecting unauthorized charges, with the overarching goal of mitigating financial losses and maintaining the trust of customers .

OBJECTIVE

- The Primary objective of the Credit Card Fraud Detection Project is to develop a robust and accurate model utilizing data science techniques to discern between legitimate and fraudulent transactions.
- Leveraging features such as transaction amount, type, and anonymized variables, the model aims to uncover patterns associated with fraudulent activities.

LITERATURE SURVEY - RESEARCH PAPER 1

Title Of Paper	Author of Paper	Summary
Credit Card Fraud Detection Based on Machine and Deep Learning	Hassan Najadat	In the Research Paper, they have performed several machine and deep learning models to detect whether an online transaction is legitimate or fraud on the IEEE-CIS Fraud Detection dataset as well built their model. They also tested several methods to deal with highly imbalanced datasets including undersampling, oversampling and SMOTE. Set of evaluation metrics used to evaluate the performance of the models

LITERATURE SURVEY - RESEARCH PAPER 2

Title Of Paper	Author of Paper	Summary
Credit Card Fraud Detection - Machine Learning methods	Dejan Varmedja, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, Andras Anderla	The main goal of the Research Paper was to compare certain machine learning algorithms for detection of fraudulent transactions. Hence, comparison was made and it was established that Random Forest algorithm gives the best results i.e. best classifies whether transactions are fraud or not. This was established using different metrics, such as recall, accuracy and precision. For this kind of problem, it is important to have recall with high value. Feature selection and balancing of the dataset have shown to be extremely important in achieving significant results.

LITERATURE SURVEY - RESEARCH PAPER 3

Title Of Paper	Author of Paper	Summary
Credit Card Fraud Detection in e-Commerce: An Outlier Detection Approach	Utkarsh Porwal , mruthi Mukun	In the Paper they propose a method that shows tremendous potential in identifying outliers by assigning a consistency score to each data point. The proposed method assumes no prior knowledge of the outliers. They showed that application of proposed method in different scenarios such as to make recommendation for potential outliers for further investigation with high precision and to create training sets for novelty detection algorithms.

DISCUSSION OF DATASET 1

Credit Card Fraud Detection: This dataset uses a September 2013 European credit card dataset, spanning two days. With only 0.172% fraudulent transactions in 284,807, it focuses on PCA-derived principal components, including 'Time' and 'Amount.' Due to confidentiality, original features are undisclosed. Features V1 to V28 represent components, 'Time' measures transaction intervals, and 'Amount' indicates the transaction amount. The 'Class' variable distinguishes fraud (Class 1) from non-fraud (Class 0). This dataset is vital for recognizing and addressing credit card fraud, emphasizing customer protection.

DISCUSSION OF DATASET 2

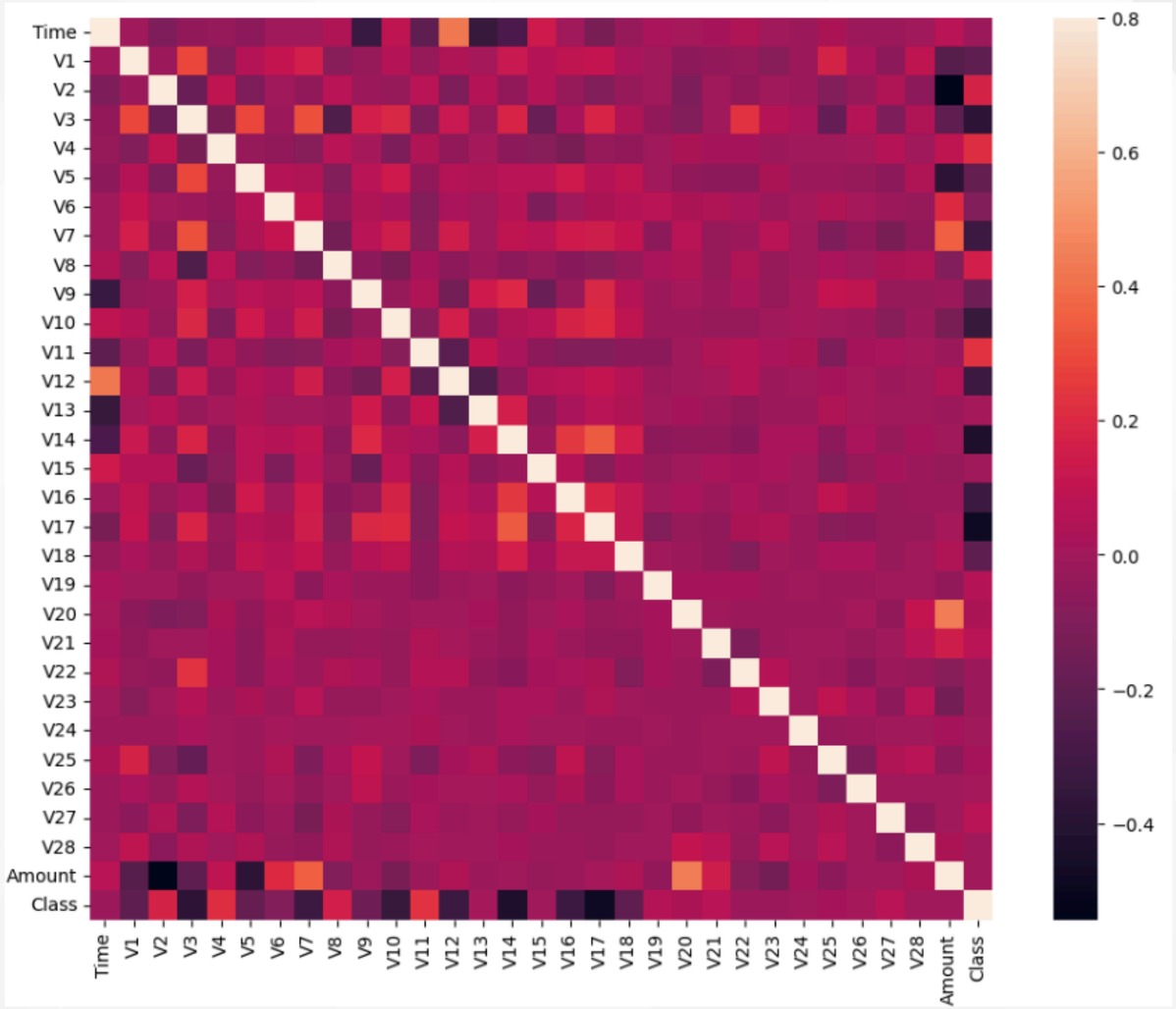
Credit Card Fraud: This dataset includes key features essential for fraud detection in digital transactions. Features like "distance_from_home" and "distance_from_last_transaction" offer spatial context. The "ratio_to_median_purchase_price" compares transaction prices to the median, revealing transaction patterns. "Repeat_retailer" flags transactions from the same retailer, indicating potential behavior patterns. "Used_chip" and "used_pin_number" detail transaction methods, whether through chip or PIN. "Online_order" identifies online transactions. Lastly, the "fraud" indicator categorizes transactions as fraudulent. These features provide a foundation for building a robust fraud detection system.

METHODOLOGY - DATASET 1 AND 2

- Library Import and Data Load: Import libraries, load credit card dataset from CSV.
- Data Exploration: Display dataset details, shape, and summary statistics. Separate into fraud and valid transactions, emphasizing fraud cases.
- Exploratory Analysis: Uncover insights, patterns, and relationships in the data.
- Descriptive Analysis: Summarize key data characteristics, highlighting fraud cases and column relations.
- Confusion Matrix Evaluation: Assess model performance using a confusion matrix.
- ML Model Implementation: Implement various Machine learning models:
- Naïve Bayes: Classify with probabilistic reasoning and feature independence.
- K-NN Classifier: Categorize using distance metrics and neighbor similarities.
- Random Forest: Boost accuracy with ensemble learning.
- Results Presentation: Showcase and analyze model effectiveness in fraud detection.

RESULT OF 1ST DATASET

Heat Map of all the Variables :



RESULT OF 1ST DATASET

Expected Result :

RF model obtained following results

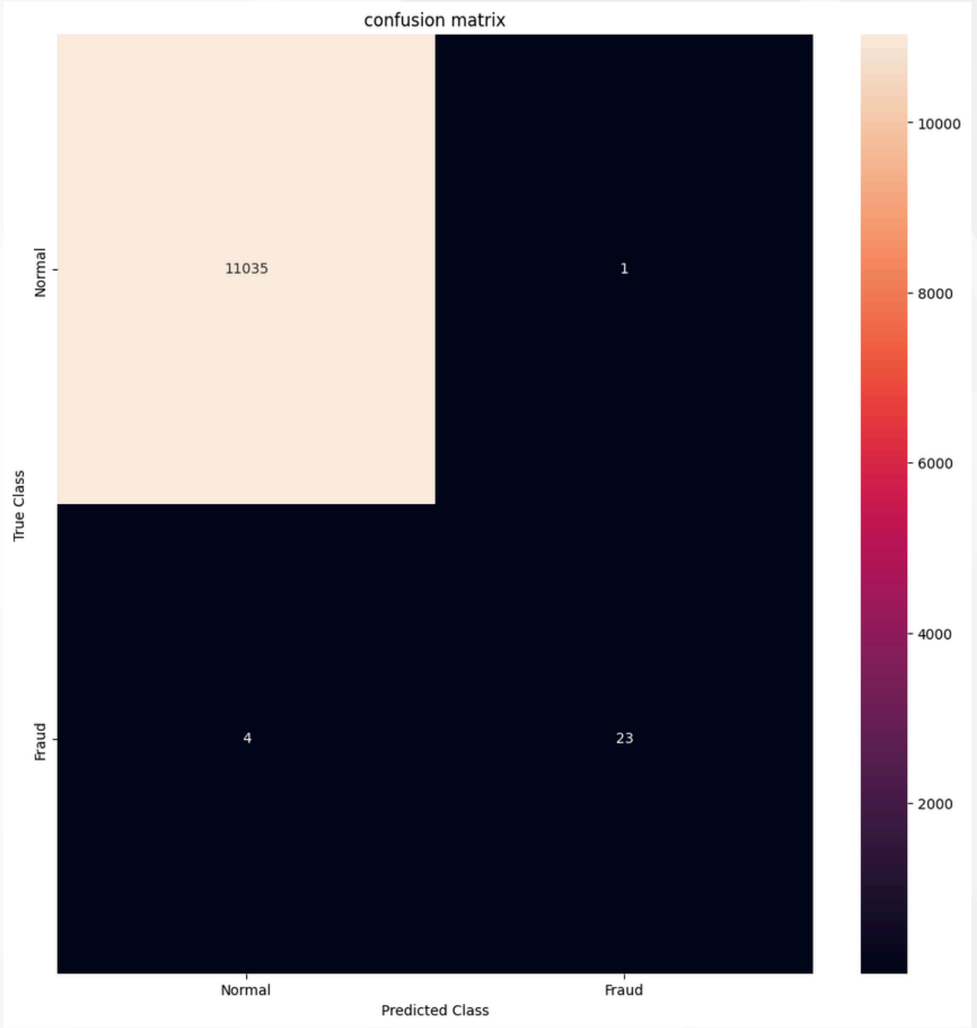
- precision: 96.38%,
- recall: 81.63%,
- accuracy: 99.96%.

TABLE 3:CONFUSION MATRIX FOR RF

		Predicted	
		0	1
Actual	0	56861	3
	1	18	80

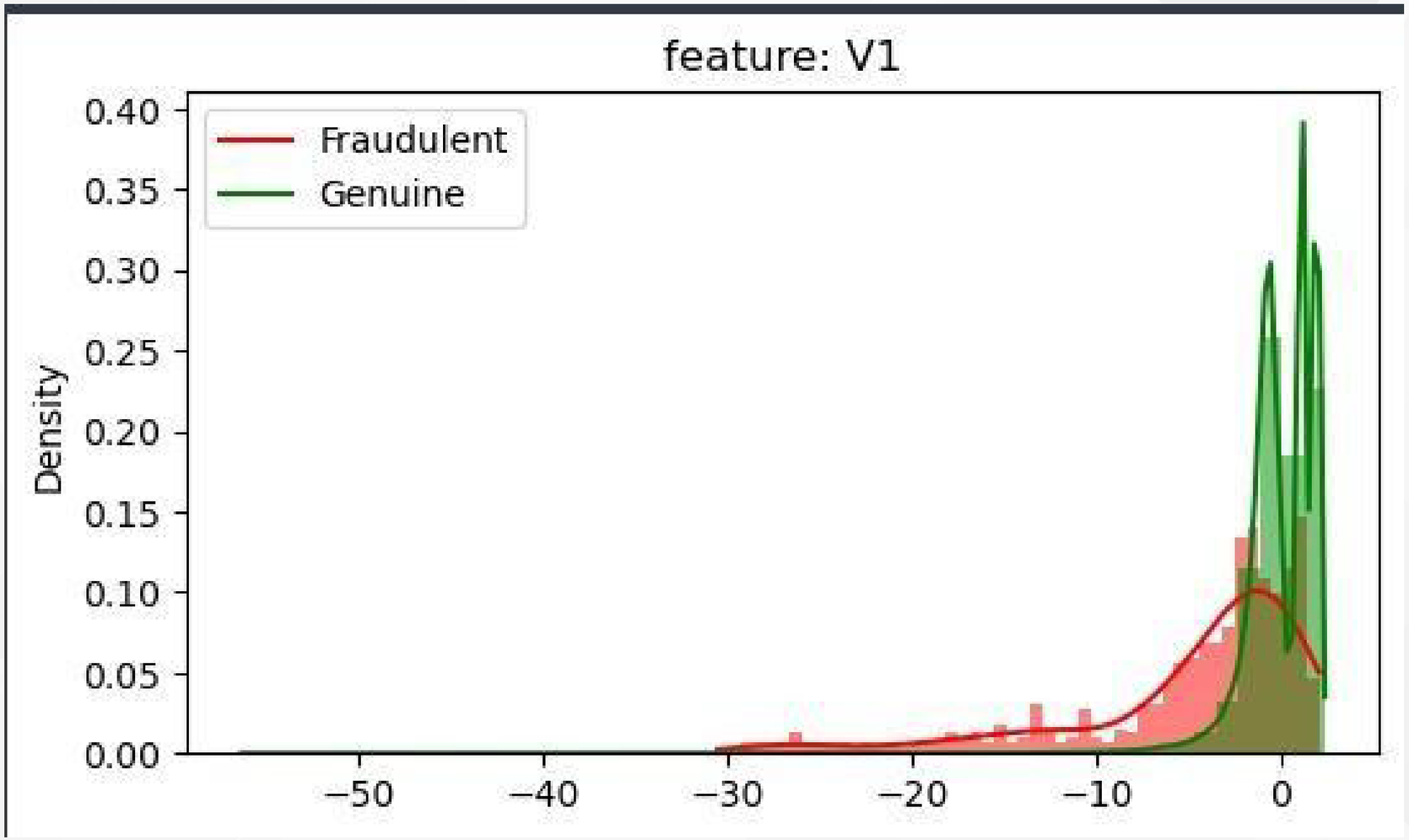
Actual Outcome:

```
[ ] acc = accuracy_score(y_test, pred)
    acc

    0.9997654172965647

[ ] prec = precision_score(y_test, pred)
    prec

    0.9672131147540983

[ ] rec = recall_score(y_test, pred)
    rec

    0.8939393939393939

[ ] f1 = f1_score(y_test, pred)
    f1

    0.9291338582677166

[ ] mcc = matthews_corrcoef(y_test, pred)
    mcc
```



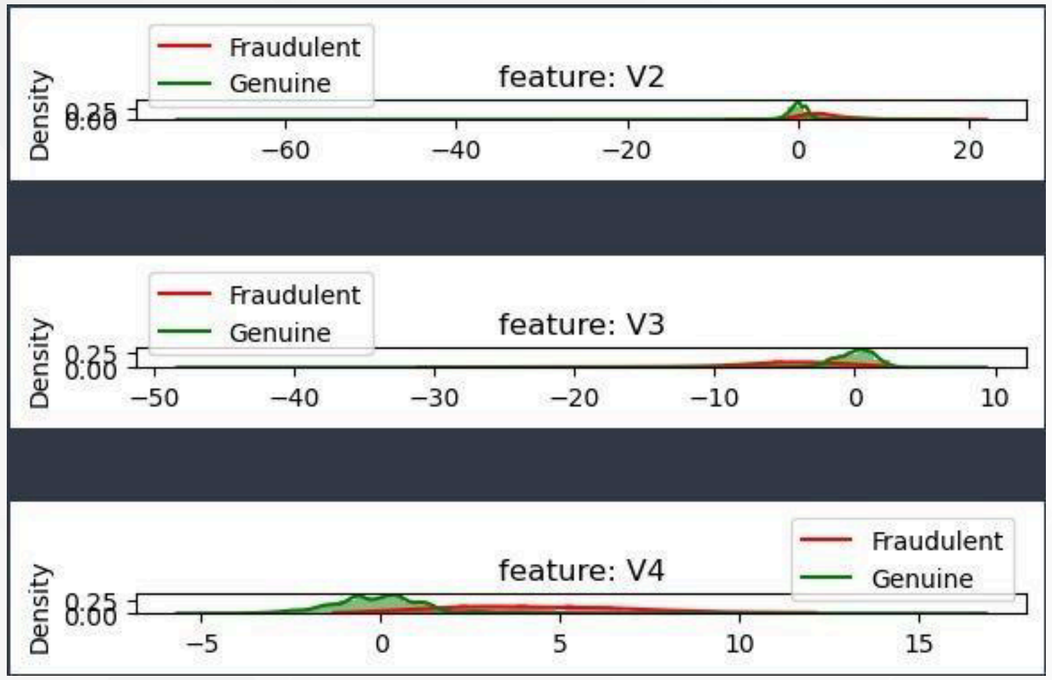
RESULT OF 1ST DATASET

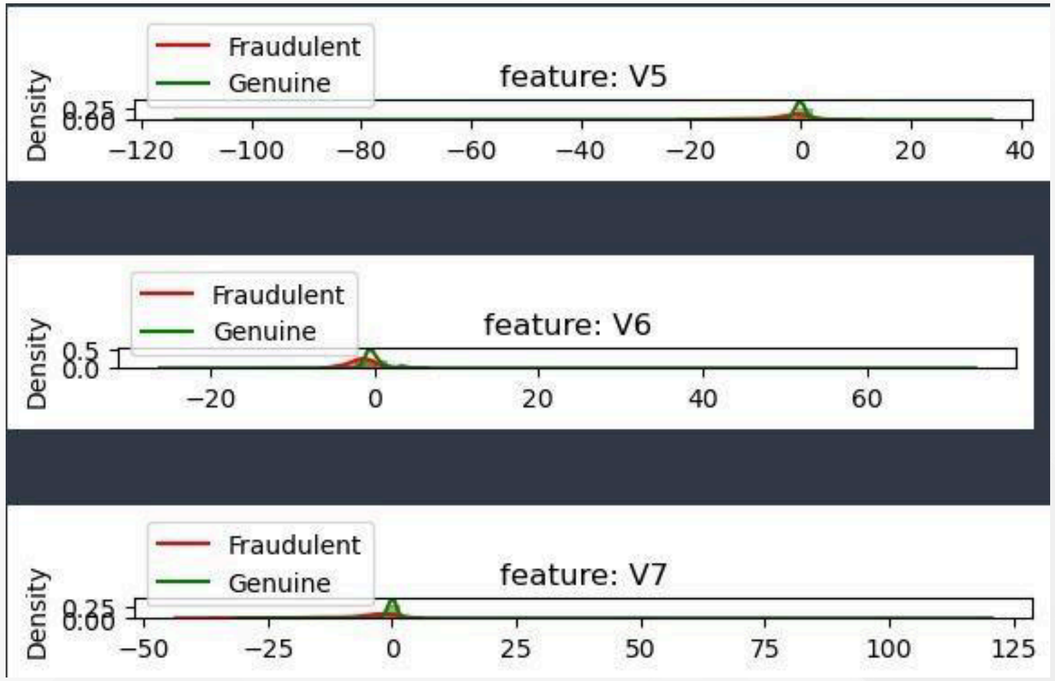
Density Graph for all features:



RESULT OF 1ST DATASET

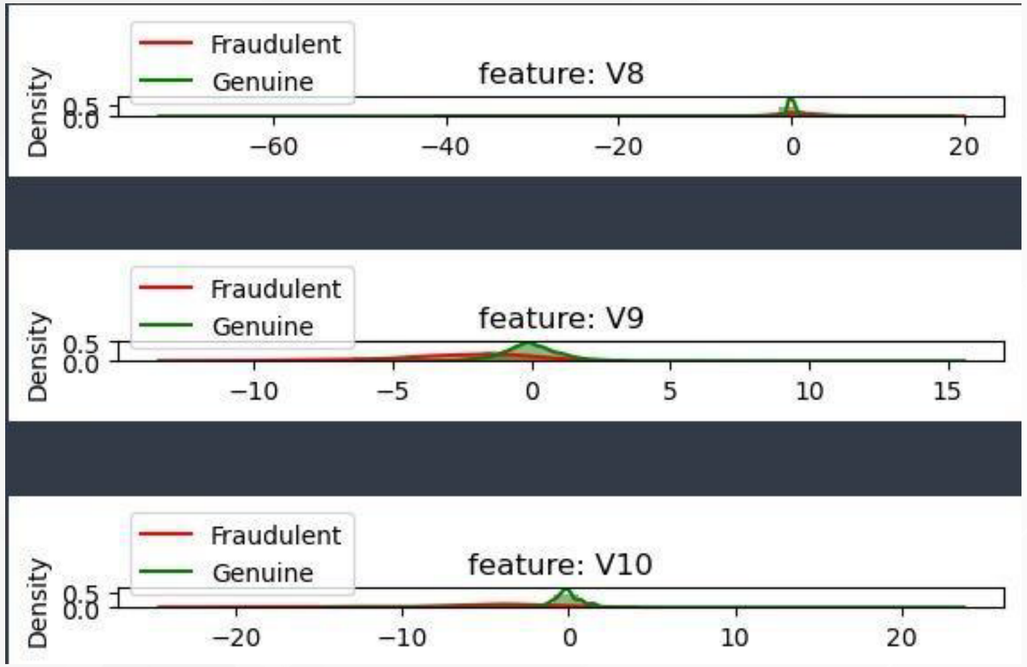
Density Graphs for all features:

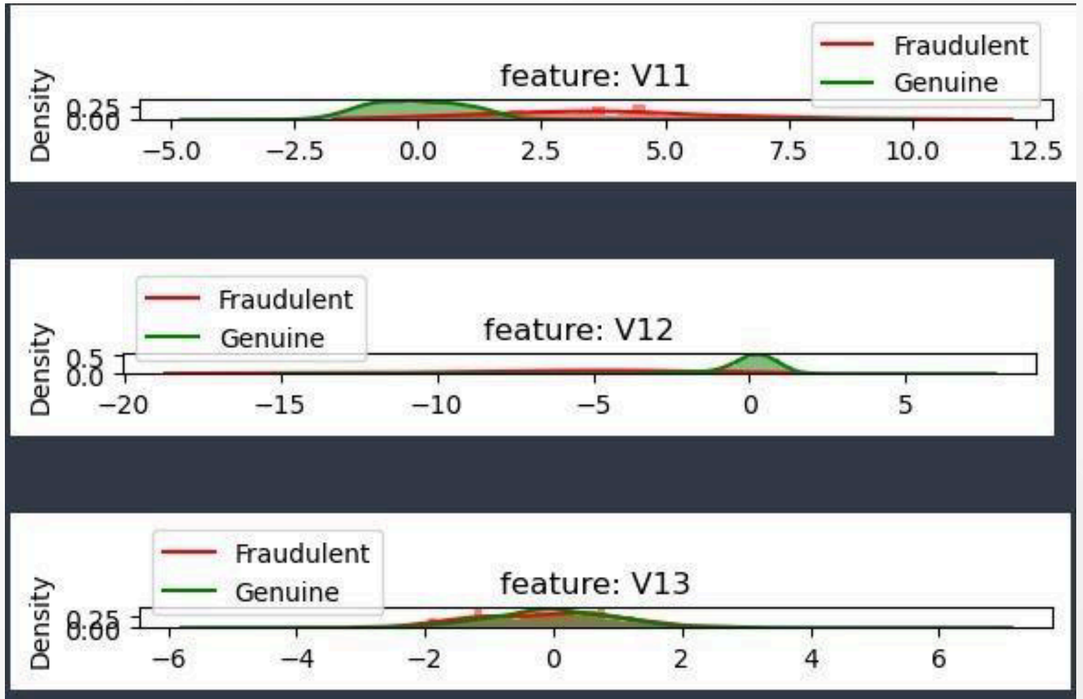




RESULT OF 1ST DATASET

Density Graphs for all features:





Similary, we infer from individual density graphs of all the 28 variables.

RESULT OF 1ST DATASET

Gaussian Naive Bayes:

Case 1: Drop List is Empty

```python
from sklearn.naive_bayes import GaussianNB
from sklearn.linear_model import LogisticRegression
# Case-NB-1 : do not drop anything
drop_list = []
X_train, X_test, y_train, y_test = split_data(df, drop_list)
y_pred, y_pred_prob = get_predictions(GaussianNB(), X_train, y_train, X_test)
print_scores(y_test,y_pred,y_pred_prob)
```

```
Index(['V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V8', 'V9', 'V10', 'V11',
       'V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19', 'V20', 'V21',
       'V22', 'V23', 'V24', 'V25', 'V26', 'V27', 'V28', 'Class', 'Time_Hr',
       'scaled_Amount'],
      dtype='object')
train-set size:  227845
test-set size:  56962
fraud cases in test-set:  98
train-set confusion matrix:
 [[222480   4971]
 [    69    325]]
test-set confusion matrix:
 [[55535  1329]
 [   15    83]]
recall score:  0.8469387755102041
precision score:  0.058781869688385266
f1 score:  0.10993377483443707
accuracy score:  0.9764053228468101
ROC AUC: 0.963247971529636
```

RESULT OF 1ST DATASET

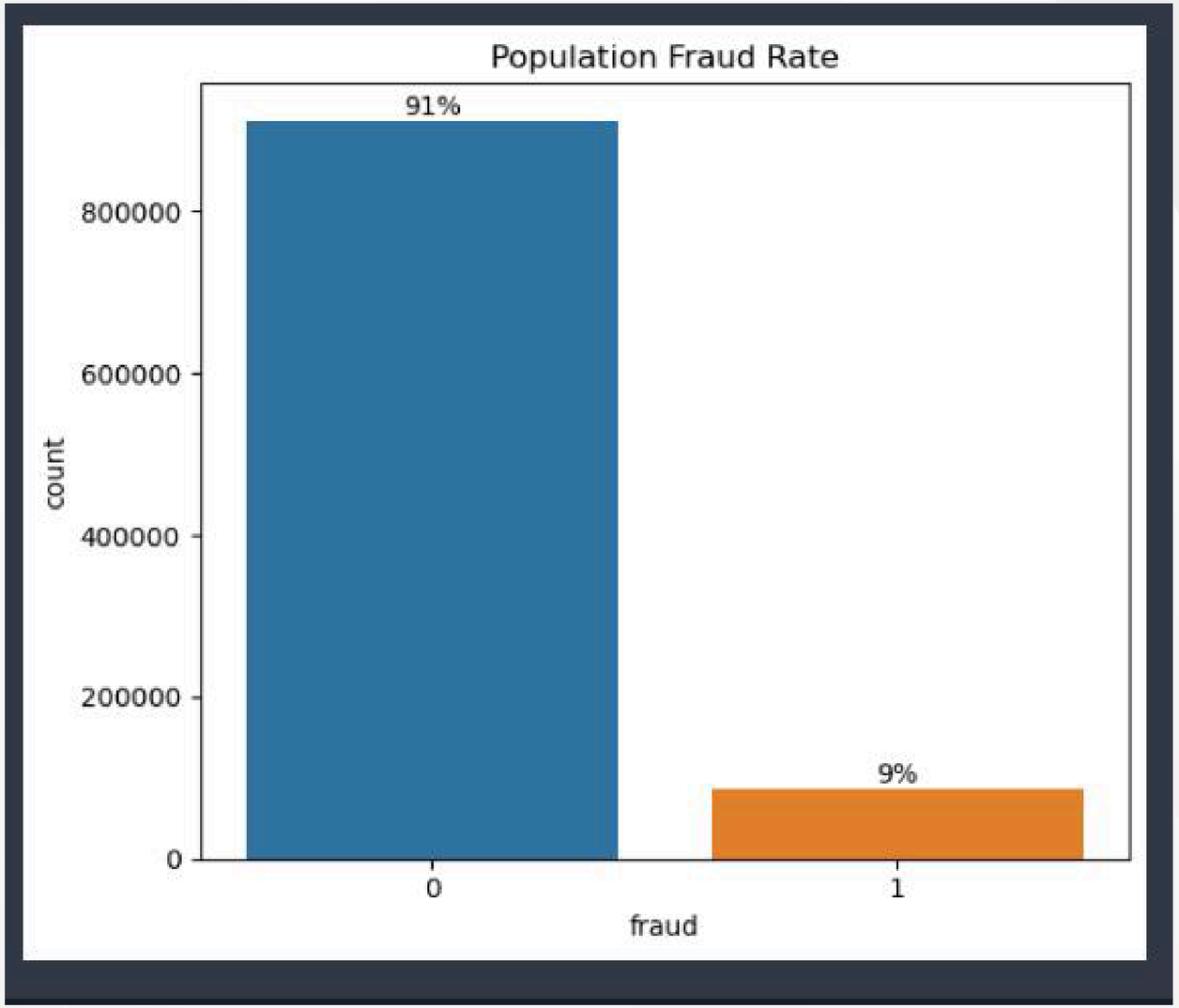
Case 2: Drop List contains features V28, V27, V26, V25, V24, V23, V22, V20, V15, V13, V8.

```
drop_list = ['V28','V27','V26','V25','V24','V23','V22','V20','V15','V13','V8']
X_train, X_test, y_train, y_test = split_data(df, drop_list)
y_pred, y_pred_prob = get_predictions(GaussianNB(), X_train, y_train, X_test)
print_scores(y_test,y_pred,y_pred_prob)
```

```
Index(['V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V9', 'V10', 'V11', 'V12',
       'V14', 'V16', 'V17', 'V18', 'V19', 'V21', 'Class', 'Time_Hr',
       'scaled_Amount'],
      dtype='object')
train-set size:  227845
test-set size:  56962
fraud cases in test-set:  98
train-set confusion matrix:
 [[223967   3484]
 [    61    333]]
test-set confusion matrix:
 [[55935   929]
 [   12    86]]
recall score:  0.8775510204081632
precision score:  0.08472906403940887
f1 score:  0.15453728661275834
accuracy score:  0.9834802148800955
ROC AUC: 0.9622034097825962
```

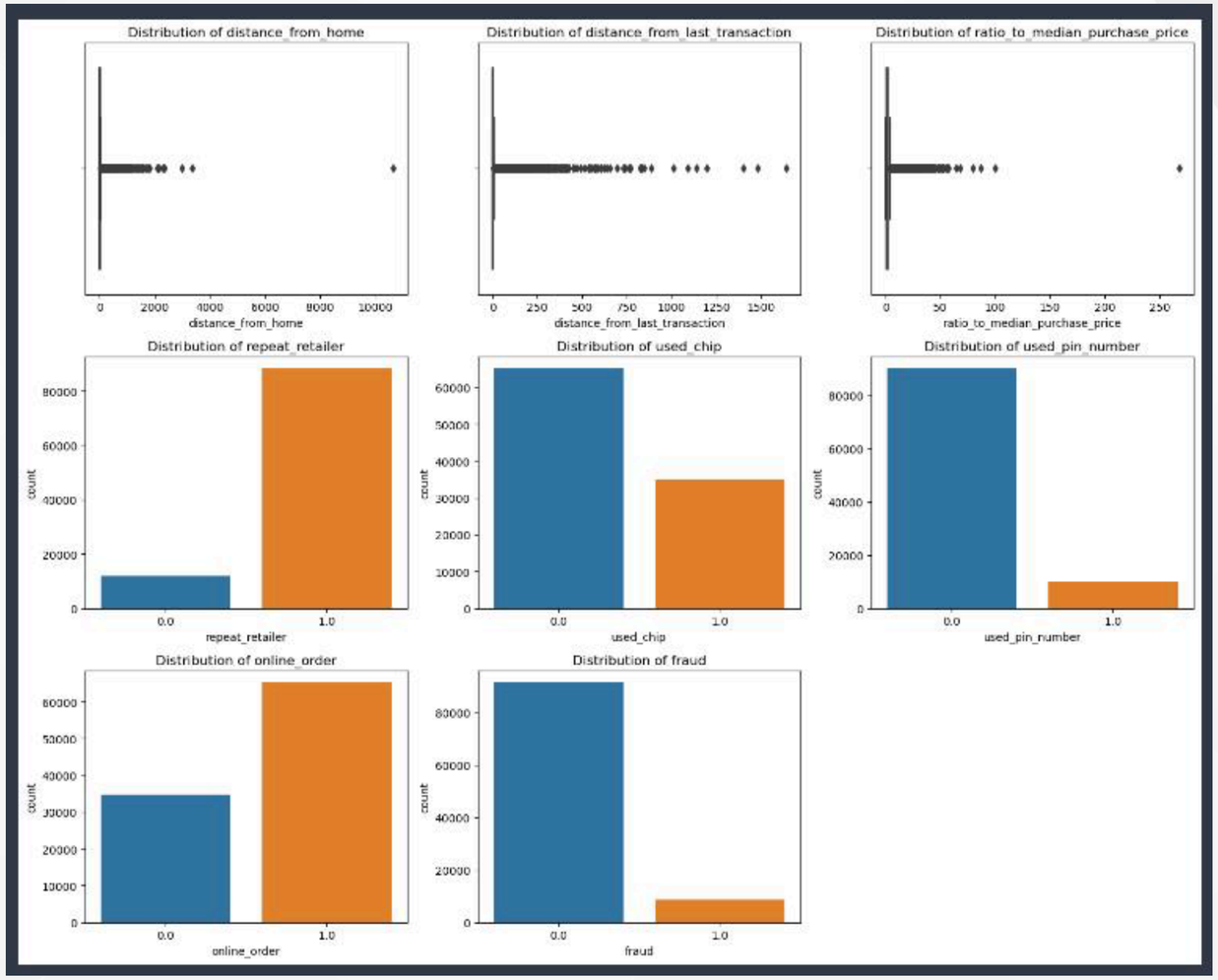
RESULT OF 2ND DATASET

Descriptive Analysis: No. of Fraud Cases



RESULT OF 2ND DATASET

Descriptive Analysis: Relation Between columns and frauds



RESULT OF 2ND DATASET

Naive Bayes:

```
Model - naive_bayes:

Test Accuracy for naive_bayes: 0.947
Train Accuracy for naive_bayes: 0.948

Confusion Matrix for naive_bayes:
 [[27866  2159]
 [ 1031 29201]]

Classification Report for naive_bayes:
              precision    recall  f1-score   support

         0.0       0.96      0.93      0.95     30025
         1.0       0.93      0.97      0.95     30232

    accuracy                           0.95     60257
   macro avg       0.95      0.95      0.95     60257
weighted avg       0.95      0.95      0.95     60257

Cross Validation scores:
 [0.94895374 0.94588851 0.94241223 0.94224874 0.95054563]
Mean Score of 0.95 with a standard deviation of 0.00
----------------------------------
```

RESULT OF 2ND DATASET

KN classifier:

```
Model - kn_classifier:

Test Accuracy for kn_classifier: 0.993
Train Accuracy for kn_classifier: 0.995

Confusion Matrix for kn_classifier:
 [[29601   424]
 [    3 30229]]

Classification Report for kn_classifier:
              precision    recall  f1-score   support

         0.0       1.00      0.99      0.99     30025
         1.0       0.99      1.00      0.99     30232

    accuracy                           0.99     60257
   macro avg       0.99      0.99      0.99     60257
weighted avg       0.99      0.99      0.99     60257

Cross Validation scores:
 [0.99305215 0.99100866 0.99288838 0.99288838 0.9912944 ]
Mean Score of 0.99 with a standard deviation of 0.00
```

RESULT OF 2ND DATASET

Random Forest:

```
Model - random_forest:

Test Accuracy for random_forest: 1.000
Train Accuracy for random_forest: 1.000

Confusion Matrix for random_forest:
 [[30023     2]
 [    0 30232]]

Classification Report for random_forest:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00     30025
         1.0       1.00      1.00      1.00     30232

    accuracy                           1.00     60257
   macro avg       1.00      1.00      1.00     60257
weighted avg       1.00      1.00      1.00     60257

Cross Validation scores:
 [1.         1.         0.99995913 1.         1.        ]
```

RESULT OF 2ND DATASET

Results of Implemented Models:

Model	Test Accuracy	Train Accuracy
Random_forest	0.999967	1.000000
Kn_classifier	0.992914	0.9
Naive_bayes	0.947060	0.9

FUTURE OUTLOOK

- Continuous Enhancement: Refine algorithms for ongoing fraud detection improvement.
- Advanced Tech Integration: Explore deep learning for sophisticated fraud pattern detection.
- Real-time Monitoring: Swift response to potential fraud through real-time detection.
- Enhanced Data Security: Implement measures for financial data integrity and confidentiality.
- User-Friendly Interfaces: Develop intuitive interfaces for efficient fraud analysis.
- Scalability Focus: Prioritize adaptability for increased transactions and changing data patterns.

CONCLUSION

- In summary, this Credit Card Fraud Detection project employed a comprehensive approach involving data preprocessing, cluster determination, and the training of Random Forest and Naive Bayes classifiers.
- Evaluation metrics revealed valuable insights into model effectiveness, while a correlation heatmap deepened our understanding of fraud patterns.
- This project presents a strong methodology for fraud detection, setting the stage for future advancements in financial security.

REFERENCES

Data Source Link: The dataset used for this analysis found at Kaggle.
Link (1):https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
Link(2):https://www.kaggle.com/datasets/dhanushnarayananr/creditcard-fraud

Research Paper URL:
Research Paper 1: Credit Card Fraud Detection in e-Commerce: An Outlier Detection Approach.
Link: https://arxiv.org/abs/1811.02196
Research paper 2: Credit Card Fraud Detection - Machine Learning methods.
Link: https://ieeexplore.ieee.org/abstract/document/8717766/references
Research paper 3: Credit Card Fraud Detection Based on Machine and Deep Learning.
Link: https://ieeexplore.ieee.org/abstract/document/9078935

THANKYOU