STUDENT NAME: TOLUWALOPE OLADIPUPO OJUROYE
STUDENT ID: @00690747
COURSE: DATA SCIENCE
MODULE TITLE: MACHINE LEARNING AND DATA MINING
LEVEL: 7

Contents

R

TITLE

Credit Card Fraud Detection Classification Modelling

(Predicting Fraud in Credit Card)

Abstract

Within the confines of the financial sector, credit card fraud is a major issue that costs both end users and financial institutions a great deal of money. To solve this trouble, the use of machine learning algorithms is being adopted more and more to recognize and stop credit card fraud. A run-through of the state of machine learning approach used for credit card fraud awareness is given in this abstract.

The creation and assessment of machines learning models, including logistic regression, and Naïve Bayes algorithms, to efficiently detect fraudulent credit card transactions has been the main emphasis of this field's study. By using past transaction data, these models may identify trends and anomalies that can be used to anticipate and identify possible fraudulent activity.

By putting strong mechanisms in place for detecting and stopping fraudulent transactions, the suggested machine learning-based credit card fraud detection systems seek to improve the security of e-payment and e-commerce systems. Using machine learning, these systems can continually learn from fresh data and adjust to changing fraud trends, which enhances their predictive powers and increases the accuracy of fraud detection.

The need for efficient fraud detection techniques is highlighted by the rising incidence of credit card fraud. To limit the effect of fraudulent actions on the financial ecosystem and safeguard consumers from financial loss, machine learning algorithms provide a possible solution to this problem. To keep ahead of increasingly complex fraudulent strategies and protect the integrity of financial transactions, continuous research and development in machine learning-based credit card fraud detection is vital as the industry continues to expand.

1.0 Introduction

The growing number of financial fraud cases has drawn a lot of interest to the application of machine learning algorithms for credit card fraud prediction. To identify and forecast fraudulent credit card transactions, machine learning models including logistic regression, random forest, decision trees, artificial neural networks, naive Bayes, and support vector machines have been tested (Afriyie et al., 2023). These algorithms are designed to enhance the detection of credit card fraud and determine which transactions are most likely to be fraudulent (Ileberi et al., 2022). On actual credit card transaction data, a variety of linear and nonlinear statistical modeling and machine learning models have been used with encouraging results; the best fraud detection performance has been exhibited by the boosted tree model (Gao et al., 2019). To choose the most reliable model for forecasting credit card default payment, the effectiveness of many machine learning algorithms has also been evaluated to analyze credit card default (Gui, 2019).

The development of a trustworthy credit card fraud detection system is the ultimate objective of this field's study, which focuses on data exploration, data cleaning, variable generation, feature selection, model algorithms, and outcomes (Gao et al., 2019). The study's challenge and intended goal determine which machine learning models are used, with an emphasis on supervised models for binary classification like logistic regression and decision trees that predict whether a transaction is fraudulent (Gui, 2019).

The goal in this research is to further the current literature by investigating the use of machine learning algorithms in credit card fraud prediction. We'll assess how well different supervised machine learning models work and how well they can identify fraudulent credit card transactions. To determine the best algorithm for credit card fraud prediction, we will also go over the procedures of feature selection, data exploration, and model comparison.

To prevent and mitigate financial fraud in e-commerce and e-payment systems, we want to create a strong predictive model that can be used to credit card fraud detection systems by using the insights and results from prior research.

1.1 Problem Statement

Credit card fraud poses an important threat to financial institutions and the end users alike, with misplacement amounting to billions of dollars annually. As technology improves, so do the techniques employed by scammers, making it imperative to develop robust and efficient fraud detection systems. The disputes lies in identifying scammers transactions promptly while minimizing false positives to ensure a seamless user experience (Mullen, 2023).

1.2 Objective

1. To develop a strong machine learning model that can distinguish between authentic and fraudulent credit card transactions by using diverse data elements and patterns.
2. To Achieve a high degree of accuracy in detecting fraudulent transactions while reducing false positives would help to prevent legitimate transactions from being mistakenly reported as fraudulent.
3. To Establish a system that can proactively identify and stop credit card fraud would help to minimize losses and safeguard the interests of companies and customers alike.
4. To examine past credit card transaction data, patterns and trends linked to fraudulent activity may be found, which can improve the model's prediction power.

1.3 Background

Numerous research works have assessed the effectiveness of various algorithms, such as decision trees, random forests, logistic regression, and deep convolution neural networks (Afriyie et al., 2023). To increase the accuracy of fraud detection, several research have suggested hybrid techniques that blend supervised and unsupervised learning (Bin Sulaiman et al., 2022). Several research have used ensemble techniques, including Xgboost and random forest, to capture prediction uncertainty (Btoush et al., 2023).

The data imbalance, whereby there are corresponding small fraudulent transactions which are proportionate to verified transactions, presents a problem in the identification of credit card fraud. In the advent of solving this barrier, several methods has been suggested by the researchers which includes oversampling and under sampling (Awoyemi et al., 2017).

Overall, research is being done to increase the accuracy and efficiency of machine learning algorithms, which are thought to be capable of detecting and predicting credit card fraud.

2.0 Data Collection and Preprocessing

This dataset has been sourced from Kaggle, a renowned platform for data science and machine learning enthusiasts. Kaggle provides a diverse collection of datasets contributed by the global data science community, fostering collaboration and innovation.

Dataset Information:

Name: credit card

Source: Kaggle

URL:https://www.kaggle.com/datasets/nelgiriyewithana/credit-cardfraud-detection-dataset-2023

In [ ]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.discrete.discrete_model as sm
```

In [ ]:
```
cc = pd.read_csv("creditcard.csv")
cc.head()
```

Out[38]:

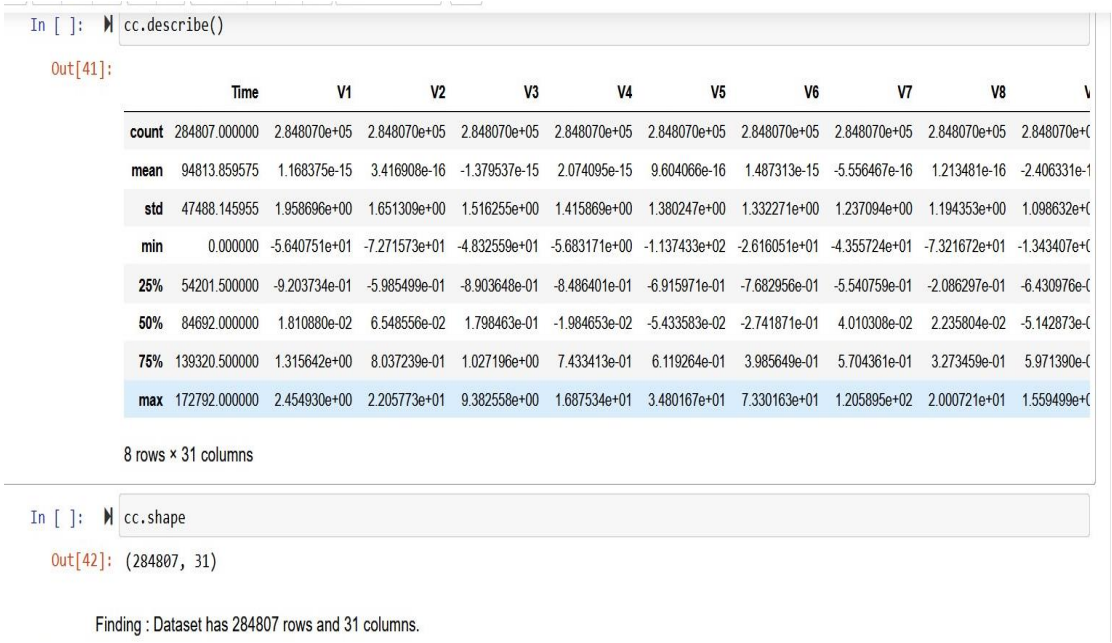
	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.1
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.1
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	-0.3
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	0.6
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	-0.2

5 rows × 31 columns

In [ ]:
```
from google.colab import drive
drive.mount('/content/drive')
```

2.1 Description

Data Source: The credit card transactions done by European cardholders in 2023 are included in this dataset. It contains more than 550,000 records in total, and to safeguard the identity of the cardholders, the data has been anonymized. This dataset's main goal is to make it easier to create models and algorithms for fraud detection that may be used to spot possibly fraudulent transactions.



```
In [ ]: cc.describe()
```

Out[41]:

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V
count	284807.000000	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+(
mean	94813.859575	1.168375e-15	3.416908e-16	-1.379537e-15	2.074095e-15	9.604066e-16	1.487313e-15	-5.556467e-16	1.213481e-16	-2.406331e-1
std	47488.145955	1.958696e+00	1.651309e+00	1.516255e+00	1.415869e+00	1.380247e+00	1.332271e+00	1.237094e+00	1.194353e+00	1.098632e+(
min	0.000000	-5.640751e+01	-7.271573e+01	-4.832559e+01	-5.683171e+00	-1.137433e+02	-2.616051e+01	-4.355724e+01	-7.321672e+01	-1.343407e+(
25%	54201.500000	-9.203734e-01	-5.985499e-01	-8.903648e-01	-8.486401e-01	-6.915971e-01	-7.682956e-01	-5.540759e-01	-2.086297e-01	-6.430976e-(
50%	84692.000000	1.810880e-02	6.548556e-02	1.798463e-01	-1.984653e-02	-5.433583e-02	-2.741871e-01	4.010308e-02	2.235804e-02	-5.142873e-(
75%	139320.500000	1.315642e+00	8.037239e-01	1.027196e+00	7.433413e-01	6.119264e-01	3.985649e-01	5.704361e-01	3.273459e-01	5.971390e-(
max	172792.000000	2.454930e+00	2.205773e+01	9.382558e+00	1.687534e+01	3.480167e+01	7.330163e+01	1.205895e+02	2.000721e+01	1.559499e+(

8 rows × 31 columns

```
In [ ]: cc.shape
```

Out[42]: (284807, 31)

Finding : Dataset has 284807 rows and 31 columns.

2.2 Data Cleaning

Data cleaning is the process of managing soiled data while maintaining a high level of detection accuracy. It includes managing missing data, scalability when necessary, and eliminating empty values.

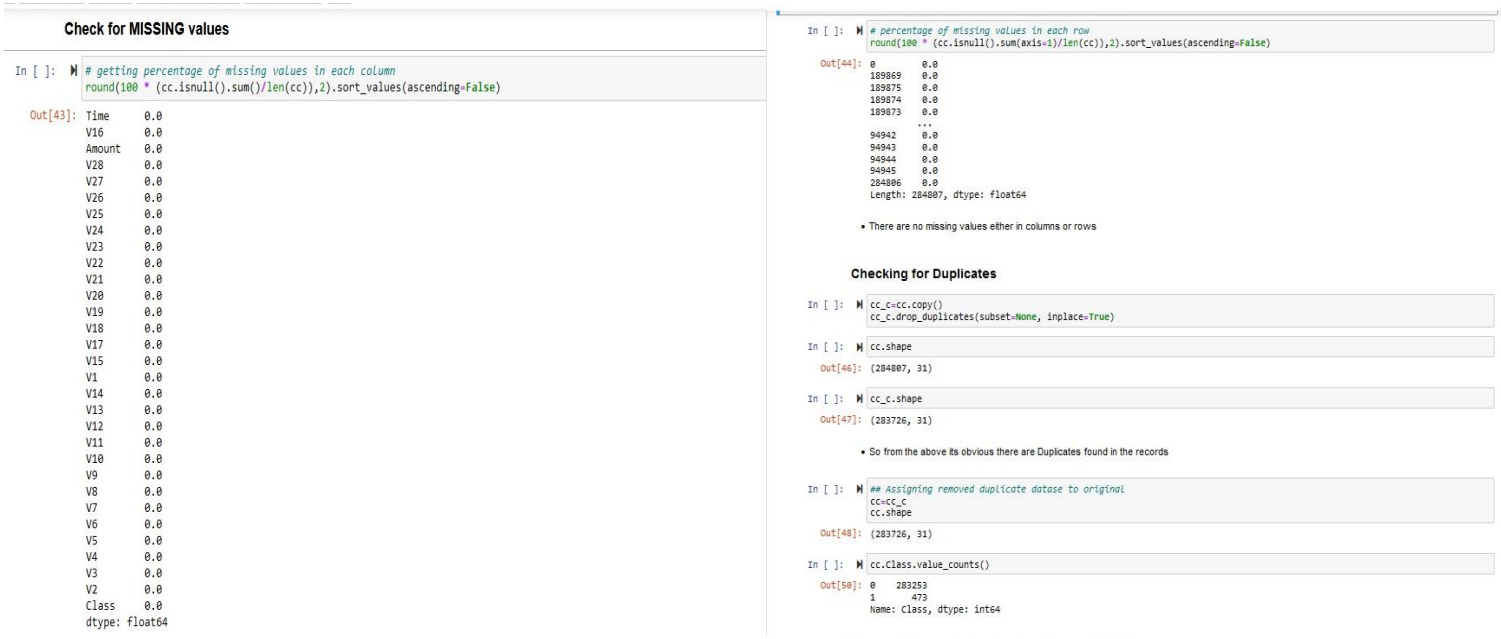
Dataset Balancing: To guarantee accurate forecasts for both fraudulent and non-fraudulent transactions, the dataset must be balanced. This stage aids in resolving the imbalanced data problem and enhances the predictive models' functionality.

Feature engineering: This stage aims to increase the model's predictive capacity by adding new features or altering current ones. To improve prediction abilities, it can include more factors in the dataset.

Resampling Techniques: To solve the problem of unbalanced datasets, resampling techniques like under sampling and oversampling are often used. To train reliable prediction models, these strategies aid in the creation of a balanced distribution of fraudulent and non-fraudulent transactions.

Algorithm Comparison and Selection: To identify the best accurate algorithm for identifying fraudulent transactions, a variety of machine learning techniques,

including decision trees, logistic regression, random forests, and ensemble tree algorithms, are assessed and contrasted.



Check for MISSING values

```
In [ ]: # getting percentage of missing values in each column
        round(100 * (cc.isnull().sum()/len(cc)),2).sort_values(ascending=False)

Out[43]: Time      0.0
         V16       0.0
         Amount    0.0
         V28       0.0
         V27       0.0
         V26       0.0
         V25       0.0
         V24       0.0
         V23       0.0
         V22       0.0
         V21       0.0
         V20       0.0
         V19       0.0
         V18       0.0
         V17       0.0
         V15       0.0
         V1        0.0
         V14       0.0
         V13       0.0
         V12       0.0
         V11       0.0
         V10       0.0
         V9        0.0
         V8        0.0
         V7        0.0
         V6        0.0
         V5        0.0
         V4        0.0
         V3        0.0
         V2        0.0
         Class     0.0
         dtype: float64
```

```
In [ ]: # percentage of missing values in each row
        round(100 * (cc.isnull().sum(axis=1)/len(cc)),2).sort_values(ascending=False)

Out[44]: 0         0.0
         189869    0.0
         189875    0.0
         189874    0.0
         189873    0.0
                  ...
         94942     0.0
         94943     0.0
         94944     0.0
         94945     0.0
         284806    0.0
         Length: 284807, dtype: float64
```

- There are no missing values either in columns or rows

Checking for Duplicates

```
In [ ]: cc_c=cc.copy()
        cc_c.drop_duplicates(subset=None, inplace=True)

In [ ]: cc.shape

Out[46]: (284807, 31)

In [ ]: cc_c.shape

Out[47]: (283726, 31)
```

- So from the above its obvious there are Duplicates found in the records

```
In [ ]: ## Assigning removed duplicate datase to original
        cc=cc_c
        cc.shape

Out[48]: (283726, 31)

In [ ]: cc.Class.value_counts()

Out[50]: 0    283253
         1       473
         Name: Class, dtype: int64
```

2.3 Exploratory Data Analysis (EDA)

Data scientists use exploratory data analysis (EDA) to examine and evaluate data sets and enumerate their primary attributes, often using techniques for data visualization. Data scientists may find trends, identify anomalies, test hypotheses, and verify assumptions more easily when they know how to effectively manipulate data sources to acquire the answers they need.

Some of the exploratory activities that were carried out are shown below:

```
In [ ]: cc.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 283726 entries, 0 to 284806
Data columns (total 31 columns):
 #   Column  Non-Null Count   Dtype  
---  ------  --------------   -----  
 0   Time    283726 non-null  float64
 1   V1      283726 non-null  float64
 2   V2      283726 non-null  float64
 3   V3      283726 non-null  float64
 4   V4      283726 non-null  float64
 5   V5      283726 non-null  float64
 6   V6      283726 non-null  float64
 7   V7      283726 non-null  float64
 8   V8      283726 non-null  float64
 9   V9      283726 non-null  float64
 10  V10     283726 non-null  float64
 11  V11     283726 non-null  float64
 12  V12     283726 non-null  float64
 13  V13     283726 non-null  float64
 14  V14     283726 non-null  float64
 15  V15     283726 non-null  float64
 16  V16     283726 non-null  float64
 17  V17     283726 non-null  float64
 18  V18     283726 non-null  float64
 19  V19     283726 non-null  float64
 20  V20     283726 non-null  float64
 21  V21     283726 non-null  float64
 22  V22     283726 non-null  float64
 23  V23     283726 non-null  float64
 24  V24     283726 non-null  float64
 25  V25     283726 non-null  float64
 26  V26     283726 non-null  float64
 27  V27     283726 non-null  float64
 28  V28     283726 non-null  float64
 29  Amount  283726 non-null  float64
 30  Class   283726 non-null  int64  
dtypes: float64(30), int64(1)
memory usage: 69.3 MB
```

```
In [ ]: def draw_histograms(dataframe, features, rows, cols):
            fig=plt.figure(figsize=(20,20))
            for i, feature in enumerate(features):
                ax=fig.add_subplot(rows,cols,i+1)
                dataframe[feature].hist(bins=20,ax=ax,facecolor='midnightblue')
                ax.set_title(feature+" Distribution",color='DarkRed')
                ax.set_yscale('log')
            fig.tight_layout()
            plt.show()

        draw_histograms(cc,cc.columns,8,4)
```

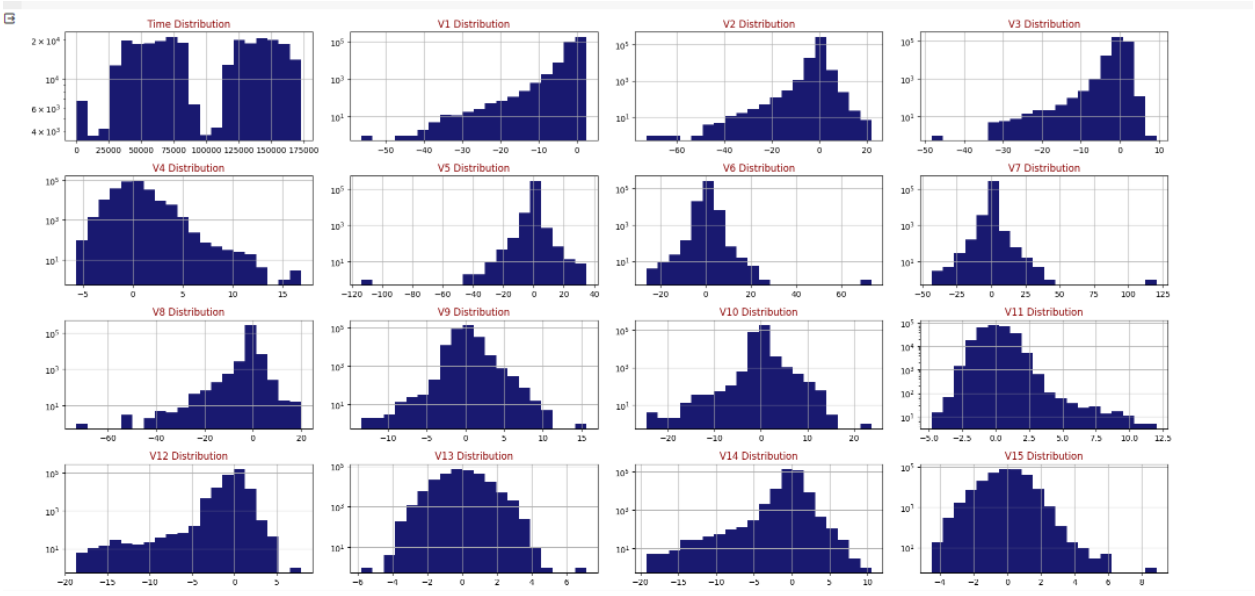


Figure 1 Distribution of Dataset

From the dataset EDA for figure 2.1 it could be seen that most of the distribution in dataset ranging from V1 − V23 from the chat it could be seen some of the datasets are normally distributed, Left Skewed and some are right skewed.

```
In [ ]: cc.Class.value_counts()

Out[54]: 0    283253
         1       473
         Name: Class, dtype: int64

In [ ]: ax = sns.countplot(x='Class',data=cc);
        ax.set_yscale('log')
```

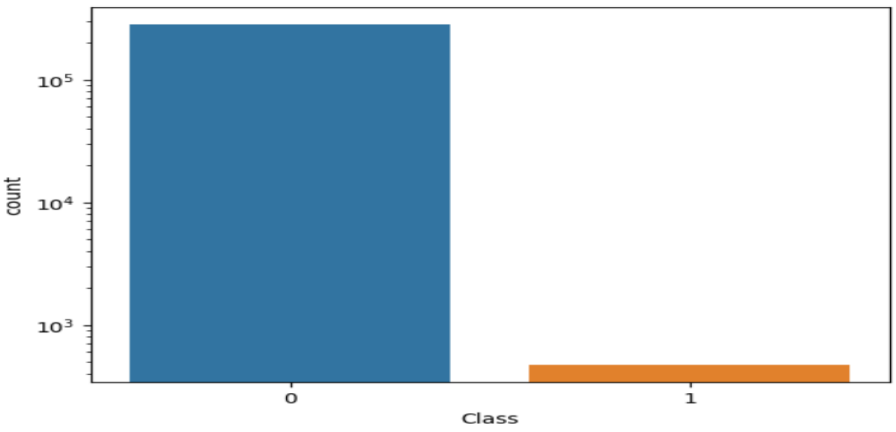


Figure 2 Class Distribution of Imbalance Dataset

Figure 2.2 Class Distribution of Imbalance dataset.

From figure 2.2 it could be identified that there is a case of data imbalance and research has shown that such data imbalance can lead to machine learning algorithms being bias in its predictions. So, to avoid such bias in machine learning algorithm under sampling was carried out to ensure class balancing in its distribution as shown in figure 2.3.

```
[ ]  sns.countplot(x='Class',data=fcc)

     <Axes: xlabel='Class', ylabel='count'>
```

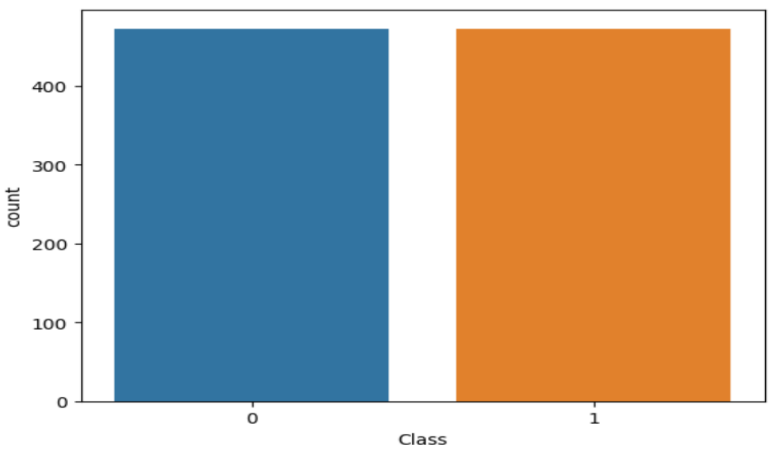


Figure 3 Class Distribution after Down sampling

3.0 Methodology

During cause of model development two algorithms were use and the two algorithms are Logistic regression (Statistical in nature) and Naïve Bayes (Probabilistic in nature)

4.0 Results

Different evaluation techniques were used to test the performance of the model as well as feature engineering techniques were applied to ensure values with P values greater > 5%. The findings indicate that several features have P values that are greater than the recommended alpha of 5%, indicating a weak statistically meaningful correlation with the likelihood of fraud. Also, Confusion Matrix, Accuracy, and ROC curve. Figure 4.1 shows the accuracy score of Logistic regression (LR) with accuracy of 0.94 which shows that logistic regression has a good score, but accuracy score may not be enough evaluation technique for classification problem.

4.1 Logistic regression

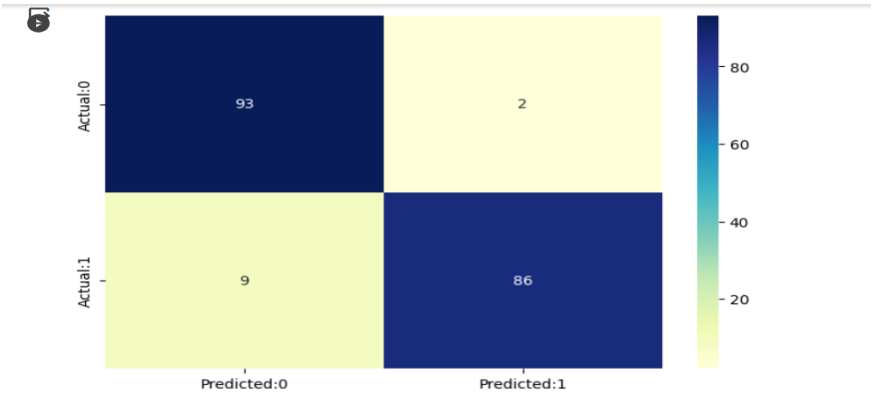
```
[ ] from sklearn.metrics import accuracy_score

    print(accuracy_score(y_test,y_pred))

0.9421052631578948
```

Figure 4 LR Accuracy Score

Figure 4.2 shows the confusion matrix of the Logistic regression. From analysis we can see the machine learning algorithm could predict correctly 179 correct predictions and 11 incorrect predictions.

Confusion matrix

```
In [ ]: from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test,y_pred)
conf_matrix = pd.DataFrame(data=cm,columns=['Predicted:0','Predicted:1'],index=['Actual:0','Actual:1'])
plt.figure(figsize = (8,5))
sns.heatmap(conf_matrix, annot=True,fmt='d',cmap="YlGnBu");
```



The confusion matrix shows 93+86 = 179 correct predictions and 2+9= 100 incorrect ones.

True Positives: 86

True Negatives: 93

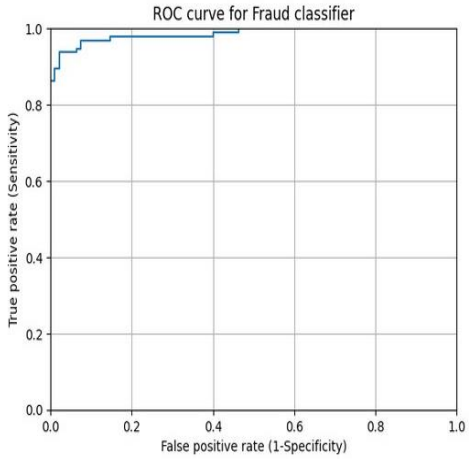
False Positives: 9 (Type I error)

False Negatives: 2 (Type II error)

Figure 5 LR Confusion Matrix

Figure 4.3 show the Performance of logistic regression (LR) using ROC Curve and from the indication the model started at a point of 8.4 and got better at point of 95% +

```
In [ ]: from sklearn.metrics import roc_curve
        fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob_yes[:,1])
        plt.plot(fpr,tpr)
        plt.xlim([0.0, 1.0])
        plt.ylim([0.0, 1.0])
        plt.title('ROC curve for Fraud classifier')
        plt.xlabel('False positive rate (1-Specificity)')
        plt.ylabel('True positive rate (Sensitivity)')
        plt.grid(True)
```




```
In [ ]: from sklearn.metrics import roc_auc_score

        roc_auc_score(y_test,y_pred_prob_yes[:,1])

Out[122]: 0.9859279778393352
```

Figure 6 LR ROC Curve

4.2 Naïve Bayes (NB)

From Figure 4.4 it could be seen that accuracy score of Naïve bayes is 0.90 which is less than the score of logistic regression.

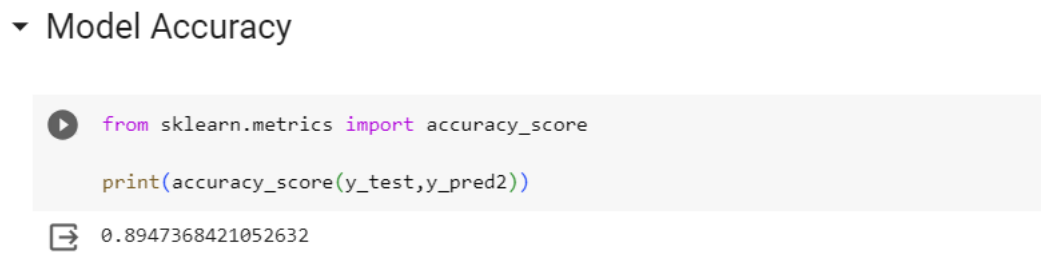


Figure 7 NB Accuracy Score

Figure 4.5 shows the confusion matrix of the Naïve Bayes and from analysis we can see the machine learning algorithm could predict 170 correct predictions and 20 incorrect predictions.

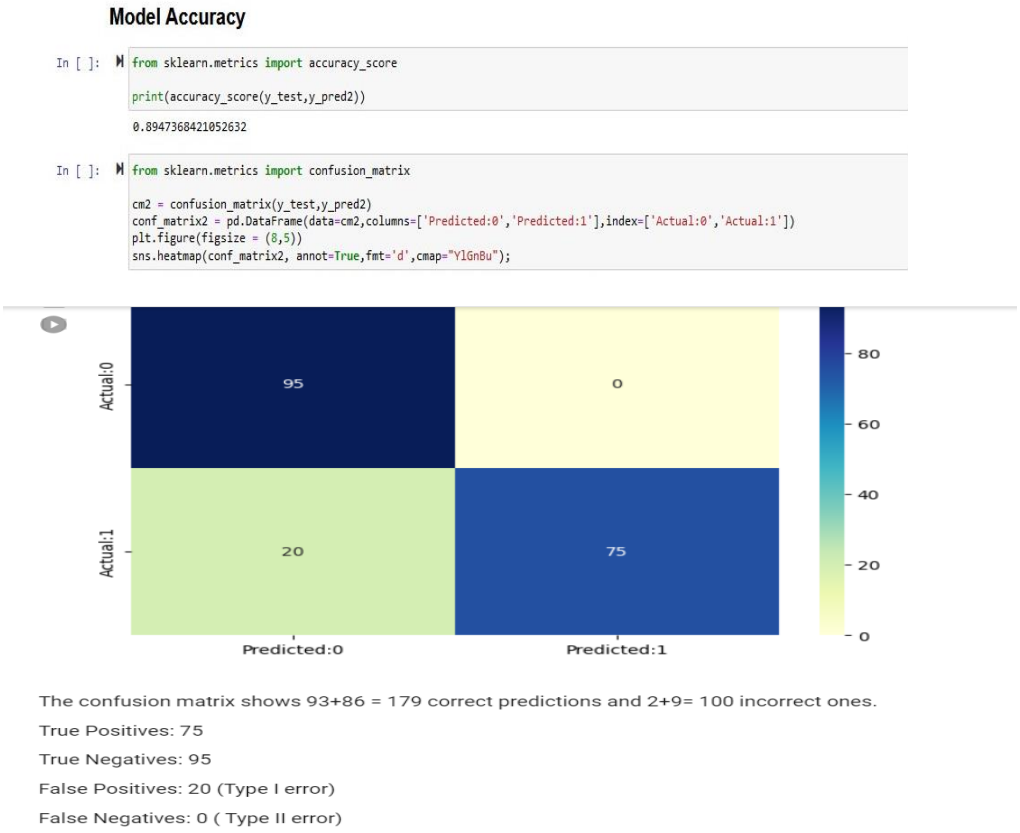
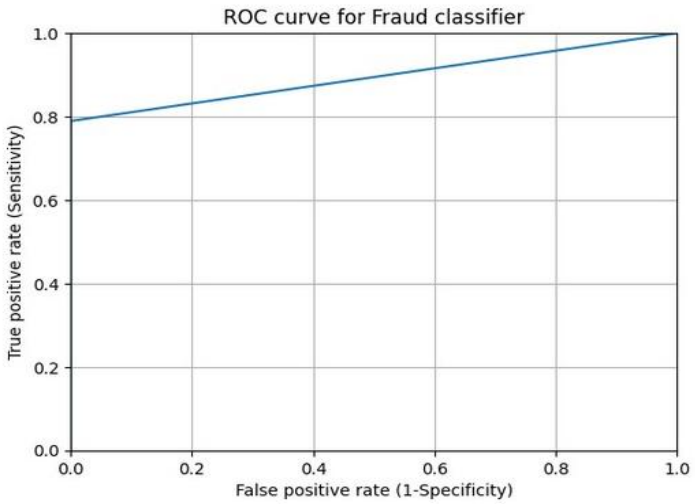


Figure 8 NB Confusion Matrix

Figure 4.6 show the Performance of Naïve Bayes (NB) using ROC Curve and from the indication the model perform poorly when compared to LR.

```
In [ ]: from sklearn.metrics import roc_curve
        fpr2, tpr2, thresholds2 = roc_curve(y_test, y_pred2)
        plt.plot(fpr2,tpr2)
        plt.xlim([0.0, 1.0])
        plt.ylim([0.0, 1.0])
        plt.title('ROC curve for Fraud classifier')
        plt.xlabel('False positive rate (1-Specificity)')
        plt.ylabel('True positive rate (Sensitivity)')
        plt.grid(True)
```




```
In [ ]: roc_auc_score(y_test,y_pred2)
Out[123]: 0.8947368421052632
```

Figure 9 NB ROC Curve

CONCLUSION

In summary, different evaluation techniques were used to test the performance of the model but the findings indicate that several features have P values that are greater than the recommended alpha of 5%, indicating a weak statistically meaningful correlation with the likelihood of fraud.

Also, the confusion matrix, Accuracy, and ROC curve shows the accuracy score of Logistic regression (LR) with accuracy of 0.94 which indicates a good score but accuracy score may not be enough to evaluate techniques for classification problem, however the result performance of Naïve Bayes (NB) using ROC Curve and from all indication the model perform poorly when compared to linear regression(LR)

PART 2

TITLE

CLUSTERING CUSTOMER DATA FROM A TRAVELING AGENTS

1.0 Introduction

As a means of providing businesses with insights into consumer behavior, preferences, and booking criteria, clustering client data from travel agents is an essential duty in the travel industry. In this procedure, clients' flight searches are grouped together based on shared attributes, such travel habits, booking preferences, and other pertinent factors. Travel companies may more efficiently evaluate and comprehend the decisions made by their online consumers by using cluster computing and feedback clustering methods. This may result in more individualized client experiences, better marketing campaigns, and increased enterprise performance.

Example research named "Feedback Clustering for Online Travel Agencies Searches" (Scaramuccia et al., 2020) emphasizes the need of grouping customer data from travel agencies. As the need to comprehend consumer choices in the travel sector grows, this research highlights the need to organize flight searches according to common booking criteria. Furthermore, cluster computing offers a generic approach to creating parallel high-performance systems, as covered in "Cluster Computing - an overview" (Chen et al., 2018). This approach may be applicable to the travel industry's analysis of consumer data.

Travel companies may discover patterns in booking behavior, get insightful knowledge about client segments, and customize their marketing and service offers to specifically cater to the demands of various consumer groups by clustering customer data efficiently. More specialized marketing efforts, tailored trip suggestions, and higher levels of consumer satisfaction are all possible outcomes of this strategy.

Ultimately, the travel industry relies heavily on the clustering of consumer data from travel agents, which helps businesses better understand and meet the varied needs and preferences of their clientele. Travel agencies may improve their client experiences, sharpen their marketing tactics, and eventually propel company development in a

market that is becoming more and more competitive by using feedback clustering methods and cluster computing.

In summary, this introduction highlights the value of grouping customer data from travel agents and highlights how it can be used to better understand consumer behavior and develop business plans for the travel sector.

Problem Synopsis

To comprehend client behavior, preferences, and booking criteria, one of the most important tasks in the travel sector is clustering customer data from travel agents. The main difficulty is in efficiently classifying flight searches made by consumers that have similar attributes, such booking inclinations and travel habits, to identify client categories and adjust marketing tactics and offerings appropriately.

The increasing significance of comprehending client selections in the travel sector highlights the need of clustering customer data from travel agents(Scaramuccia et al., 2020). The problem lies in consolidating all flight searches conducted by consumers who have the same booking criteria, as flight search queries become the major source of customer information. To efficiently segment consumer data and extract useful insights, this calls for the creation of novel clustering algorithms and techniques.

Moreover, the use of clustering techniques to identify business travelers (Tabianan et al., 2022) emphasizes the necessity to identify underlying patterns in the data to differentiate between various client categories, including business travelers, and customize services to meet their unique needs. This emphasizes how crucial it is to have strong clustering methods to precisely recognize and target different client categories in the travel sector.

Furthermore, the application of clustering algorithms for intelligent customer segmentation (Tabianan et al., 2022) highlights the need of utilizing cutting-edge methods, like K-Means clustering, to evaluate customer purchase behavior data and improve sales by offering personalized experiences for each segment. This highlights even more how important it is to have efficient clustering techniques to enhance business results and extract useful insights from consumer data.

To summarize, the challenge of grouping customer data from travel agents is centered on the creation of novel algorithms and techniques that can efficiently classify flight searches, recognize discrete customer categories, and customize marketing tactics and offerings to cater to the requirements of various travel industry clientele. Travel agencies must overcome this obstacle if they want to increase client satisfaction, get a competitive advantage, and expand their company.

Background

To comprehend client preferences and customize marketing tactics, clustering consumer data from travel agencies entails putting clients into groups according to their search habits, booking requirements, and purchase behavior. This procedure aids in creating itineraries, customizing travel experiences, and focusing on certain clientele. Customer data is analyzed using a variety of clustering methods, including K-Means and K-Modes, to find trends and commonalities.

Segmentation is essential in the travel and tourism sector for creating itineraries and marketing materials that are directed towards certain clientele (Dolnicar, 2022).

The act of grouping all the data into groups (clusters) according to the patterns seen in the data is known as clustering (Tabianan et al., 2022). For instance, it was noted as critical in a case study on online travel firms to combine all flight searches made by the same kind of consumer using the same booking parameters.

A K-Means clustering technique was presented in research on customer purchase behavior data for intelligent customer segmentation to comprehend consumer behavior and concentrate on highly lucrative categories (Tabianan et al., 2022). Like this, clustering algorithms were used to identify business travelers by grouping passengers based on airline data. This showed the possibility of finding new market groups and assigning passengers to those categories for focused marketing.

To improve marketing tactics and provide individualized travel experiences, clustering consumer data from travel agents entails using clustering algorithms to get insights into customer behavior, preferences, and characteristics.

Overview of the process: Using sophisticated data analysis tools, clients are grouped according to shared travel habits, preferences, and behavioral patterns. A wide range

of information is gathered by travel agencies, such as past reservations, preferred destinations, frequency of travel, budget preferences, and customer reviews. After that, clustering techniques are used to this heterogeneous dataset to find organic patterns and data segments.

Vital Steps:

Data Gathering**:**

collecting thorough consumer information from a range of sources, including websites, travel agencies, and customer reviews.

gathering data on preferences for travel, ideal locations, financial concerns, frequency of travel, and other pertinent aspects.

Preprocessing of Data:

To guarantee correctness and consistency, the data should be cleaned and arranged.

managing partial or missing data pieces and normalizing variables to ensure consistency in analysis.

Feature Choice:

determining the salient characteristics that support client segmentation.

Considering variables such frequency of travel, favorite class of travel, choice of lodging, and financial limitations.

Use of Clustering Algorithm:

grouping clients based on shared traits by using clustering techniques such as DBSCAN, k-means, or hierarchical clustering. Clusters are iteratively refined to produce segments that are relevant and useful.

Segmentation Analysis:

constructing thorough profiles that showcase shared traits and inclinations for every consumer group. Recognizing the requirements and demands of every group to customize service offerings and marketing tactics.

Advantages:

Tailored Promotion:

using data on market segments to develop marketing efforts that are relevant to each target audience.

The optimization of operations:

simplifying processes considering the discovered clusters, and allocating resources and inventories as efficiently as possible.

Improved Client Experience

providing individualized services, incentives, and travel advice depending on the tastes of each group.

Making Strategic Decisions:

enabling travel businesses to decide on collaborations, product development, and growth plans with knowledge.

2.0 Data Collection and Preprocessing

The process of classifying and arranging massive amounts of data gathered from different travel agencies is known as clustering client data from travel agents. Through the identification of patterns, trends, and unique groupings within the consumer data, this analytical technique eventually assists travel companies in making well-informed business choices. This is a thorough explanation of the procedure:

2.1 Description

To comprehend client preferences and customize marketing tactics, clustering consumer data from travel agencies entails putting clients into groups according to their search habits, booking requirements, and purchase behavior.

An effective technique for consumer segmentation is cluster analysis, which finds patterns and similarities in customer data using a variety of methods, including K-

Modes and K-Means. Segmentation is essential in the travel and tourism sector for creating itineraries and marketing materials that are aimed at certain clientele.

The potential for finding new market segments and grouping passengers into those categories for targeted marketing has been shown using clustering algorithms to cluster passengers based on data acquired from airlines. In a similar vein, an intelligent customer segmentation strategy based on K-Means clustering was put out to comprehend consumer behavior and concentrate on highly lucrative areas. To improve marketing tactics and provide individualized travel experiences, clustering consumer data from travel agents entails using clustering algorithms to get insights into customer behavior, preferences, and characteristics.

2.2 Data Cleaning

Data cleaning is the process of managing soiled data while maintaining a high level of detection accuracy. It includes managing missing data, scalability when necessary, and eliminating empty values.

Dataset Balancing: To guarantee accurate forecasts for both fraudulent and non-fraudulent transactions, the dataset must be balanced. This stage aids in resolving the imbalanced data problem and enhances the predictive models' functionality.

Feature engineering: This stage aims to increase the model's predictive capacity by adding new features or altering current ones. To improve prediction abilities, it can include more factors in the dataset.

Resampling Techniques: To solve the problem of unbalanced datasets, resampling techniques like under sampling and oversampling are often used. To train reliable prediction models, these strategies aid in the creation of a balanced distribution of fraudulent and non-fraudulent transactions.

Algorithm Comparison and Selection: To single out the best precise algorithm for recognizing scammers transactions, a variety of machine learning approach, including decision trees, logistic regression, random forests, and ensemble tree algorithms, are evaluated and contrasted.

2.3 Exploratory Data Analysis (EDA)

Data scientists use exploratory data analysis (EDA) to investigate and assess data sets and itemize their primary impute, often making use of techniques for data visualization. Data scientists may find trends, identify anomalies, test hypotheses, and verify assumptions more easily when they know how to effectively manipulate data sources to acquire the answers they need.

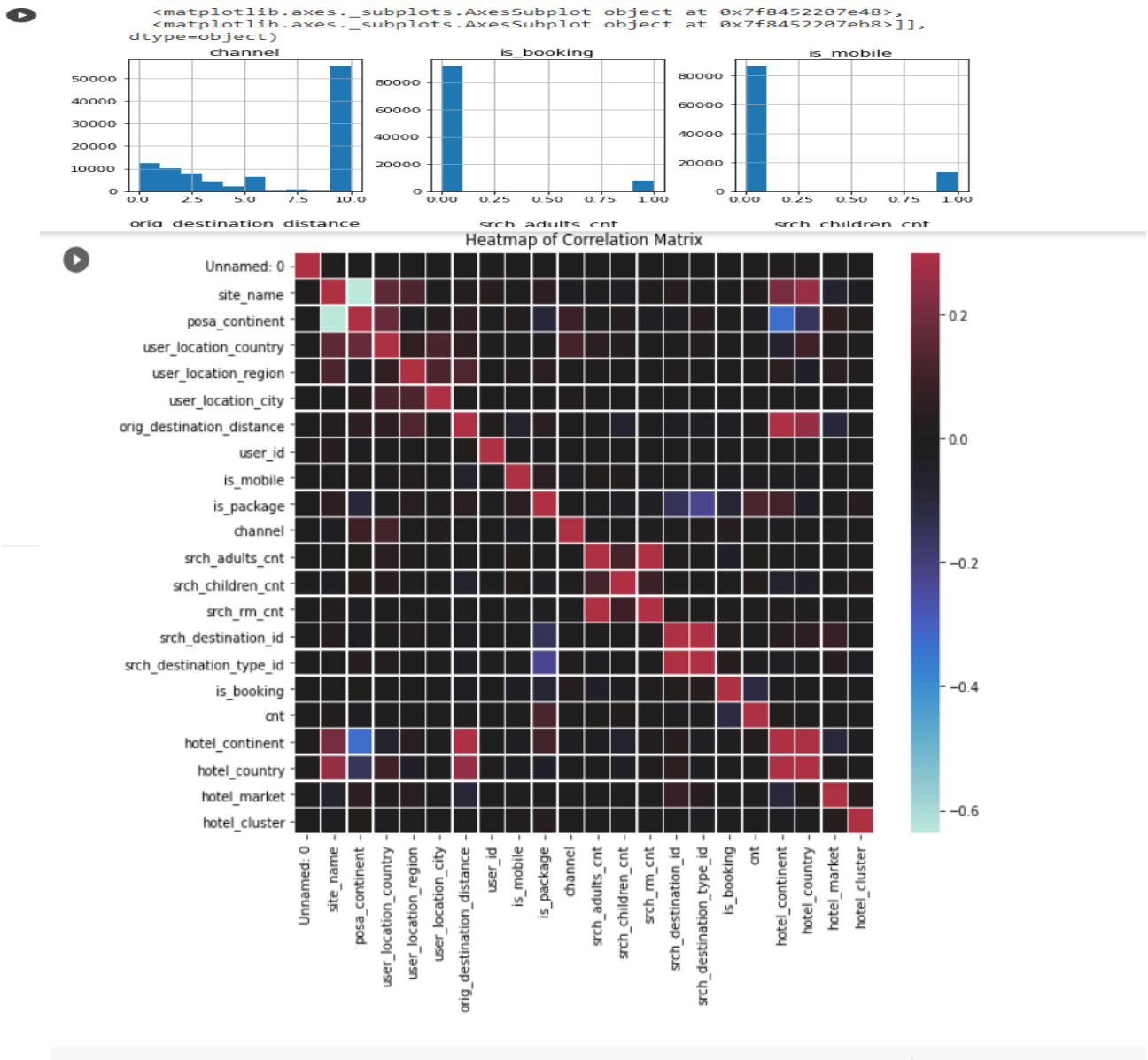
Some of the exploratory activities that were carried out are shown below: Figure 10 presents a heatmap that shows the degree of correlation between variables. red-colored regions indicate a strong positive association with other variables. The heatmap's darker areas indicate higher degrees of positive correlation, highlighting how closely related the variables under analysis are to one another.

```
In [ ]: # Generate our corelation plot or heatmap
        plt.figure(figsize = (10,10))
        cmap = sns.diverging_palette(220,10,as_cmap = True)

        sns.heatmap(corr,xticklabels=corr.columns.values,
                    yticklabels=corr.columns.values,cmap=cmap,vmax=.3, center=0, square=True, linewidths=.5, cbar_kws={"shrink": .82
        plt.title('Heatmap of Correlation Matrix')

Out[8]: Text(0.5, 1.0, 'Heatmap of Correlation Matrix')
```

Figure 10 Heat Map Correlation of Dataset Involved

Also, to find more patterns and information in the data, new features were designed.

To enhance the performance and comprehension of the model, polynomial features, interaction terms, and domain-specific characteristics were developed.



Following the feature engineering challenge, datasets of sizes 2k, 3k, and 4k were subjected to K-means clustering. But the investigation showed that utilizing datasets

of sizes 2k and 3k produced the best clustering results. Figure 12 provides further information and examples of this phenomenon.

Figure 11 Distribution of Dataset

```
plt.show()
```

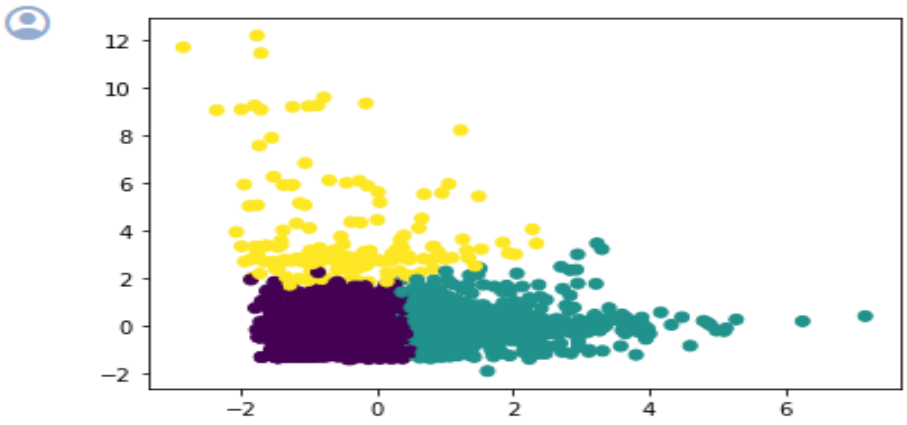


Figure 12 Clustering segmentation

PART 3

TITLE

Sentiment Analysis of Airline Tweets

https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

1.0 Introduction

Twitter US Airline Sentiment Analysis is a project that involves analyzing tweets from Twitter regarding major US airlines to classify them as positive, negative, or neutral. The dataset is coined out from Kaggle and the project was created in February, 2015 (Hosseini, 2020). The purpose of this project is to furnish airlines with free feedback that can help them upgrade their services and end users' satisfaction. Sentiment analysis is performed on the tweets using various techniques such as text processing, vectorization, and parameter tuning. The project involves both descriptive

and predictive analysis to understand the sentiments expressed by travelers on Twitter (Hosseini, 2020).

One of the most important techniques in social media marketing is sentiment analysis and categorization of tweets, which entails keeping an eye on the feelings expressed in Twitter discussions.

Companies may get insights into their audience, stay updated about mentions of their brand and rivals, and identify emerging trends in the market by monitoring consumer feedback on Twitter.

As part of a project called Twitter US Airline Sentiment Analysis, tweets about significant US airlines are analyzed and categorized as positive, negative, or neutral. The dataset is available for download via Kaggle, and the project was completed in February 2015. The project's goal is to provide airlines free input so they may raise consumer happiness and enhance services. The tweets are subjected to sentiment analysis using a variety of methods, including text processing, vectorization, and parameter tuning. Descriptive and predictive analysis are used in this study to comprehend the opinions that passengers have shared on Twitter4.

As social media has become a fundamental part of our everyday lives, one of the most widely used social media tenets is Twitter. Tweets are short messages sent by prospects on the microblogging place Twitter. Sentiment analysis is a crucial tool in social media marketing, and Twitter has appeared as a useful information resource for companies. Sentiment analysis is the process of classifying tweets according to the emotions that users of Twitter are expressing. By keeping an eye on customer comments on Twitter (Hosseini, 2020), businesses can get insights into their target market, remain informed about mentions of their brand and competitors, and spot new trends in the industry.

One great example of how sentiment analysis can be utilized to provide companies with insightful information is the Twitter US Airline Sentiment Analysis project. The dataset comprises the airline name, the text of tweet5, the mood of the tweets (good, negative, and neutral), and the reasons for the bad tweets. For sentiment analysis, several methods have been used, such as TFIDF with Naive Bayes, Word2Vec with Naive Bayes, Bag of Words with Naive Bayes, and TFIDF weighted Word2Vec with Naive Bayes (Hosseini, 2020).

Both descriptive and predictive analysis are included in the project.
Using a descriptive analysis, one may comprehend the opinions shared by passengers on Twitter by condensing and illustrating the data. Using text-based predictions to infer a tweet's emotion is known as predictive analysis. To create prediction models, the project has used a variety of machine learning methods, including Naive Bayes, Logistic Regression, Random Forest, and Support Vector Machines.

1.2 Background on Airline Tweet Sentiment Analysis

Using natural language processing (NLP) methods, sentiment analysis of airline tweets determines the sentiment conveyed in these tweets. Determining if the tweets express neutral, positive, or negative attitudes can help this study determine consumer happiness (Das et al., 2017).

Several research works, including case studies and systematic reviews, have examined sentiment analysis of airline tweets. The aim of these research is to assist airlines in understanding consumer feedback and improving their offerings appropriately (Zahoor and Rohilla, 2020).

To categorize airline-related tweets into positive, negative, or neutral categories, sentiment analysis of airline tweets entails gathering relevant tweets, preprocessing the data, and using machine learning techniques. Different methods are used to filter away irrelevant text 5 because of the possible noise in the data.

An example of how machine learning algorithms work in this situation is shown by research that employed an ensemble CNN and LSTM architecture for sentiment analysis of airline tweets (Das et al., 2017).

Further research used the Naive Bayes algorithm to categorize the emotions of recent tweets from various airlines. Sentiment analysis of airline tweets has also been investigated to provide airlines the ability to evaluate the positivity or negativity of tweets according to the day of the week.

All things considered, sentiment analysis of airline tweets is crucial to the airline industry's ability to learn from consumer input and boost customer happiness via improved services.

1.2 Objective

To categorize tweets about the main US airlines into good, negative, and neutral categories, a sentiment analysis was conducted on Twitter.

1.3 Data Synopsis: •

An examination of attitudes about the issues facing every major American airline. Contributors were requested to initially categorize good, negative, and neutral tweets; they were then asked to identify negative causes (such "late flight" or "rude service") after Twitter data from February 2015 was scraped.

Dataset:

The dataset has the following columns:

tweet_id

airline sentiment

airline_sentiment_confidence

Negative reason

Negative reason confidence

airline

airline_sentiment_gold

name

Negative reason_gold

retweet_count

text

tweet_coord

tweet_created.

tweet_location

user_timezone

2.0 Data Collection and Preprocessing

This dataset has been sourced from Kaggle, a renowned platform for data science and machine learning enthusiasts. Kaggle provides a diverse collection of datasets contributed by the global data science community, fostering collaboration and innovation.

Dataset Information:

Name: credit card

Source: Kaggle

URL: *https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment*

```
text = " ".join(review for review in airline_tweets.text)
print ("There are {} words in the combination of all review.".format(len(text)))
```

```
There are 1534594 words in the combination of all review.
```

Figure 13 Total dataset count

From the figure 10 statistical analysis was done to get the number word that is involved in the entire dataset, and we could see from the analysis that the total word in combination is 1534594 words. Also, in figure 11-word cloud chat was used to show what most of the tweets were about.

```
[ ]  # Generate a word cloud image
     wordcloud = WordCloud(background_color="white").generate(text)

     # Display our Word Cloud
     plt.figure(figsize=(12,8))
     plt.imshow(wordcloud, interpolation='bilinear')
     plt.axis("off")
     plt.show()
```

Figure 14 Word cloud

So, from the information on the word cloud, we can see that most of the users are tweeting more of the letters that are in bold like AmericanAir, Flight, JetBlue, United, USAirway and SouthwesAir, these are the major trending topics in the dataset. Also, the word cloud in figure 12 shows the most frequent word in the dataset tweets.

```
stopwords = set(STOPWORDS)

stopwords.update(["United", "AmericanAir", "SouthwestAir", "JetBlue", "USAirways", "VirginAmerica"])

# Generate a word cloud image
wordcloud = WordCloud(stopwords = stopwords, background_color="white").generate(text)

# Display our Word Cloud
plt.figure(figsize=(12,8))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

Figure 15 Word Cloud for Selected Country

From figure 13 the datasets are split based on the number of Airways that are available, and it shows tweet of the frequency of what people are saying about the different airline that are available. From the analysis in figure 13 it could be seen that American, US Airways and United trend the most.

```
plot_size[1] = 6
plt.rcParams["figure.figsize"] = plot_size

airline_tweets.airline.value_counts().plot(kind='pie', autopct='%1.0f%%')
```

```
8.0
6.0
<matplotlib.axes._subplots.AxesSubplot at 0x7ffb30972898>
```

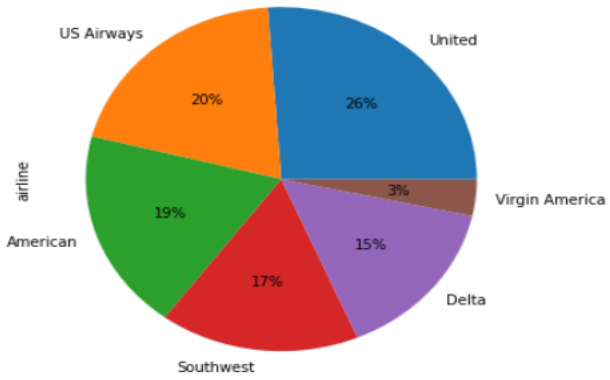


Figure 16 Frequency of Airline

From the figure 14 segregation of the dataset was done as shown to be able to identify the negative, positive and neutral comments that was in the picked airline.

```
[ ] # Ratio of Negative to Neutral to Postive sentiments
    airline_tweets.airline_sentiment.value_counts().plot(kind='pie', autopct='%1.0f%%', colors=["red", "yellow", "green"])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ffb3171be80>
```

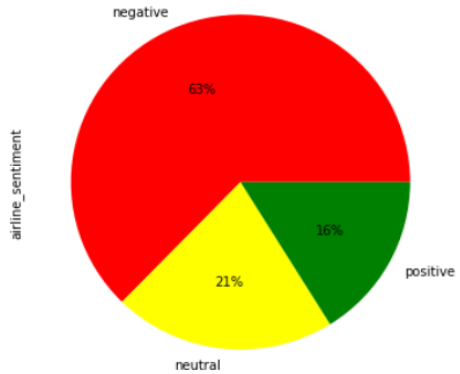


Figure 17 Classification of Sentiment

Figure 15 shows the accuracy score performance for the NLP. The image in figure 15 shows the accuracy of the different segregation for neutral, positive and negative with

the total accuracy score of 75.99% this show good level of accuracy score however this accuracy can be worked upon to improve the score.

Figure 18 Accuracy Score of NLP

```
# Let's display the results of our classifier on our test dataset

from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

print(confusion_matrix(y_test,predictions))
print(classification_report(y_test,predictions))
print("Accuracy {0:.2f}%".format(100*accuracy_score(y_test, predictions)))
```

```
[[1723  108   39]
 [ 326  248   40]
 [ 132   58  254]]
              precision    recall  f1-score   support

    negative       0.79      0.92      0.85      1870
     neutral       0.60      0.40      0.48       614
    positive       0.76      0.57      0.65       444

    accuracy                           0.76      2928
   macro avg       0.72      0.63      0.66      2928
weighted avg       0.75      0.76      0.74      2928

Accuracy 75.99%
```

CONCLUSION

In conclusion, the Natural Language processing (NLP) show the accuracy performance of the different segregation of neutral, positive and negative with total accuracy score of 75.99% this show good level of accuracy score, however this accuracy can be worked upon to improve the score.

REFERENCES

AFRIYIE, J. K., TAWIAH, K., PELS, W. A., ADDAI-HENNE, S., DWAMENA, H. A., OWIREDU, E. O., AYEH, S. A. & ESHUN, J. 2023. A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. *Decision Analytics Journal,* 6**,** 100163.

AWOYEMI, J. O., ADETUNMBI, A. O. & OLUWADARE, S. A. Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 international conference on computing networking and informatics (ICCNI), 2017. IEEE, 1-9.

BIN SULAIMAN, R., SCHETININ, V. & SANT, P. 2022. Review of machine learning approach on credit card fraud detection. *Human-Centric Intelligent Systems,* 2**,** 55-68.

BTOUSH, E. A. L. M., ZHOU, X., GURURAJAN, R., CHAN, K. C., GENRICH, R. & SANKARAN, P. 2023. A systematic review of literature on credit card cyber fraud detection using machine and deep learning. *PeerJ Computer Science,* 9**,** e1278.

CHEN, C., LI, K., OUYANG, A., ZENG, Z. & LI, K. 2018. GFlink: An in-memory computing architecture on heterogeneous CPU-GPU clusters for big data. *IEEE Transactions on Parallel and Distributed Systems,* 29**,** 1275-1288.

DAS, D. D., SHARMA, S., NATANI, S., KHARE, N. & SINGH, B. Sentimental analysis for airline twitter data. IOP conference series: materials science and engineering, 2017. IOP Publishing, 042067.

DOLNICAR, S. 2022. Market segmentation for e-tourism. *Handbook of e-Tourism.* Springer.

GAO, J., ZHOU, Z., AI, J., XIA, B. & COGGESHALL, S. 2019. Predicting credit card transaction fraud using machine learning algorithms. *Journal of Intelligent Learning Systems and Applications,* 11**,** 33-63.

GUI, L. 2019. *Application of machine learning algorithms in predicting credit card default payment*, University of California, Los Angeles.

HOSSEINI, S. 2020. *Twitter US Airline Sentiment Analysis* [Online]. Available: https://towardsdatascience.com/twitter-us-airline-sentiment-analysis-91caa7a22a93 [Accessed November 17 2023].

ILEBERI, E., SUN, Y. & WANG, Z. 2022. A machine learning based credit card fraud detection using the GA algorithm for feature selection. *Journal of Big Data,* 9**,** 1-17.

MULLEN, C. 2023. *Card industry's fraud-fighting efforts pay off: Nilson Report Published Jan. 5, 2023* [Online]. Available: https://www.paymentsdive.com/news/card-industry-fraud-fighting-efforts-pay-off-nilson-report-credit-debit/639675/ [Accessed November 14 2023].

SCARAMUCCIA, S., NANTY, S. & MASSEGLIA, F. 2020. Feedback Clustering for Online Travel Agencies Searches: a Case Study. *arXiv preprint arXiv:2007.07073*.

TABIANAN, K., VELU, S. & RAVI, V. 2022. K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. *Sustainability,* 14**,** 7243.

ZAHOOR, S. & ROHILLA, R. Twitter sentiment analysis using machine learning algorithms: a case study. 2020 International Conference on Advances in Computing, Communication & Materials (ICACCM), 2020. IEEE, 194-199.