Contents

Contents 1
1 Abstract 2
2 Introduction 2
2.1 Credit Card Fraud Detection 2
2.2 Heart Failure Prediction 2
2.3 Black Friday Sales Prediction 3
3 Data Sources 3
3.1 Dataset-1 – Credit Card Fraud Detection 3
3.2 Dataset-2 – Heart Failure Prediction 3
3.3 Dataset-3 – Black Friday Sale Prediction 3
4 Methodology 3
5 Data Cleaning and Transformation 4
5.1 Dataset-1 Credit Card Fraud Detection 4
5.2 Dataset-2 Heart Failure Prediction 5
5.3 Dataset-3 Black Friday Sales Prediction 6
6 Models Prediction and Analysis 7
6.1 Dataset 1: Credit Card Fraud Detection 7
6.1.1 MODEL 1: XG Boost 7
6.1.2 MODEL 2: Random Forest 7
6.2 Dataset 2: Heart Failure Prediction 7
6.2.1 MODEL 1: Support Vector Machine(SVM) 7
6.2.2 MODEL 2: kNN Classifier 8
6.3 Dataset 3: Black Friday Sales Prediction 8
6.3.1 MODEL 1: XG Boost 8
7 Conclusions 8
References 8

Credit Card Fraud Detection, Heart Failure and Black Friday Sales Prediction Using Data Mining and Machine Learning Models

Heena Chopra
Student No: x19205309
National College of Ireland

.

1 ABSTRACT

Three Datasets were imported, cleaned, explored, analyzed, patterned and modelled by various machine learning models using the KDD methodology. Dataset 1 was chosen from Kaggle that helped in analyzing whether the performed Credit Card transaction is fraud (CLASS '1') or not (CLASS '0'). Dataset 2 was taken from Kaggle which contains features depicting a heart failure i.e.; Will a patient get a heart failure or not based on its previous medical history? Dataset 3 was chosen from Kaggle, initially taken from Analytics Vidhya that helps in predicting the purchase amount to be spent by the customers in the upcoming Black Friday Sale. For Dataset 1, XGBoost and Random forest achieved 99.96% accuracy respectively. For Dataset 2, KNN Classifier achieved 93.33% accuracy and Support Vector Machine (SVM) achieved 90.0% Success Rate. For 3rd Dataset, Xg boost model gave the RMSE value of 2895.87.
Keywords: Credit Card Fraud, Heart Failure, Black Friday, Death Rate, Purchase Amount, Machine Learning, regression, classification, data exploration.

2 INTRODUCTION

2.1 Credit Card Fraud Detection

- Banking is a financial sector that accepts money from public and keep it as a deposit with them. In order to deposit a certain amount in bank, one needs to create an account in the bank. On creating an account in bank (Government/Private), we are provided with a number of services from bank which includes Credit Card, Debit Card, Online Banking and so on.
- Credit card is a payment card issued to account holders that enables them to make payments at various merchants for goods or services, keeping in consideration that the total amount needs to be returned to the bank along with the agreed charges at a certain time [1].
- Credit card transactions are continuing to grow in number, taking away the biggest share of payment system. Today, E-commerce and almost all online services have migrated to online payment system through credit and debit cards, increasing the risk for online frauds [2].
- Credit Card Fraud is the illegal use of credit card (e.g.: Unusual Transaction behavior) without consent of the card holder. We have a number of credit card fraud types, few of them are listed below(refer to 1).
- Credit card frauds costs consumers and financial institutions billions of dollars every year. Therefore, it has become very important to know for a bank as well as the card holder that Is the transaction made fraud or Not? Data Mining and Machine Learning is the new emerging technology that can help in detecting whether a transaction is fraud or not by detecting the hidden patterns and relationship between its previous Data Set. An attempt is made to build a classifier machine learning model which can detect the fraudulent transactions based on the previous transaction data and hence recognize the credit card fraud.

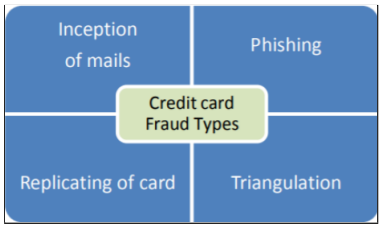

Figure 1: Types of credit card frauds

2.2 Heart Failure Prediction

- The Heart is a very essential organ of a human body. We rely on a human heart completely for oxygen and energy supply to all the other organs of the body. A Heart disease can cause irregular blood circulation all over the body that can further risk life of the person. Therefore, if heart of a human body stops working properly (as it is supposed to be), then all of the system of the human body may become dead [3].
- A Heart Failure is caused due to various risk factors which can be divided into 2 further categories; One are the factors that are caused by a person's behaviour such as Smoking, Drinking, Tobacco Use, Physical Inactivity, Unhealthy Diet i.e., these can be altered. Other factors are the ones that can't be altered such as Gender, Age and Patients Family History. Since we are well aware of the factors that cause a heart attack, we can keep a record of these risk factors for various patients who've suffered a heart attack and by looking at the patterns we can recognize or predict that Will a person get a heart attack or not?
- Prediction of a heart failure could help the medical doctors who are aiming to understand if a patient is going to have a heart failure or not based on his previous history. Keeping this question in mind a machine learning model is created using a Dataset from Kaggle, published by Davide Chicco and Giuseppe Jurman [4].

2.3 Black Friday Sales Prediction

- i.Black Friday is an informal name given to Friday that is accompanied by the Thanksgiving Day, celebrated on the fourth Thursday of November in United States of America. This day generally marks as the beginning of Christmas Shopping Season in USA. During Black Friday Sale, many shopping stores offer a highly promoted sales and give a huge discount on a number of goods and services. At this point of time of the year, retailers have high chances of getting the biggest sale of the year.
- One of the biggest applications of Computer industry is that it helps the retail industry to predict how much a customer is going to spend in a store based on its previous shopping patterns using Machine learning algorithms. If the retailers understand their customers well in terms of their shopping behaviours in the previous shopping seasons, they can enhance and develop better marketing techniques for their specific customers.
- Therefore, Machine learning models helps retail companies in analysing the consumer behaviour against the products of various types and predicting How much money will be spent by customers on purchase of products of different categories? The algorithm of the model is mainly going to predict the purchase amount going to be spent by the customers on a particular day. An attempt is made to build a machine learning model which can predict the amount to be spent by the customers based on their historic transactional data.

3 DATA SOURCES

3.1 Dataset-1 – Credit Card Fraud Detection

The Dataset for Credit Card Fraud Detection model has been taken from Kaggle. The dataset was initially collected and analysed during a research collaboration of worldline and machine learning group of ULB (Université Libre de Bruxelles) [5] on big data mining and fraud detection. The dataset holds the data for credit card transactions done by European Cardholders in September 2013.This dataset contains the transactions data for two days, where we have found 492 frauds out of 284,807. The dataset is highly unbalanced i.e., the fraud transactions accounts for 0.172% of the total transactions.

The dataset contains only numerical variables and no categorical variables. Due to Confidentiality issues, information about the original data could not be provided and therefore the numerical values are obtained by the result of PCA (Principal Component Analysis) transformation done of features. V1, V2, V3.... V28 are the principal components acquired after PCA. Two of features that have not been transformed with PCA are 'Time' and 'Amount' where time is the number of seconds passed between the first transaction and the latest transaction and Amount is the transactional amount of the credit card. Class is a dichotomous variable that can either have value 0 or 1.

3.2 Dataset-2 – Heart Failure Prediction

The Dataset for Heart Failure Prediction was taken from Kaggle where it was published by Larxel and authorised by Davide Chicco and Giuseppe Jurman. The dataset is in .csv format and contains 13 features and 299 rows. These features are some of the factors that affect the death event of a person. The dataset contains information about heart failure patients who have had a left ventricular systolic dysfunction and a heart failure of stage 1 or 2. Out of 299 patients in total, 105 are women and 194 are men and their age ranges from between 40 to 95 years old.

3.3 Dataset-3 – Black Friday Sale Prediction

The Dataset for Black Friday sale prediction was taken from Kaggle. The dataset was initially sourced from Analytics Vidhya from the Black Friday Contest. The dataset is in .csv file and contains 12 features and 5,50,069 rows. The model performance is measured by the accuracy shown while predicting the purchase amount for the test data(test.csv), that contains all the features same as that of train data except the purchase amount which needs to be predicted. The variables of the dataset are explained by Fig. 2

Variables	Description
User_ID	User ID
Product_ID	Product ID
Gender	Sex of User
Age	Age in bins
Occupation	Occupation (Masked)
City_Category	Category of the City (A,B,C)
Stay_In_Current_City_Years	Number of years stay in current city
Marital_Status	Marital Status
Product_Category_1	Product Category (Masked)
Product_Category_2	Product may belongs to other category also (Masked)
Product_Category_3	Product may belongs to other category also (Masked)
Purchase	Purchase Amount (Target Variable)

Figure 2: Data description for dataset 3

4 METHODOLOGY

The methodology that was followed for creating this report is Knowledge discovery in databases (KDD). The sequence of steps followed are as below:

- The Dataset was imported and the response variables were identified.
- Data is pre-processed, cleaned and explored using Exploratory Data Analysis to stabilize the data for further understanding of the logic to choose the best type of machine learning model.
- Data after being explored is moulded and transformed into the way, that is necessary for the application of machine learning models.
- Data Manipulation techniques such as Principal Component Analysis, Normalisation, Encoding /decoding etc. were performed on the dataset for further modelling.
- Data Mining and Machine learning techniques were then applied to the final dataset to analyse all the patterns in the data.
- At Last, all the conclusions were observed and reported to gain knowledge.

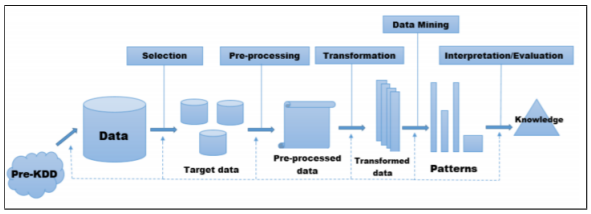

Figure 3: KDD methodology

5 DATA CLEANING AND TRANSFORMATION

5.1 Dataset-1 Credit Card Fraud Detection

- **Data Description**: A brief Overview of the Dataset is given as below:
  - The dataset contains 284,807 rows and 31 columns (Figure 4).

```
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
 19  V19     284807 non-null  float64
 20  V20     284807 non-null  float64
 21  V21     284807 non-null  float64
 22  V22     284807 non-null  float64
 23  V23     284807 non-null  float64
 24  V24     284807 non-null  float64
 25  V25     284807 non-null  float64
 26  V26     284807 non-null  float64
 27  V27     284807 non-null  float64
 28  V28     284807 non-null  float64
 29  Amount  284807 non-null  float64
 30  Class   284807 non-null  int64
dtypes: float64(30), int64(1)
memory usage: 67.4 MB
```

Figure 4: Data Overview for dataset 1

  - **NULL values**: 2.There are no null values in the dataset. (Figure 5)

	Class	V14	V1	V2	V3	V4	V5	V6	V7	V8	...	V20	V21	V22	V23	V24	V25	V26	V27	V28	Time
Total	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Percent	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

2 rows × 31 columns

Figure 5

  - **Target Variable**: The Target Variable in this dataset is 'Class'. The values for Class can either be '1' or '0' which tells whether the transaction is Fraud or Not. (Figure 6,7).

As we can see in Fig.7, the data is highly unbalanced with respect to the target variable Class.

```
Not Fraud    284315
Fraud           492
Name: Class, dtype: int64
```

Figure 6

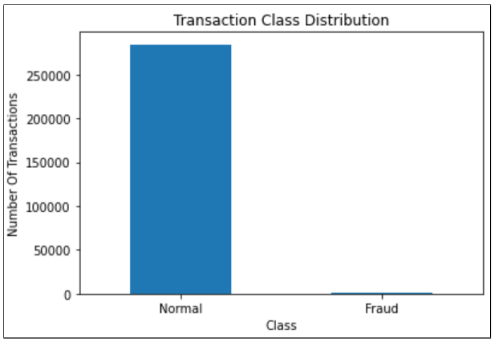

Figure 7

- **Data Exploration and Cleaning**
  - **Exploring categorical Variables**: We don't have any categorical variables in the dataset.
  - **Exploring Numerical Variables**: All the features in this dataset are numerical in nature except Class i.e., dichotomous in nature.
  - **Outliers**: The Dataset was observed carefully to see if there are any outliers in the dataset for any of the features. The real transactions had a larger number of outliers as compared to the fraudulent transactions. The outliers were represented using distribution plot and box plot as shown in figure 8.

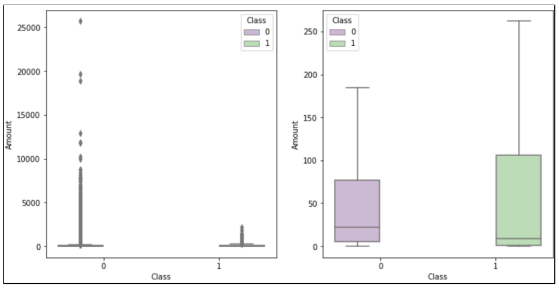

Figure 8

  - **Correlations**: There are no notable correlations between features V1-V28 as shown in Figure 9. There exists a correlation value between some of these features. Amount is directly correlated with V7 and V20 and inversely correlated with V1 and V5. Time is inversely correlated with V3. Ref to figure 10 to have a better look at the correlations using a correlation heat map.

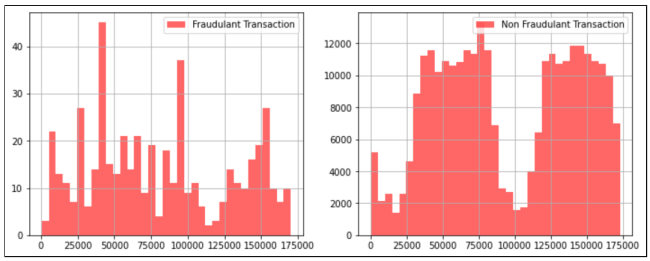

Figure 9

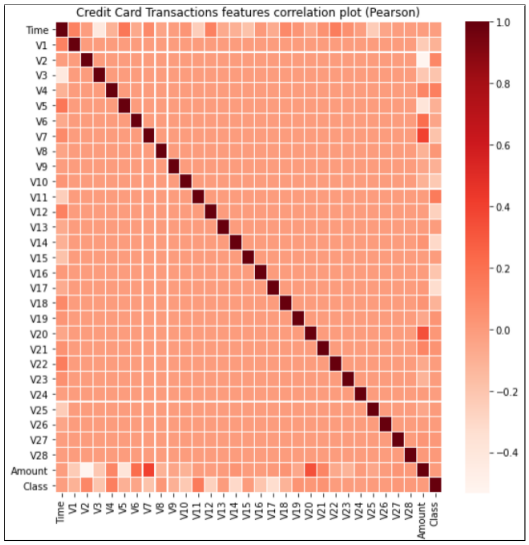

Figure 10

5.2 Dataset-2 Heart Failure Prediction

- **Data Description**: A brief Overview of the Dataset is given as below:
  - **Data Overview**: The dataset contains 299 rows and 13 columns (Figure 11).

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 299 entries, 0 to 298
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   age                       299 non-null    float64
 1   anaemia                   299 non-null    int64
 2   creatinine_phosphokinase  299 non-null    int64
 3   diabetes                  299 non-null    int64
 4   ejection_fraction         299 non-null    int64
 5   high_blood_pressure       299 non-null    int64
 6   platelets                 299 non-null    float64
 7   serum_creatinine          299 non-null    float64
 8   serum_sodium              299 non-null    int64
 9   sex                       299 non-null    int64
 10  smoking                   299 non-null    int64
 11  time                      299 non-null    int64
 12  DEATH_EVENT               299 non-null    int64
dtypes: float64(3), int64(10)
memory usage: 30.5 KB
```

Figure 11

  - **Null Values**: There are no null values in the dataset. (Figure 12)

	Total	Percent
DEATH_EVENT	0	0.0
time	0	0.0
smoking	0	0.0
sex	0	0.0
serum_sodium	0	0.0
serum_creatinine	0	0.0
platelets	0	0.0
high_blood_pressure	0	0.0
ejection_fraction	0	0.0
diabetes	0	0.0
creatinine_phosphokinase	0	0.0
anaemia	0	0.0
age	0	0.0

Figure 12

  - Target Variable: The Target Variable in this dataset is 'Death Event'. The values for Death Event can either be '1' or '0' which depicts the patient's survival from heart failure. (Figure 13, 14)

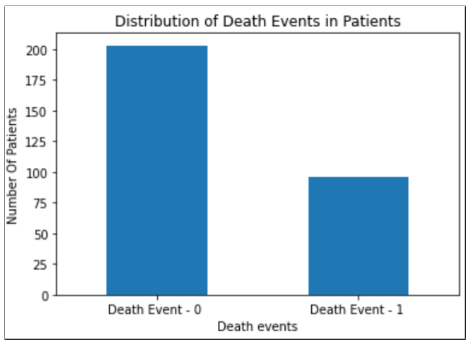

Figure 13

```
Death event - 0    203
Death Event - 1     96
Name: DEATH_EVENT, dtype: int64
```

Figure 14

- **Data Exploration and Cleaning**:
  - **Exploring categorical Variables**: There were 5 categorical variables in the dataset i.e.; anaemia, diabetes, high_blood_pressure, sex and smoking. Death_Event is the response variable but also categorical in nature.
  - **Exploring Numerical Variables**: There were 7 variables in the dataset that were numerical in nature (Age, creatinine_phosphokinase, ejection_fraction, platelets, serum_creatinine, serum_sodium, time). None having any null values.
  - **Correlations**: There were correlations between few of the features of the dataset as shown in Figure 15. There exists a considerable correlation between age, ejection_fraction, serum_creatinine, serum_sodium, and time with the DEATH_EVENT. This also shows that not considering age column for further analysis would improve the accuracy of the model.

Ref to figure 16 to have a better look at the correlations of all the features using a correlation heat map.

```
age                       0.253729
ejection_fraction        -0.268603
serum_creatinine          0.294278
serum_sodium             -0.195204
time                     -0.526964
DEATH_EVENT               1.000000
Name: DEATH_EVENT, dtype: float64
```

Figure 15

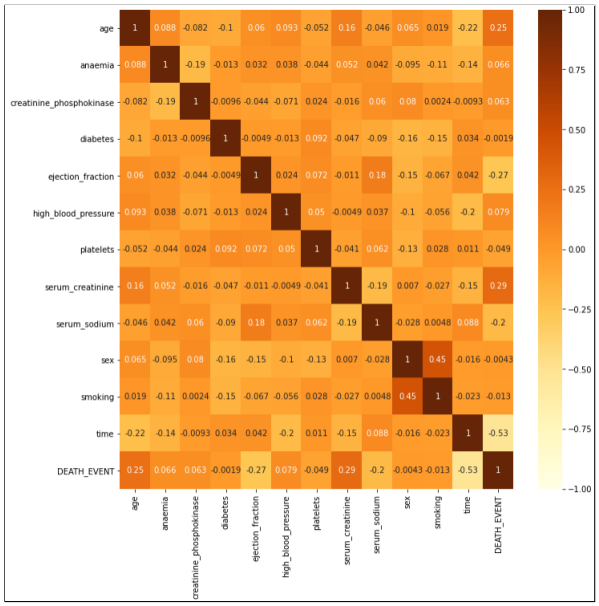

Figure 16

5.3 Dataset-3 Black Friday Sales Prediction

- **Data Description**: A brief Overview of the Dataset is given as below:
  - **Data Overview**: The dataset contains 55,068 rows and 12 columns (Figure 18). There were a few columns in dataset that are irrelevant for the analysis (User_ID and Purchase_ID), therefore these columns were removed. There were a few special characters in Age and Stay_In_Current_City_Years which were removed.

```
traindata.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 550068 entries, 0 to 550067
Data columns (total 10 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   Gender                      550068 non-null  object
 1   Age                         550068 non-null  object
 2   Occupation                  550068 non-null  int64
 3   City_Category               550068 non-null  object
 4   Stay_In_Current_City_Years  550068 non-null  object
 5   Marital_Status              550068 non-null  int64
 6   Product_Category_1          550068 non-null  int64
 7   Product_Category_2          550068 non-null  float64
 8   Product_Category_3          550068 non-null  float64
 9   Purchase                    550068 non-null  int64
dtypes: float64(2), int64(4), object(4)
memory usage: 42.0+ MB
```

Figure 17

  - **Null Values**: There are 38247 null values in the Product_Category_3 column and 173638 null values in Product_Category_2 column. (Figure 18). Null values for Product_category2 and product_category3 are filled in using the mean values of these columns. Mean value of Product_category_2 is 9.8 and Product_category_3 is 12.6. (Figure 19, 20)

	Total	Percent
Product_Category_3	383247	69.672659
Product_Category_2	173638	31.566643
Purchase	0	0.000000
Product_Category_1	0	0.000000
Marital_Status	0	0.000000
Stay_In_Current_City_Years	0	0.000000
City_Category	0	0.000000
Occupation	0	0.000000
Age	0	0.000000
Gender	0	0.000000

Figure 18

	Occupation	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
count	550068.000000	550068.000000	550068.000000	376430.000000	166821.000000	550068.000000
mean	8.076707	0.409653	5.404270	9.842329	12.668243	9263.968713
std	6.522660	0.491770	3.936211	5.086590	4.125338	5023.065394
min	0.000000	0.000000	1.000000	2.000000	3.000000	12.000000
25%	2.000000	0.000000	1.000000	5.000000	9.000000	5823.000000
50%	7.000000	0.000000	5.000000	9.000000	14.000000	8047.000000
75%	14.000000	1.000000	8.000000	15.000000	16.000000	12054.000000
max	20.000000	1.000000	20.000000	18.000000	18.000000	23961.000000

Figure 19

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 550068 entries, 0 to 550067
Data columns (total 10 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   Gender                      550068 non-null  object
 1   Age                         550068 non-null  object
 2   Occupation                  550068 non-null  int64
 3   City_Category               550068 non-null  object
 4   Stay_In_Current_City_Years  550068 non-null  object
 5   Marital_Status              550068 non-null  int64
 6   Product_Category_1          550068 non-null  int64
 7   Product_Category_2          376430 non-null  float64
 8   Product_Category_3          166821 non-null  float64
 9   Purchase                    550068 non-null  float64
dtypes: float64(3), int64(3), object(4)
memory usage: 42.0+ MB
```

Figure 20

  - **Target Variable**: The Target Variable in this dataset is 'Purchase Amount'.

Data Exploration and Cleaning:

  - **Exploring categorical Variables**: There were 7 categorical variables in the dataset i.e., Gender, Age, City_Category, Stay_In_Current_City_Years, Marital_Status, Product_Category_1, Product_Category_2, Product_Category_3. Two of them contained NULL values which were later removed.

- **Exploring Numerical Variables**: There were 2 variables in the dataset that were numerical in nature i.e., Occupation and the response variable Purchase Amount. None having any null values.
- **Correlations**: The key takes from the correlation graph is that Purchase Amount is positively correlated with 3 of the features i.e., Occupation, Stay_In_Current_City_Years and Marital Status. Increase or change in these will increase the Purchase Amount for the customer. Refer Figure 21.

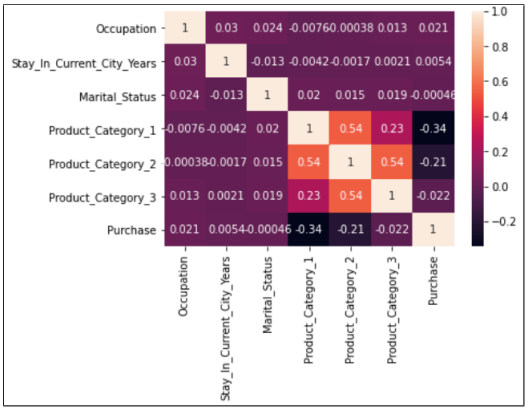

Figure 21

The same EDA Procedure was followed for the test data.

6 MODELS PREDICTION AND ANALYSIS

6.1 Dataset 1: Credit Card Fraud Detection

6.1.1 MODEL 1: XG Boost .

- **Target Variable**: 'Class'
- **Analysis**:
  - Train Accuracy: 100%
  - Test Accuracy: 99.96%
  - Confusion Matrix: Refer Figure 22.
  - Classification Report: Refer Figure 22.

  As we can see from figure 22,

```
Train Result:
================================================
Accuracy Score: 100.00%
________________________________________________
Classification Report:
                    0      1  accuracy  macro avg  weighted avg
precision         1.0    1.0       1.0        1.0           1.0
recall            1.0    1.0       1.0        1.0           1.0
f1-score          1.0    1.0       1.0        1.0           1.0
support    159204.0  287.0       1.0   159491.0      159491.0
________________________________________________
Confusion Matrix:
 [[159204      0]
 [     0    287]]

Test Result:
================================================
Accuracy Score: 99.96%
________________________________________________
Classification Report:
                     0           1   accuracy     macro avg  weighted avg
precision     0.999707    0.948718   0.999637      0.974212      0.999626
recall        0.999930    0.816176   0.999637      0.908053      0.999637
f1-score      0.999818    0.877470   0.999637      0.938644      0.999624
support    85307.000000  136.000000   0.999637  85443.000000  85443.000000
________________________________________________
Confusion Matrix:
 [[85301     6]
 [   25   111]]
```

Figure 22

- **Model Analysis**: As we can see from the figure 22, that the model achieves 100% accuracy on the training dataset and 99.6% accuracy on the test dataset. The high percentages of accuracy can be owed to the fact that xgboost implements a lot of decision trees and at a much faster rate. We can see that there are no false positives and true negative classifications for the training data since the accuracy is 100 %.

6.1.2 MODEL 2: Random Forest.

- **Target Variable**: 'Class'
- **Analysis**:
  - Train Accuracy: 100%
  - Test Accuracy: 99.96%
  - Confusion Matrix: Refer Figure 23.
  - Classification Report: Refer Figure 23.
- **Model Analysis**: The same scenario as observed in the XG Boost algorithm case, can be observed in this case too. The training accuracy is 100%, while the test accuracy is 99.6%.

```
Train Result:
================================================
Accuracy Score: 100.00%
________________________________________________
Classification Report:
                    0      1  accuracy  macro avg  weighted avg
precision         1.0    1.0       1.0        1.0           1.0
recall            1.0    1.0       1.0        1.0           1.0
f1-score          1.0    1.0       1.0        1.0           1.0
support    159204.0  287.0       1.0   159491.0      159491.0
________________________________________________
Confusion Matrix:
 [[159204      0]
 [     0    287]]

Test Result:
================================================
Accuracy Score: 99.96%
________________________________________________
Classification Report:
                     0           1   accuracy     macro avg  weighted avg
precision     0.999730    0.926230   0.999625      0.962980      0.999613
recall        0.999894    0.830882   0.999625      0.915388      0.999625
f1-score      0.999812    0.875969   0.999625      0.937891      0.999615
support    85307.000000  136.000000   0.999625  85443.000000  85443.000000
________________________________________________
Confusion Matrix:
 [[85298     9]
 [   23   113]]
```

Figure 23

6.2 Dataset 2: Heart Failure Prediction

6.2.1 MODEL 1: Support Vector Machine(SVM).

- **Target Variable**: 'Death_Event'
- **Analysis**:
  - Accuracy: 90.00
  - Confusion Matrix: Refer Figure 24.
  - Classification Report: Refer Figure 24.

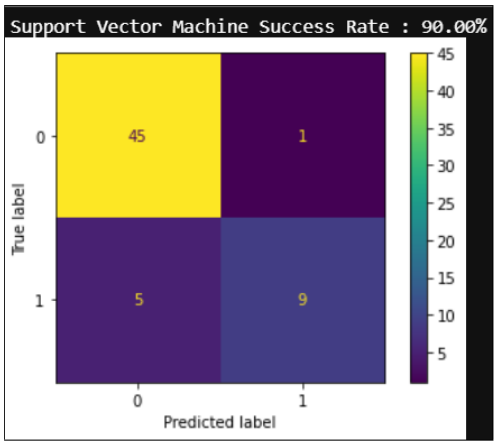

Figure 24

- **Model Analysis**: As we can see from the figure, the Support Vector Algorithm achieves an accuracy of 90% which can be considered as optimum. The confusion matrix depicts this scenario.

6.2.2 MODEL 2: kNN Classifier.

- **Target Variable**: 'Death_Event'
- **Analysis**:
  - Accuracy: 93.33%
  - Confusion Matrix: Refer Figure 25.
  - Classification Report: Refer Figure 25.

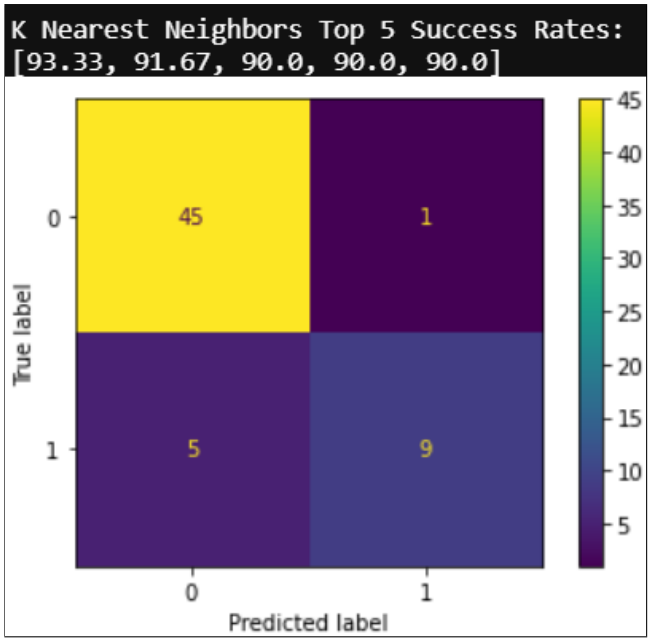

Figure 25

- **Model Analysis**: The top 3 success rates for the kNN classifier are 93.33%, 91.67%, 90.0%. These three scores correspond to different values of k. These scores can be considered as optimum for our case.

6.3 Dataset 3: Black Friday Sales Prediction

6.3.1 MODEL 1: XG Boost.

- **Target Variable**: 'Purchase_Amount'
- **Analysis**:
  - **RMSE**: 2895.87
  - **Confusion Matrix**: Refer Figure 26.
  - **Classification Report**: Refer Figure 26.
- **Model Analysis**: The XG boost gives the RMSE value of 2895.87 which is pretty good considering the range of the target variables in our data.

```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
             importance_type='gain', interaction_constraints='',
             learning_rate=1.0, max_delta_step=0, max_depth=6,
             min_child_weight=40, missing=nan, monotone_constraints='()',
             n_estimators=100, n_jobs=8, num_parallel_tree=1, random_state=0,
             reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=0, subsample=1,
             tree_method='exact', validate_parameters=1, verbosity=None)
RMSE value for XGBoost is 2895.871824558747
```

Figure 26

7 CONCLUSIONS

For the dataset 1, both models showed similar training and test accuracy. The accuracy for training set was 100% while the accuracy for test was 99.6%. Hence both models i.e *Random Forest* and *XG Boost* fit the data extremely well. For dataset 2 which was a classification task, the SVM showed the accuracy of 90% while the kNN classifier showed an accuracy of 93.3%. These can very well be considered as optimum values for our data. For the dataset 3, the XG boost gives the RMSE score of 2895.87.

REFERENCES

[1] Wikipedia, "*'Credit Card Fraud'*," Available: https://en.wikipedia.org/wiki/Credit_card, 2020.
[2] V. N. Dornadula and S. Geetha, "Credit card fraud detection using machine learning algorithms," *Procedia Computer Science*, vol. 165, pp. 631–641, 2019.
[3] A. Javeed, S. S. Rizvi, S. Zhou, R. Riaz, S. U. Khan, and S. J. Kwon, "Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification," *Mobile Information Systems*, vol. 2020, 2020.
[4] D. Chicco and G. Jurman, "*'Heart Failure Data'*," Available: https://www.kaggle.com/andrewmvd/heart-failure-clinical-data, 2020.
[5] ULB, "*'Credit Card Fraud'*," Available: https://www.kaggle.com/mlg-ulb/creditcardfraud, 2020.