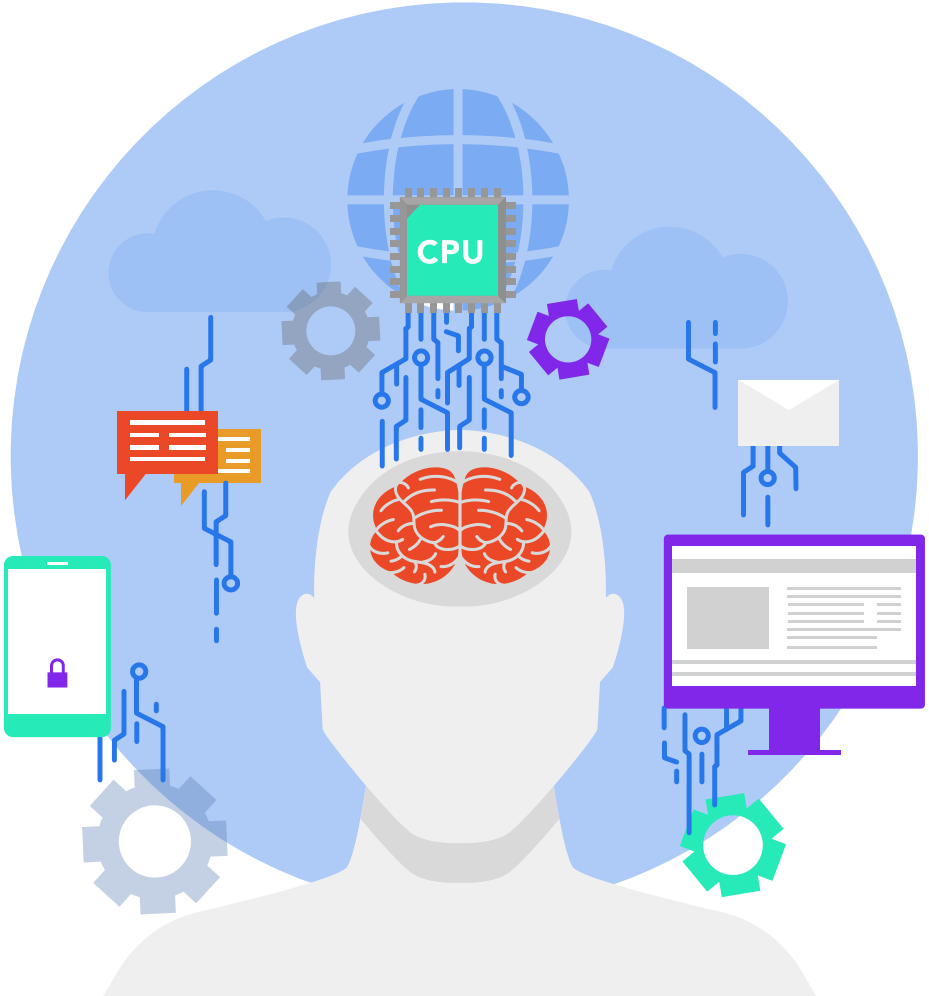
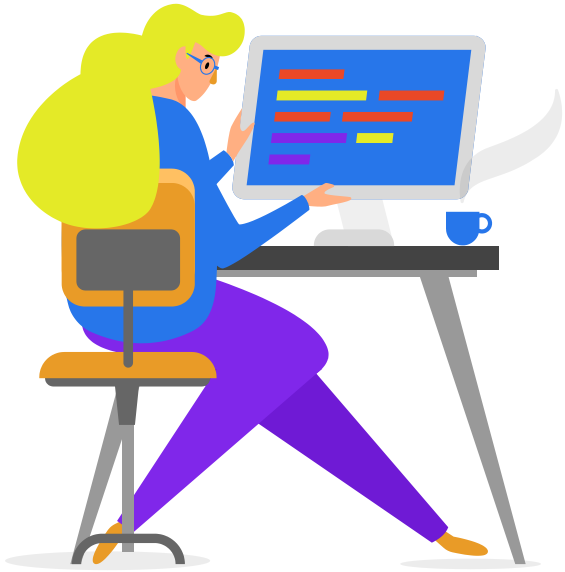


Bank Account Fraud

António Pedro Pinheiro
up201704931

Table of Contents

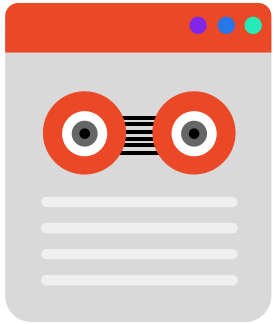


Business Understanding: Problem Definition

Goal

Problem Statement

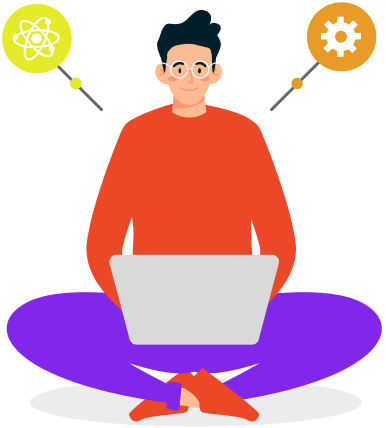
Detect fradulent applications to bank accounts by training a machine learning model that accurately predicts if we're dealing with credit fraud



Output

Classification Problem

Binary: Fraud vs. No Fraud



How?

ML Models

Predicting the outcome given a set of predictors

Success

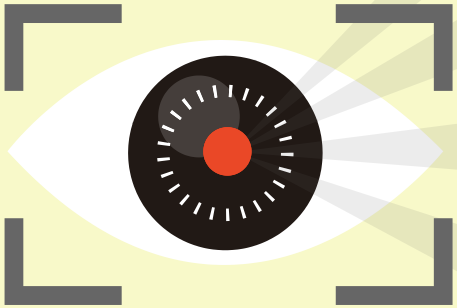
Metrics

Area under Curve

Data Understanding: Summarization

Dataset
36 predictors
1 target variable

Analysis & Summary



Significant portion of missing values
prev_address_months_count
bank_months_count

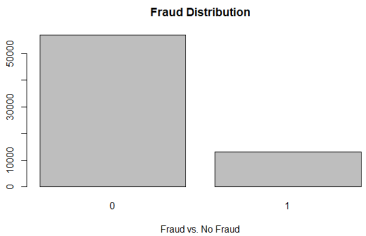
Device Fraud Count
All cases with value 0

Customer Age
Cases <10 & > 90

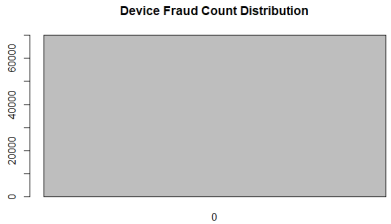
Imbalanced Domain Learning
81% of cases on the train dataset have “no fraud” status

Data Understanding: Visualization

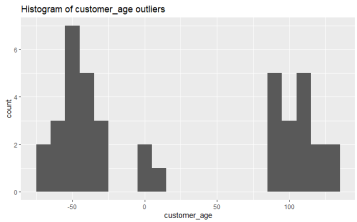
Imbalanced Domain Learning



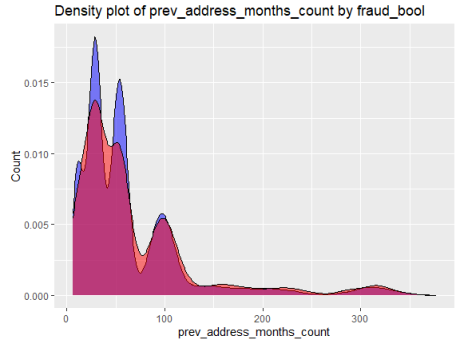
Device Fraud Count

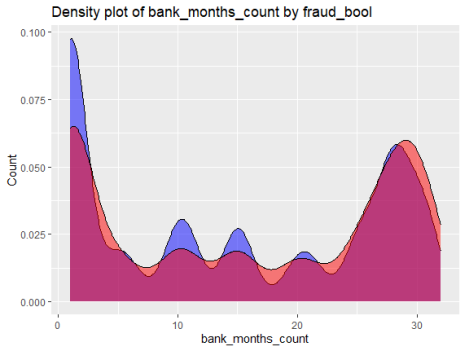


Customer Age



Significant portion of missing values





Data Preparation: Data Quality & Transformation

01

Sampling Data

Stratified Train/Validation Sets (70/30)

02

Training Set: Data Cleaning

Elimination of cases with customer_age <10 & >90

03

Missing Values

Imputation of Mean and Mode in numeric and categorical variables, respectively

04

Data Normalization

Z-Score normalization

05

Training Set: Removal of Redundant & Unecessary Attributes

- ID
- prev_address_months_count
- bank_month_count
- device_fraud_count

Data Preparation: Feature Selection

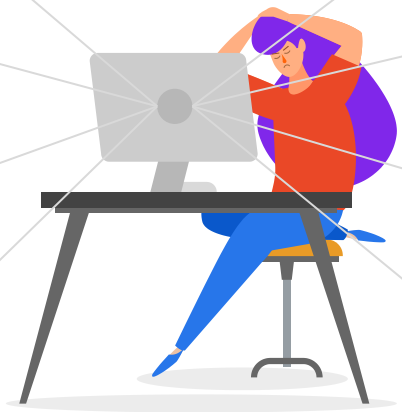
Methods

01 Correlation Matrix
Numeric Variables

02 Chi-Square Test
Categorical Variables

03 ANOVA One-Way Test
Numeric/Categorical Variables

04 Information Gain
Reduction in entropy



05 Variable Importance with Random Forests
Mean Decreasing Gini

06 Boruta Algorithm (Wrapper Method)
Copied shuffled features

07 Variable Importance with Rpart
Recursive Partitioning and Regression Trees

08 PCA
Principal Component Analysis

Predictive Modelling:

Training + Hyperparameter Tuning + Evaluation Methodologies + Prediction

Prediction on Test Set

Submission on Kaggle

04

03

Evaluation on Validation Set

- Accuracy
- Precison/Recall /F-score
- **AUC (chosen metric)**

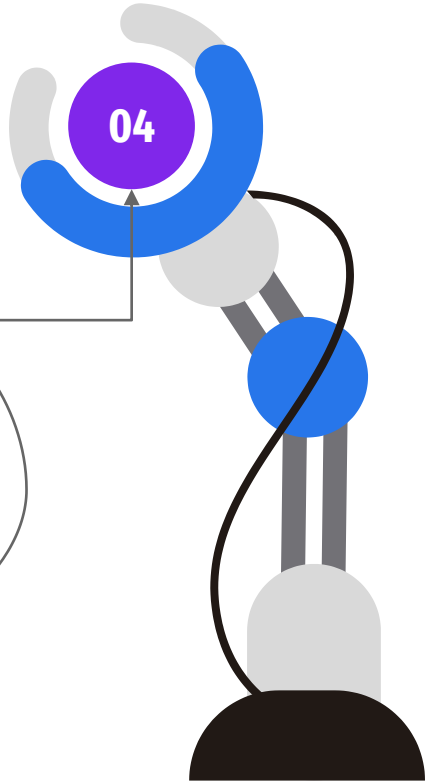
Model Training on Train Set

- k-fold Cross Validation
- Grid/Random Search Hyperparameter Tuning

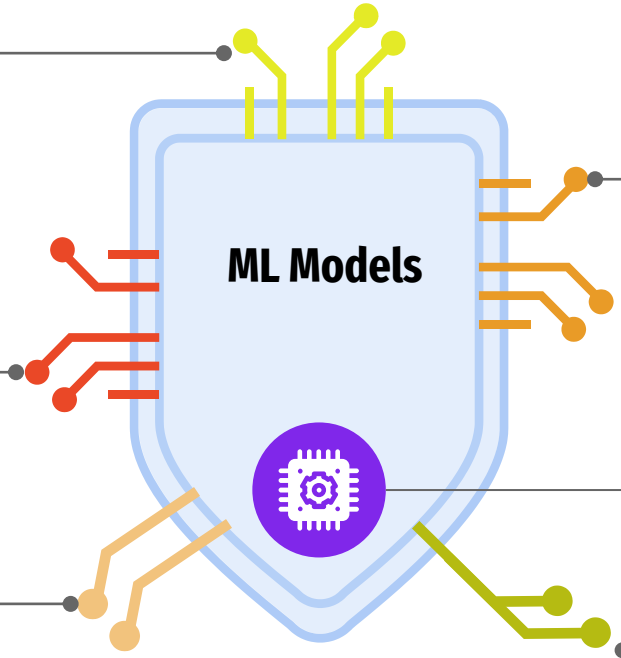
02

Stratified Train/Validation Sets

01



Predictive Modelling: Moldel Comparison



01

k-Nearest Neighbours

- k = 1
- AUC = 0,72518

02

XGBoost

- mtry =13
- min_n = 6
- tree_depth = 10
- learn_rate = 3,4 x 10^{7}
- loss_Reduction = 1,1 x 10^{9}
- AUC = 0,74559

03

Naïve-Bayes

- laplace = 0
- usekernel = TRUE
- adjust = 1
- AUC = 0,74124

04

Random Forest

- mtry = 5
- AUC = 0,74751

05

Bagged CART Trees

- AUC = 0,73469

06

CART Trees

- cp = 0
- AUC = 0,71013

Conclusion

Conclusions & Limitations

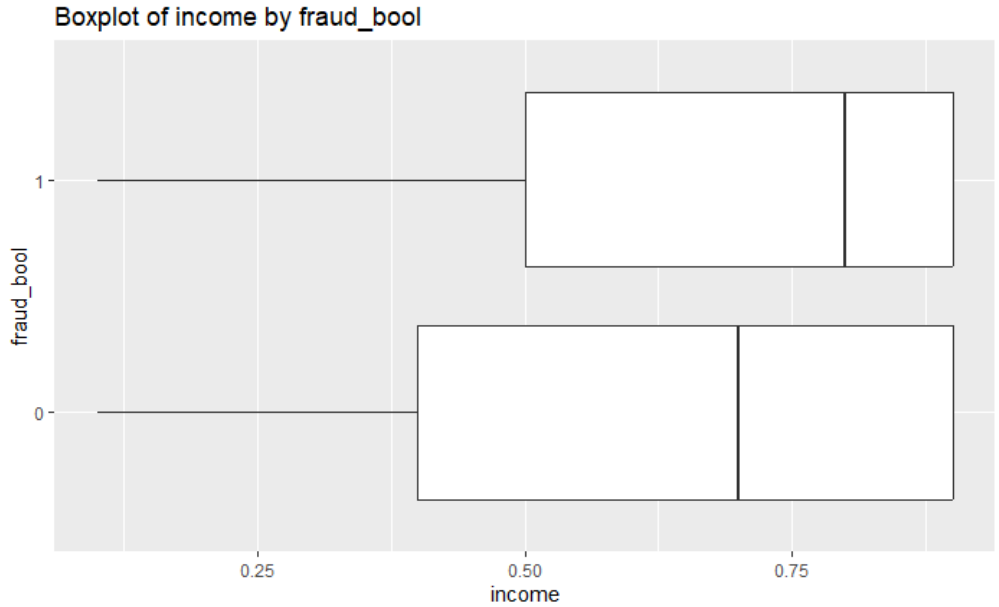
- Imbalanced Learning is a prevalent issue
- Recall is too low because the prevalent class is 0 in the target variable
- Accuracy is not a relevant metric
- AUC is more suited for this analysis
- Ensemble Models have better scores than primary models
- Final Chosen Model: Random Forest

Future Work

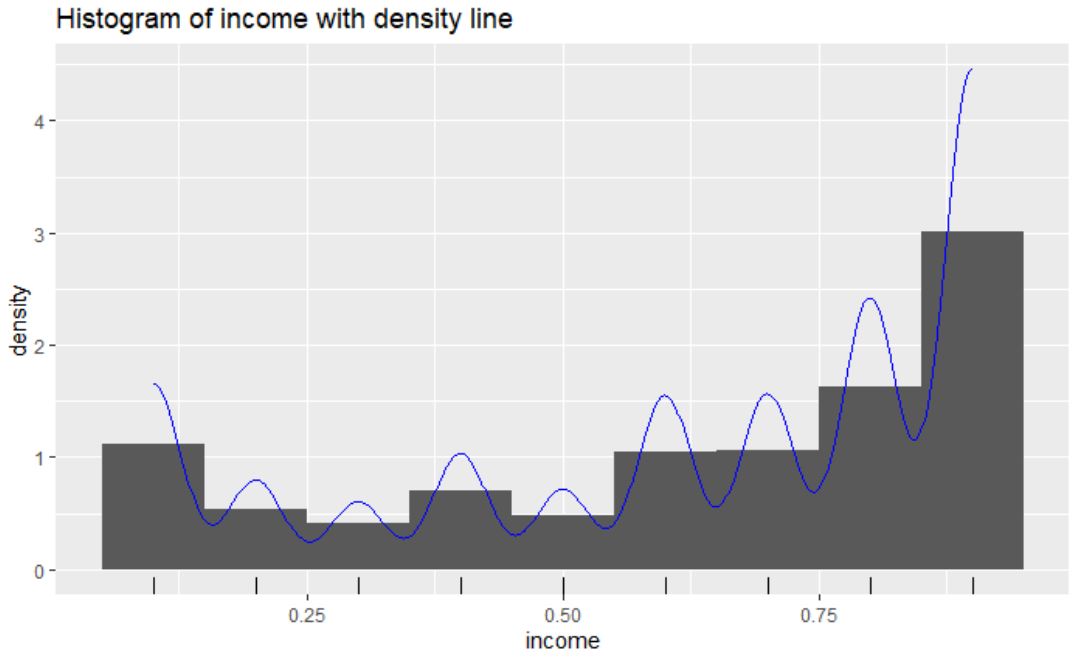
- Explore other strategies for imbalanced domains
- Train with more diverse models
- In-Depth Study to advance the feature selection techniques
- Try extracting insightful information from other clustering techniques

Annexes

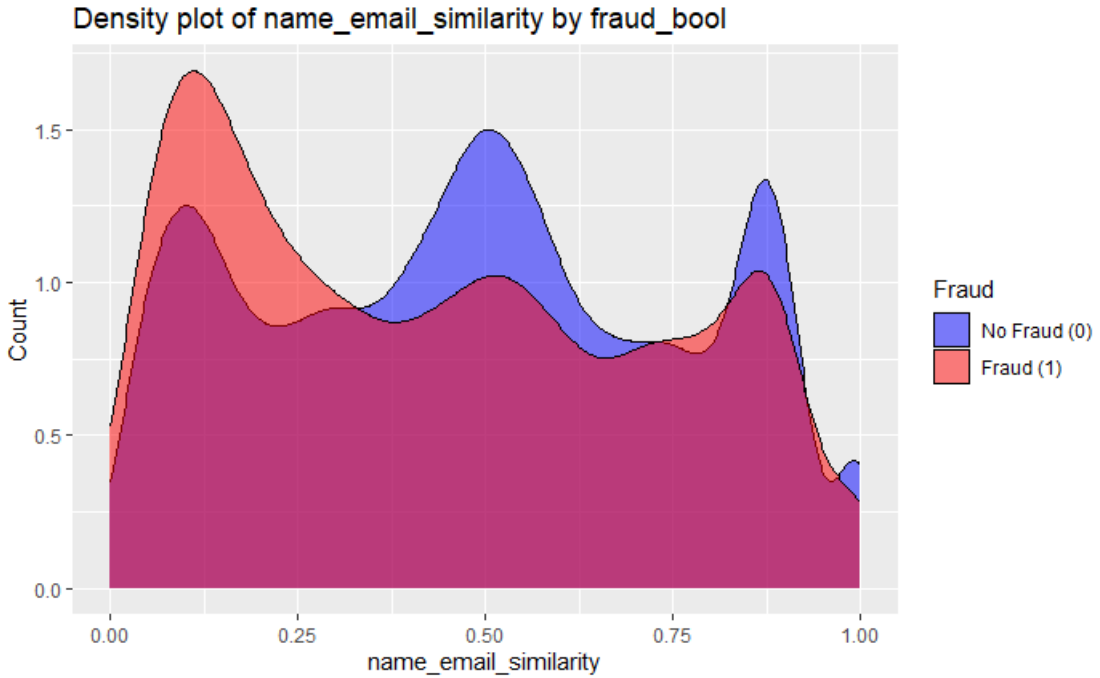
Annex A: Income Boxplot by fraud_bool



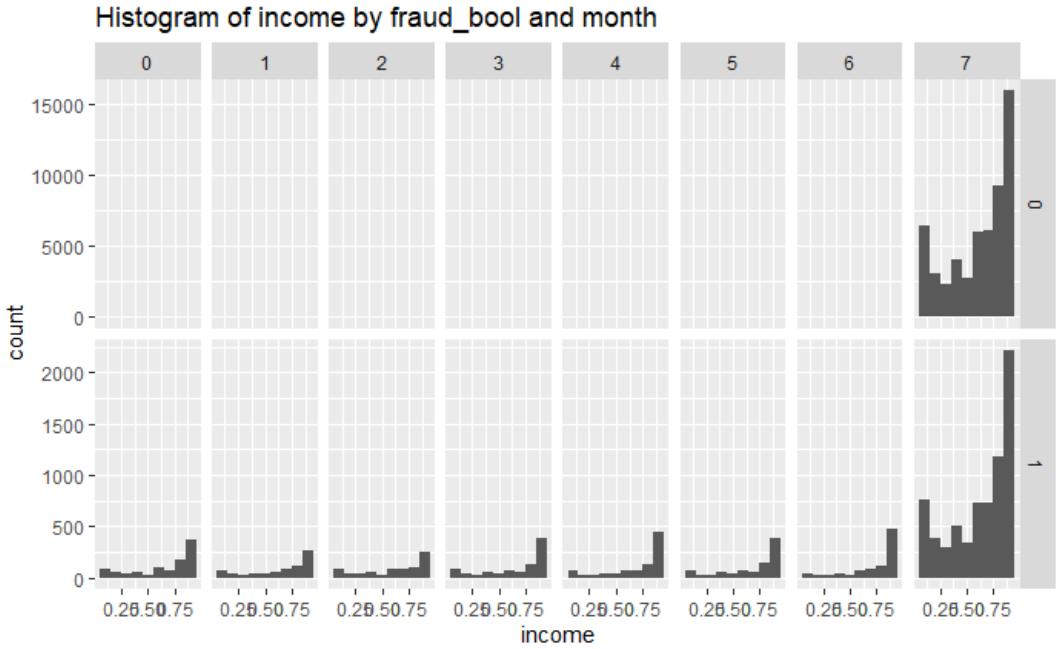
Annex B: Income Histogram with Density Line



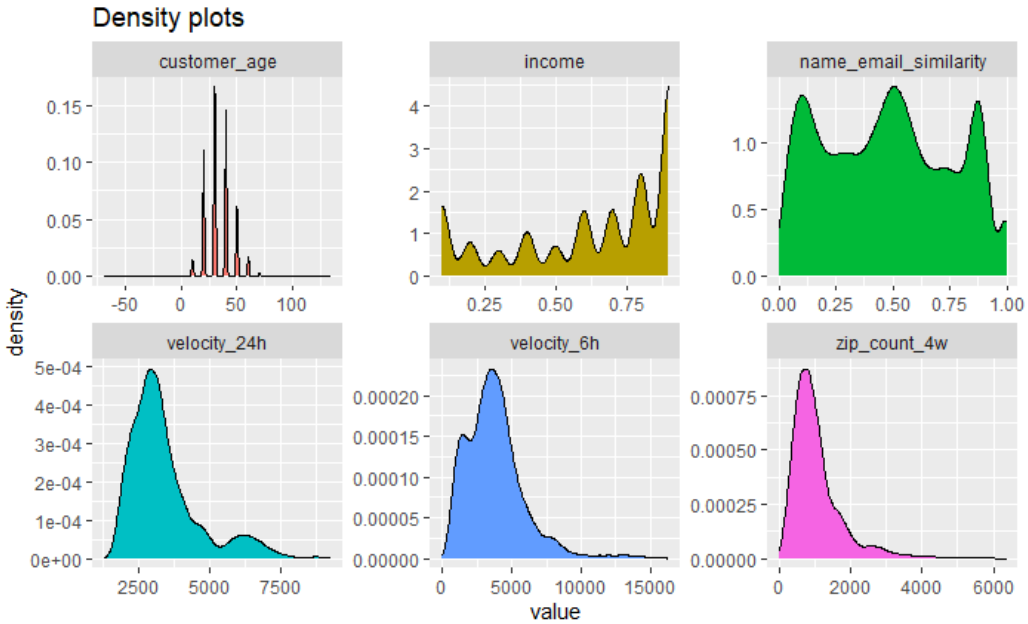
Annex C: Density Plot of name_email_similarity by fraud_bool



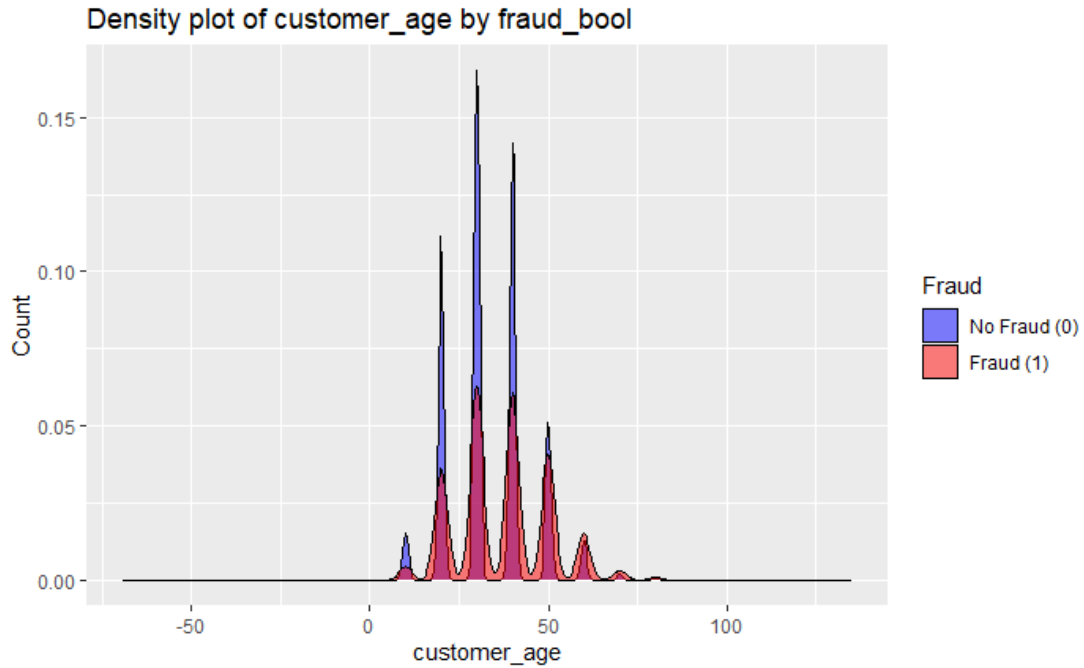
Annex D: Income Histogram by fraud_bool + month



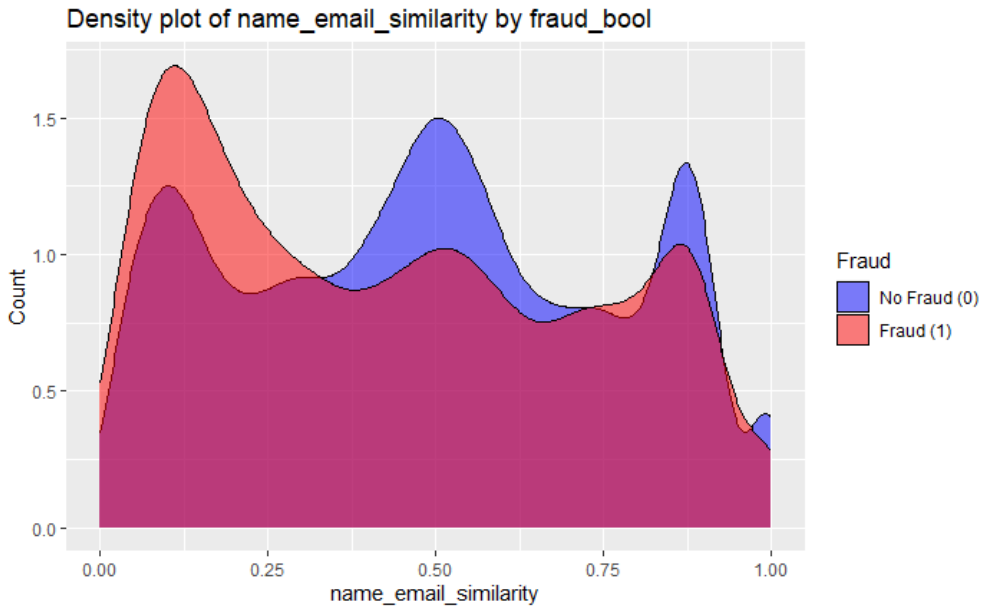
Annex E: Density Plots



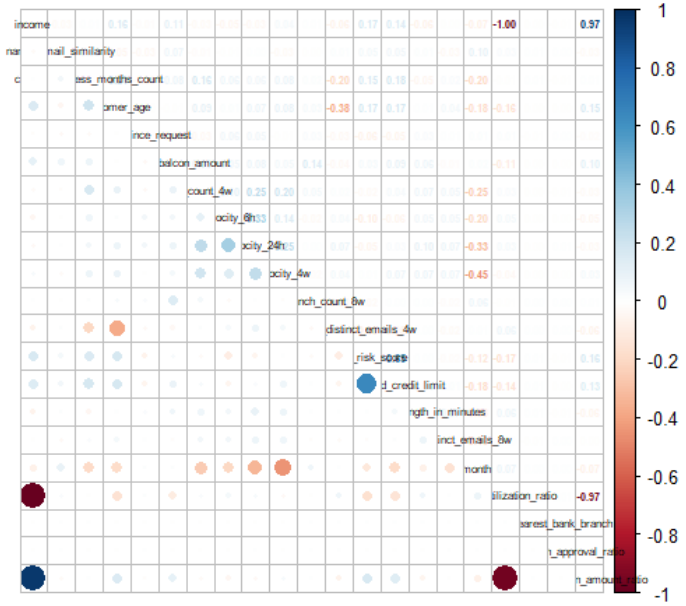
Annex F: Density Plot of customer_age by fraud_bool



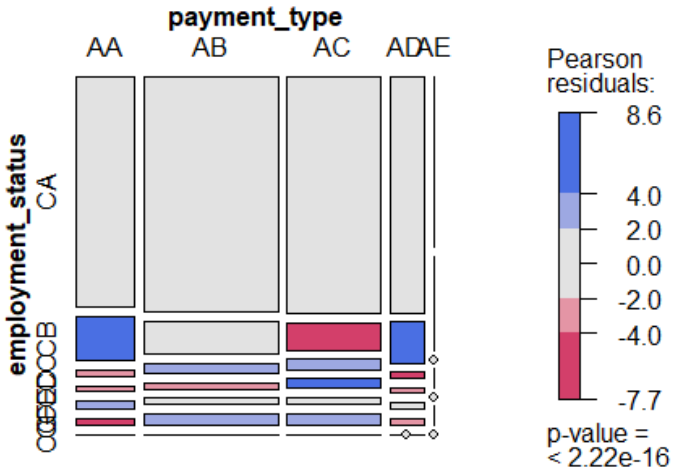
Annex G: Density Plot of customer_age by fraud_bool



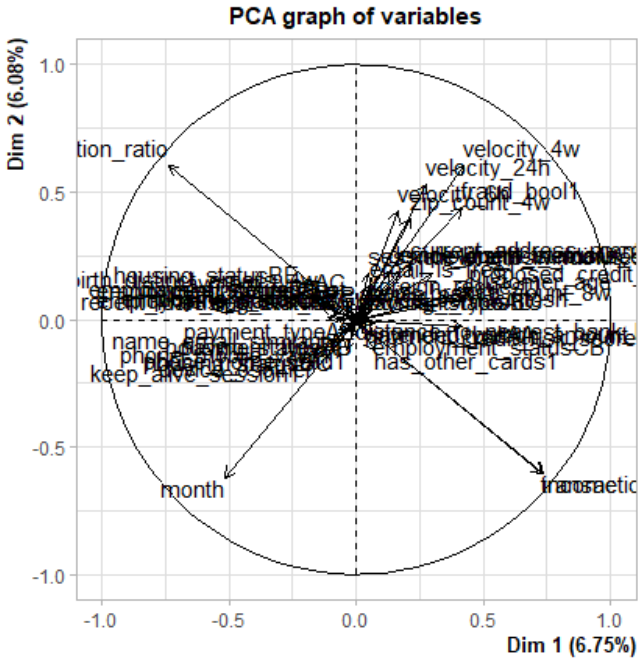
Annex H: Correlation Matrix (Spearman Coefficients)



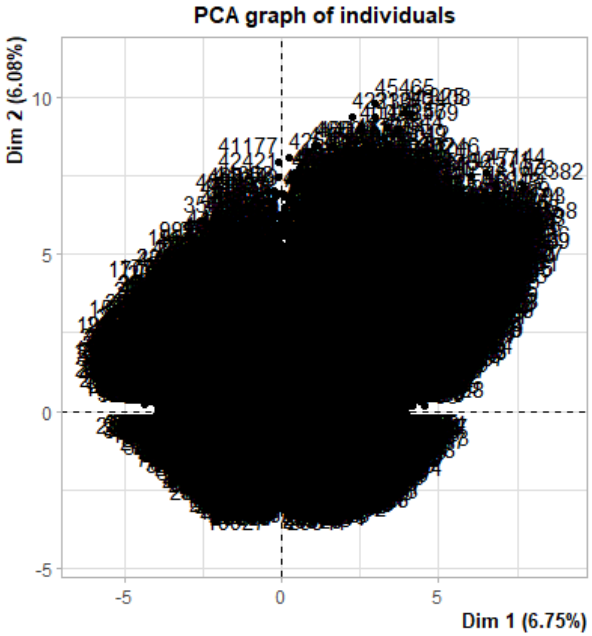
Annex I: Chi-Square Test between payment_type & emplyment status



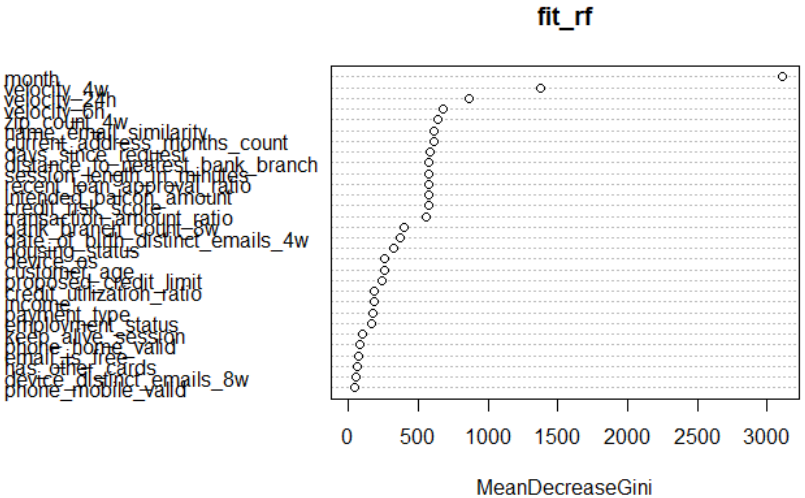
Annex J: PCA Graph of Variables/Dimensions



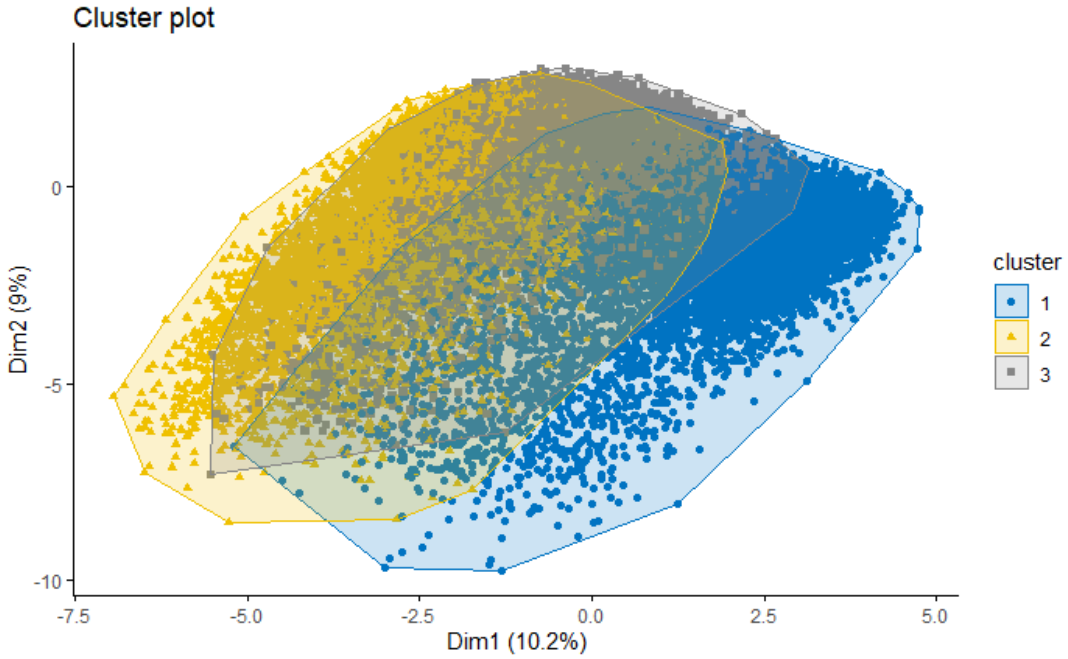
Annex K: PCA Graph of Individuals



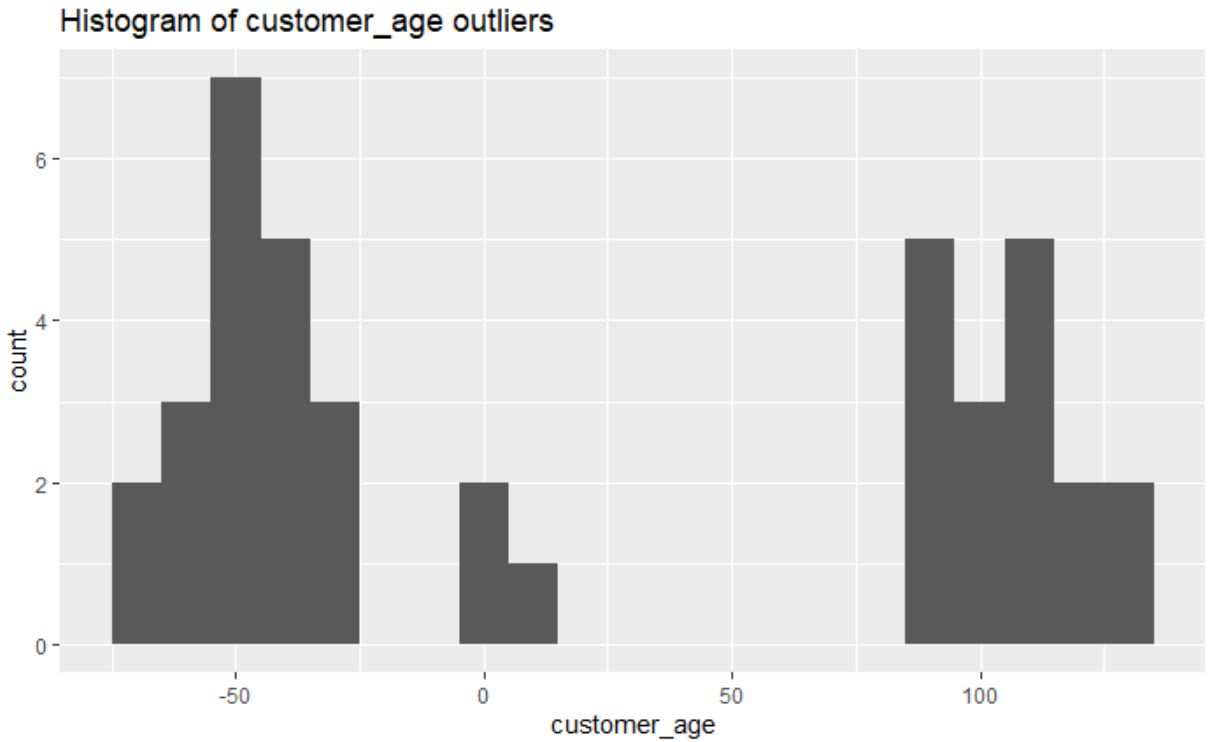
Annex L: Variable Importance in a Random Forest Model



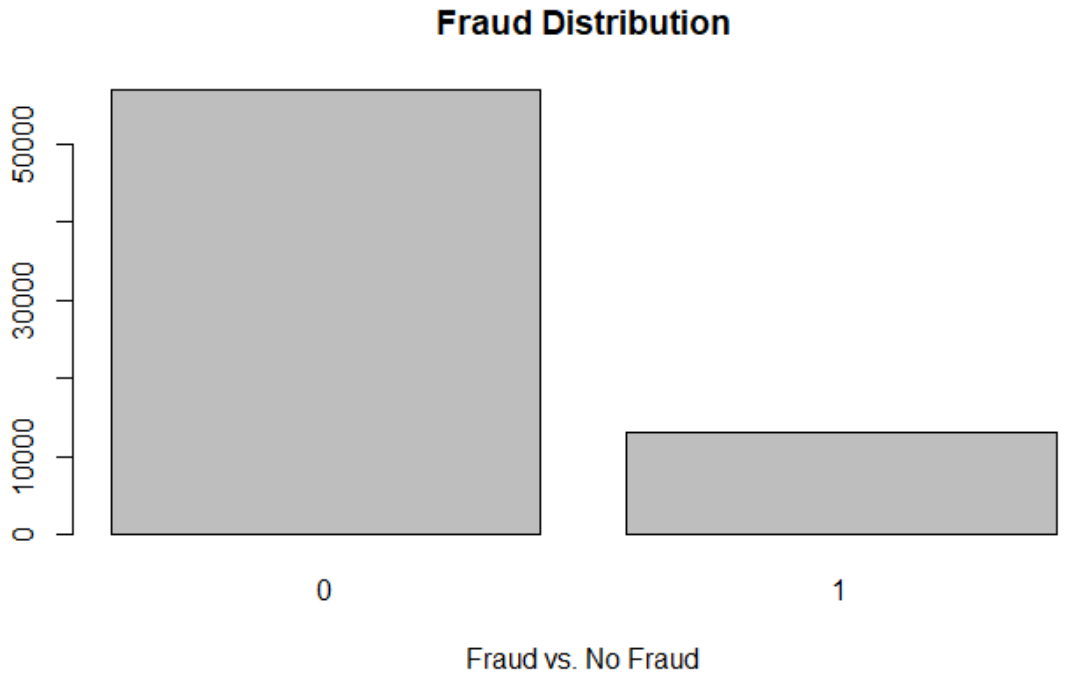
Annex M: Cluster Plot for CLARA Clustering Algortihm with k=3



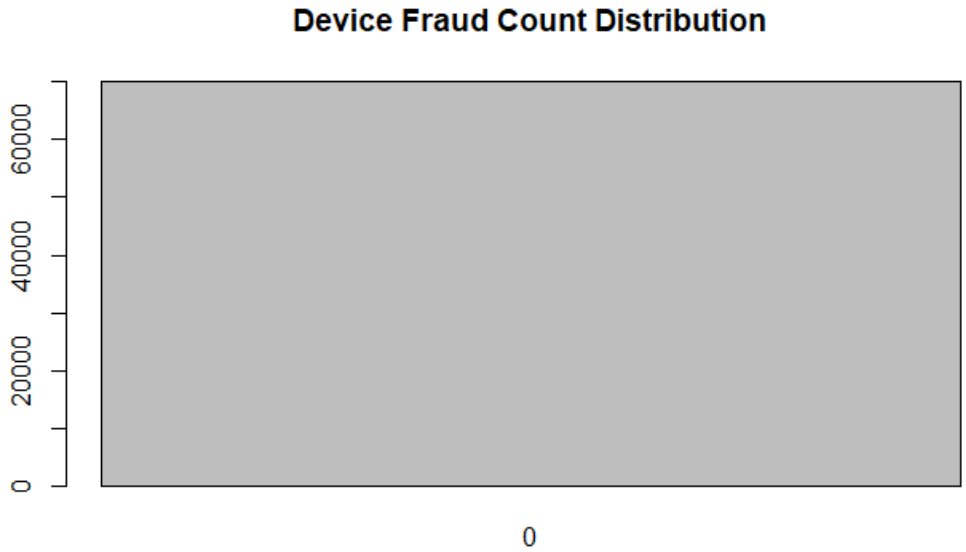
Annex N: Histogram of customer_age outliers



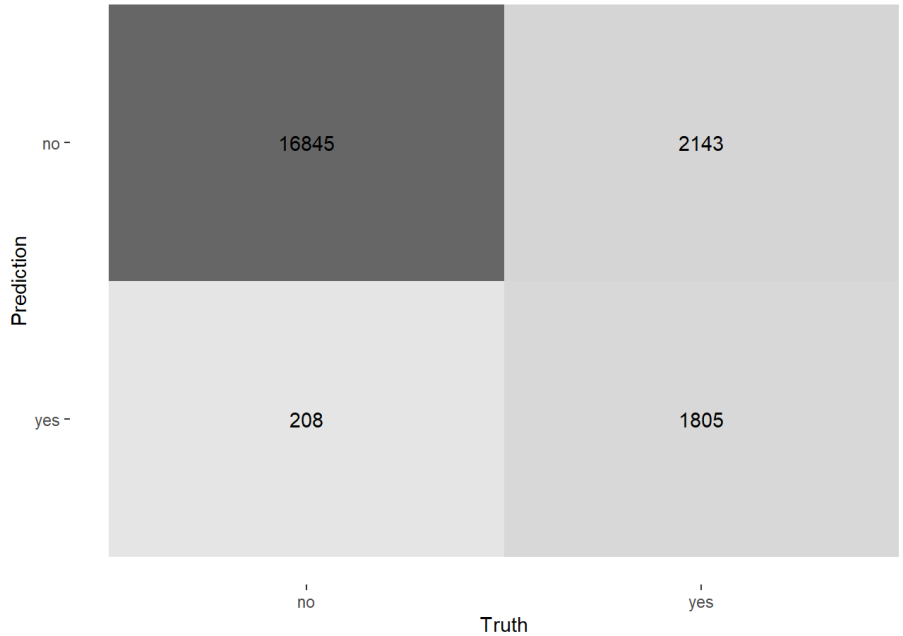
Annex 0: fraud_bool dataset distribution



Annex P: device_fraud_count dataset distribution



Annex Q: Confusion Matrix – Random Forest



Annex Q: ROC Curve – Random Forest