CS 571: Data Preparation & Analysis Credit Card Fraud Detection Spring 2023

Pradyothan Govrineni	A20438408
Uttam Kotadiya	A20480934
Harsh Dungrani	A20514062
Sohaib Syed	A20439074

Executive Summary

- We aimed to develop a machine learning classifier system using R to identify fraudulent credit card transactions.
- The project utilized a step-by-step approach, including data preparation, exploratory data analysis, model training, validation, testing, and evaluation.
- The Card Transactions dataset was the primary source of data for the development of the classifier system.
- Various machine learning algorithms were explored during the model training phase, including: logistic regression, and decision trees.
- We concluded the project by determining the accuracy and interpretability of the developed classifier system, as well as its ability to identify fraudulent credit card transactions with high confidence.

Overview

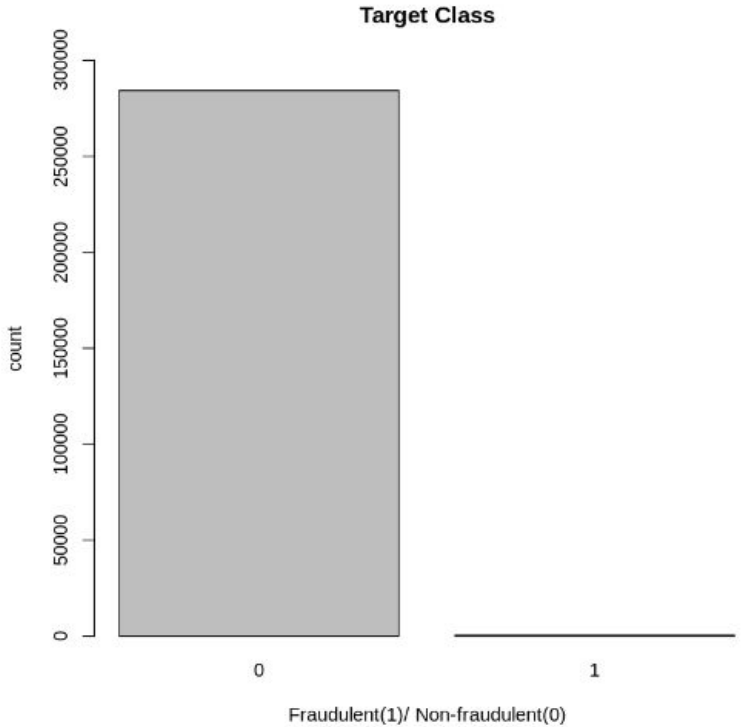
- Project plan and details includes the timeline, methodology, and tasks we divided among ourselves
- Most of the tasks were carried out as expected
- Due to confidentiality issues, in dataset 1 the original features and more background information about the data was not provided
- Boruta algorithm was not applied successfully to find the relevant features from the dataset 2

Data Exploration

Distribution of Target class in dataset 1

'Target Class Distribution: 0: Non-Fraudulent, 1: Fraudulent'

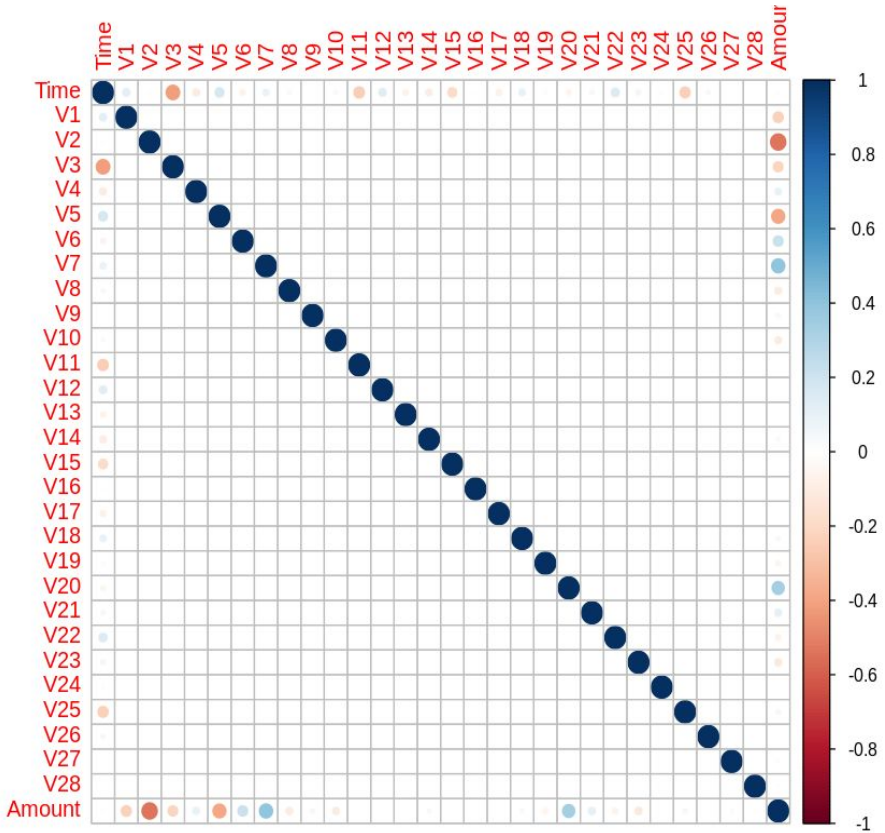
0	1
284315	492



Data Exploration

The correlation map for dataset 1 shows that the variables ‘V2’ and ‘Amount’ have a negative correlation with each other

Rest of the features have weak correlation among each other

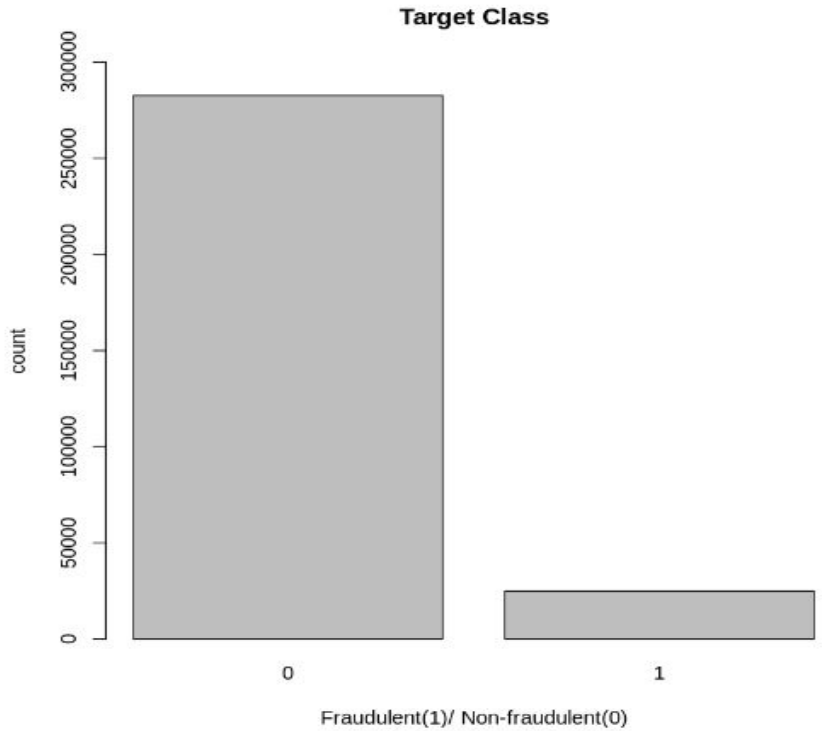


Data Exploration

Distribution of Target class in dataset 2

'Target Class Distribution: 0: Non-Fraudulent, 1: Fraudulent'

0	1
282686	24825



Data Preprocessing

- For dataset 1:
  - Number of rows: 284807, Number of columns: 31
  - Features V1,V2,...,V28 were principal components obtained with PCA.
  - No NA values were present
  - Variable datatypes were 30 numeric variables and 1 integer type variable
- For dataset 2:
  - Number of rows: 307511, Number of columns: 122
  - NA values were present
  - Variable datatypes were 65 numeric variables and 41 integer type variables and 16 character type variables

Data Preprocessing

- Dropped the columns which had large number of NA values
- Omitted the rows with any NA values
- The final size of the dataset was:
  - Number of rows: 306199
  - Number of columns: 63
- Data types in the dataframe were as follows: 11 character type, 40 integer type and 12 numeric type variables
- Transformed the 11 character type variables to numeric type variables

Model Training (Dataset 1)

Logistic Model:

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 56857    29
         1     6    69

               Accuracy : 0.9994
                 95% CI : (0.9991, 0.9996)
    No Information Rate : 0.9983
    P-Value [Acc > NIR] : 1.953e-13

                  Kappa : 0.7974

 Mcnemar's Test P-Value : 0.0002003

            Sensitivity : 0.9999
            Specificity : 0.7041
         Pos Pred Value : 0.9995
         Neg Pred Value : 0.9200
             Prevalence : 0.9983
         Detection Rate : 0.9982
   Detection Prevalence : 0.9987
      Balanced Accuracy : 0.8520

       'Positive' Class : 0
```

Imbalanced data

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 56062     3
         1   816    80

               Accuracy : 0.9856
                 95% CI : (0.9846, 0.9866)
    No Information Rate : 0.9985
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1612

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.963855
            Specificity : 0.985654
         Pos Pred Value : 0.089286
         Neg Pred Value : 0.999946
             Prevalence : 0.001457
         Detection Rate : 0.001404
   Detection Prevalence : 0.015730
      Balanced Accuracy : 0.974754

       'Positive' Class : 1

>
```

SMOTE data

Model Training (Dataset 1)

Decision Tree Model:

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 56845    21
         1    18    77

               Accuracy : 0.9993
                 95% CI : (0.9991, 0.9995)
    No Information Rate : 0.9983
    P-Value [Acc > NIR] : 9.766e-12

                  Kappa : 0.7976

 Mcnemar's Test P-Value : 0.7488

            Sensitivity : 0.9997
            Specificity : 0.7857
         Pos Pred Value : 0.9996
         Neg Pred Value : 0.8105
             Prevalence : 0.9983
         Detection Rate : 0.9980
   Detection Prevalence : 0.9983
      Balanced Accuracy : 0.8927

       'Positive' Class : 0
```

Imbalanced data

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 56303     7
         1   575    76

               Accuracy : 0.9898
                 95% CI : (0.9889, 0.9906)
    No Information Rate : 0.9985
    P-Value [Acc > NIR] : 1

                  Kappa : 0.205

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.915663
            Specificity : 0.989891
         Pos Pred Value : 0.116743
         Neg Pred Value : 0.999876
             Prevalence : 0.001457
         Detection Rate : 0.001334
   Detection Prevalence : 0.011429
      Balanced Accuracy : 0.952777

       'Positive' Class : 1
```

SMOTE data

Model Training (Dataset 2)

Logistic Model:

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 56282  4952
         1     4     1

               Accuracy : 0.9191
                 95% CI : (0.9169, 0.9212)
    No Information Rate : 0.9191
    P-Value [Acc > NIR] : 0.5215

                  Kappa : 2e-04

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9999289
            Specificity : 0.0002019
         Pos Pred Value : 0.9191299
         Neg Pred Value : 0.2000000
             Prevalence : 0.9191202
         Detection Rate : 0.9190549
   Detection Prevalence : 0.9999184
      Balanced Accuracy : 0.5000654

       'Positive' Class : 0
```

Imbalanced data

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2
         1 36096 20082
         2 20190 34407

               Accuracy : 0.6365
                 95% CI : (0.6336, 0.6393)
    No Information Rate : 0.5081
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.2727

 Mcnemar's Test P-Value : 0.5939

            Sensitivity : 0.6413
            Specificity : 0.6314
         Pos Pred Value : 0.6425
         Neg Pred Value : 0.6302
             Prevalence : 0.5081
         Detection Rate : 0.3258
   Detection Prevalence : 0.5071
      Balanced Accuracy : 0.6364

       'Positive' Class : 1
```

SMOTE data

Model Training (Dataset 2)

Decision Tree Model: The decision tree for the imbalanced dataset two was obtained as follows which was not at all acceptable and was highly biased to the majority class in the dataset

Model Training (Dataset 2)

Performance of Decision Tree Model on balanced data:

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2
         1 55630 13389
         2   656 41100

               Accuracy : 0.8732
                 95% CI : (0.8712, 0.8752)
    No Information Rate : 0.5081
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7454

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9883
            Specificity : 0.7543
         Pos Pred Value : 0.8060
         Neg Pred Value : 0.9843
             Prevalence : 0.5081
         Detection Rate : 0.5022
   Detection Prevalence : 0.6231
      Balanced Accuracy : 0.8713

       'Positive' Class : 1
```

Model Training (Dataset 2)

Feature Selection Based on Variable Importance

by Decision Tree Model:

Variable	Importance
<chr>	<dbl>
NAME_EDUCATION_TYPE	4.950860e+04
REG_CITY_NOT_WORK_CITY	3.303368e+04
LIVE_CITY_NOT_WORK_CITY	2.682830e+04
REGION_RATING_CLIENT	2.267733e+04
FLAG_PHONE	1.725110e+04
REGION_RATING_CLIENT_W_CITY	1.699629e+04
REG_CITY_NOT_LIVE_CITY	1.244447e+04
FLAG_OWN_CAR	1.242587e+04
REGION_POPULATION_RELATIVE	3.878306e+03
CNT_CHILDREN	3.172776e+03
OBS_60_CNT_SOCIAL_CIRCLE	1.370675e+03
OBS_30_CNT_SOCIAL_CIRCLE	1.364034e+03
CNT_FAM_MEMBERS	1.346038e+03
FLAG_OWN_REALTY	1.344775e+03
NAME_INCOME_TYPE	1.312754e+03
REG_REGION_NOT_WORK_REGION	1.123527e+03
LIVE_REGION_NOT_WORK_REGION	1.086017e+03
REG_REGION_NOT_LIVE_REGION	5.515178e+02
HOUR_APPR_PROCESS_START	1.447517e+02
FLAG_DOCUMENT_8	1.440828e+02
FLAG_WORK_PHONE	1.429277e+02
NAME_HOUSING_TYPE	1.198994e+02

Model Training (Dataset 2)

Based on the selected 18

features logisitic and decision tree

model:

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2
         1 33813 23524
         2 22465 30973

               Accuracy : 0.5848
                 95% CI : (0.5819, 0.58
    No Information Rate : 0.508
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1692

 Mcnemar's Test P-Value : 8.075e-07

            Sensitivity : 0.6008
            Specificity : 0.5683
         Pos Pred Value : 0.5897
         Neg Pred Value : 0.5796
             Prevalence : 0.5080
         Detection Rate : 0.3052
   Detection Prevalence : 0.5176
      Balanced Accuracy : 0.5846

       'Positive' Class : 1
```

Logistic

```
Confusion Matrix and Statistics

          Reference
Prediction     1     2
         1 55589 13383
         2   689 41114

               Accuracy : 0.873
                 95% CI : (0.871, 0.8749)
    No Information Rate : 0.508
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7449

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9878
            Specificity : 0.7544
         Pos Pred Value : 0.8060
         Neg Pred Value : 0.9835
             Prevalence : 0.5080
         Detection Rate : 0.5018
   Detection Prevalence : 0.6226
      Balanced Accuracy : 0.8711

       'Positive' Class : 1
```

Decision Tree

Results (Dataset 1)

- SMOTE technique was especially helpful in decreasing the False Negatives
  - Fraud domain, better to be safe
- Both models improved
  - Logistic model was better in terms of Balanced Accuracy
  - DT split on only 1 variable, V14

Results (Dataset 2)

- We applied preprocessing technique to handle the NA values and transform the dataset into a dataframe which is usable for fitting logistic and decision tree models
- Performance of both the models on imbalanced data was not acceptable
- Models were highly biased to the majority class
- SMOTE helped in enhancing the performance of the models in performing the classification task
- Variable importance from Decision Tree helped in identifying the relevant features
- Performance of both models on the selected features dataframe was similar to the previous dataframe consisting of all the features

Future Work

- Future work could improve interpretability and explore ensemble methods to enhance model performance
- Real-time data feeds could be incorporated to improve the model's ability to detect fraudulent transactions in real-time
- This could prevent financial loss due to fraudulent activities
- Try to get some successful output from the Boruta algorithm and get the features deemed important and unimportant
- Use Recursive Feature Elimination (RFE) algorithm to perform feature selection