FRAUD DETECTION

STUDI INDEPENDEN MBKM 2022
FINAL PROJECT

SHELLYA NUR ATQIYA
UPDATED ON DECEMBER 2023

SHELLYA NUR ATQIYA

Analytical thinker who loves to get lost in data.

Mathematics, Universitas Pendidikan Indonesia
2019 - 2023
GPA : 3.73/4.00

Experienced as:

- Data Analyst Intern at Cakap
- Assistant Mentor at Data Analysis Bootcamp by DQLab

+6283876583628

shellyanura@gmail.com

linkedin.com/in/shellyanra/

shellyanat.github.io

FRAUD DETECTION

Total Rows

640K++

Total Columns

29

Explanation

The Fraud Detection dataset consist of more than 640k credit account transaction records. Every record includes some detailed information, such as the user info, the transaction date, even the merchant info. It’s also already comes with the fraud label to differentiate a fraud transaction with a non-fraud transaction.

Output Target

FInd the variable that has the most influence to the fraud label (target variable) is the purpose of this project.

linkedin.com/in/shellyanra/

PROJECT BREAKDOWN

01

EXPLORATORY DATA ANALYSIS

02

DATA PREPROCESSING

03

DATA MODELLING

04

CONCLUSION

linkedin.com/in/shellyanra/

01

EXPLORATORY DATA ANALYSIS

MISSING VALUES

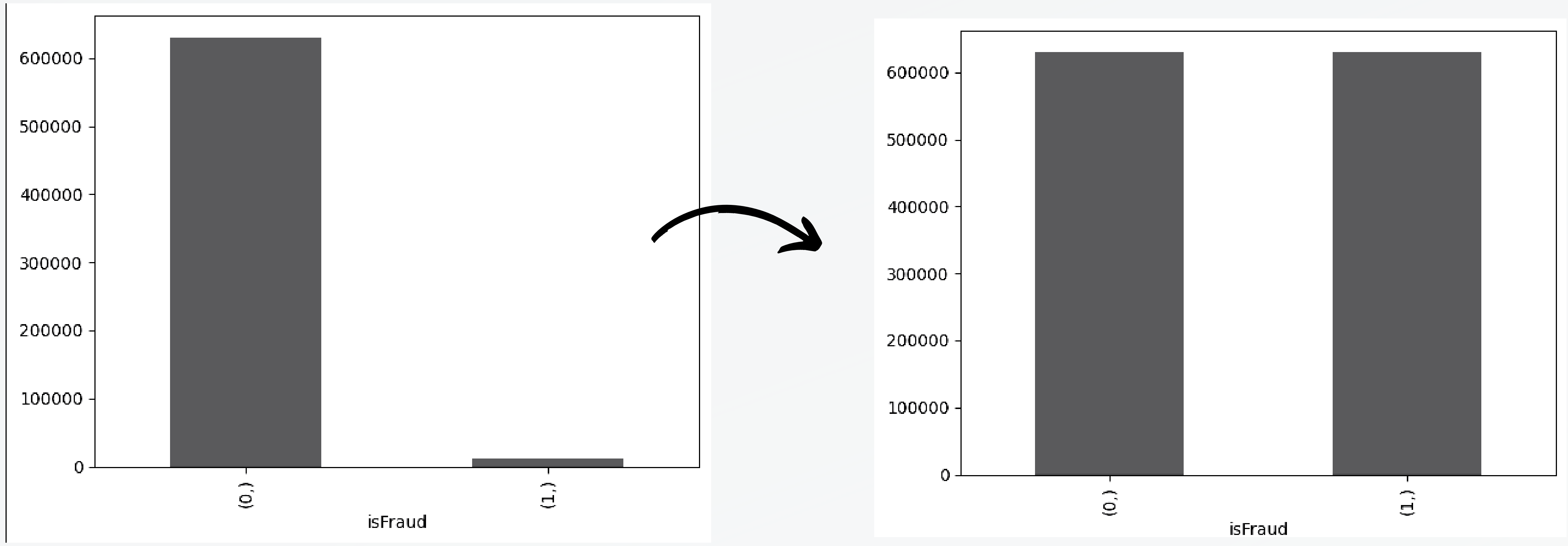
Six columns were removed because all the records are Null. For the other columns that contains Null, it's been calculated that the missing values was below 1% of total records so the null was replaced with the mode of each column.

OUTLIERS

Outliers was seen in numerical records from the transaction such as for creditLimit, availableMoney, transactionAmount, and currentBalance. Since the purpose of this project is to detect the fraud transaction, it's better to keep the outliers since it could be one of the crucial indication of fraud.

linkedin.com/in/shellyanra/

02 DATA PREPROCESSING



ENCODING

New column [’CVVNotMatch’] created to identify the transaction where the card CVV and the entered CVV was not match. In addition, columns with boolean datatypes has been changed to object datatypes so it’s easier for data modelling.

BALANCING DATASET

The target variable, isFraud, has two values: 0 (not a fraud) and 1 (fraud], the number of fraud transaction is very small compared to the non-fraudulent ones. Using SMOTE, the imbalanced fixed in hopes the algorithm would predict better with balanced records between the fraud and non-fraud transaction.

03

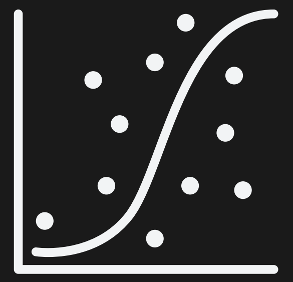
DATA MODELLING

Decision Tree capture non-linear relationships and interactions between variables without any prior assumptions about the data structure.

DECISION TREE

Random Forest combine multiple decision trees, often leading to improve accuracy. This method reduce the risk of overfitting to the training data.

RANDOM FOREST



Logistic Regression model coefficients can be directly interpreted to understand the influence of each predictor variable on the outcome.

LOGISTIC REGRESSION

linkedin.com/in/shellyanra/

MODEL EVALUATION

DECISION TREE

Accuracy	*Specificity*
93.77%	0.923

RANDOM FOREST

Accuracy	*Specificity*
96.09%	0.947

LOGISTIC REGRESSION

Accuracy	*Specificity*
69.52%	0.696

Why Accuracy and Specificity?

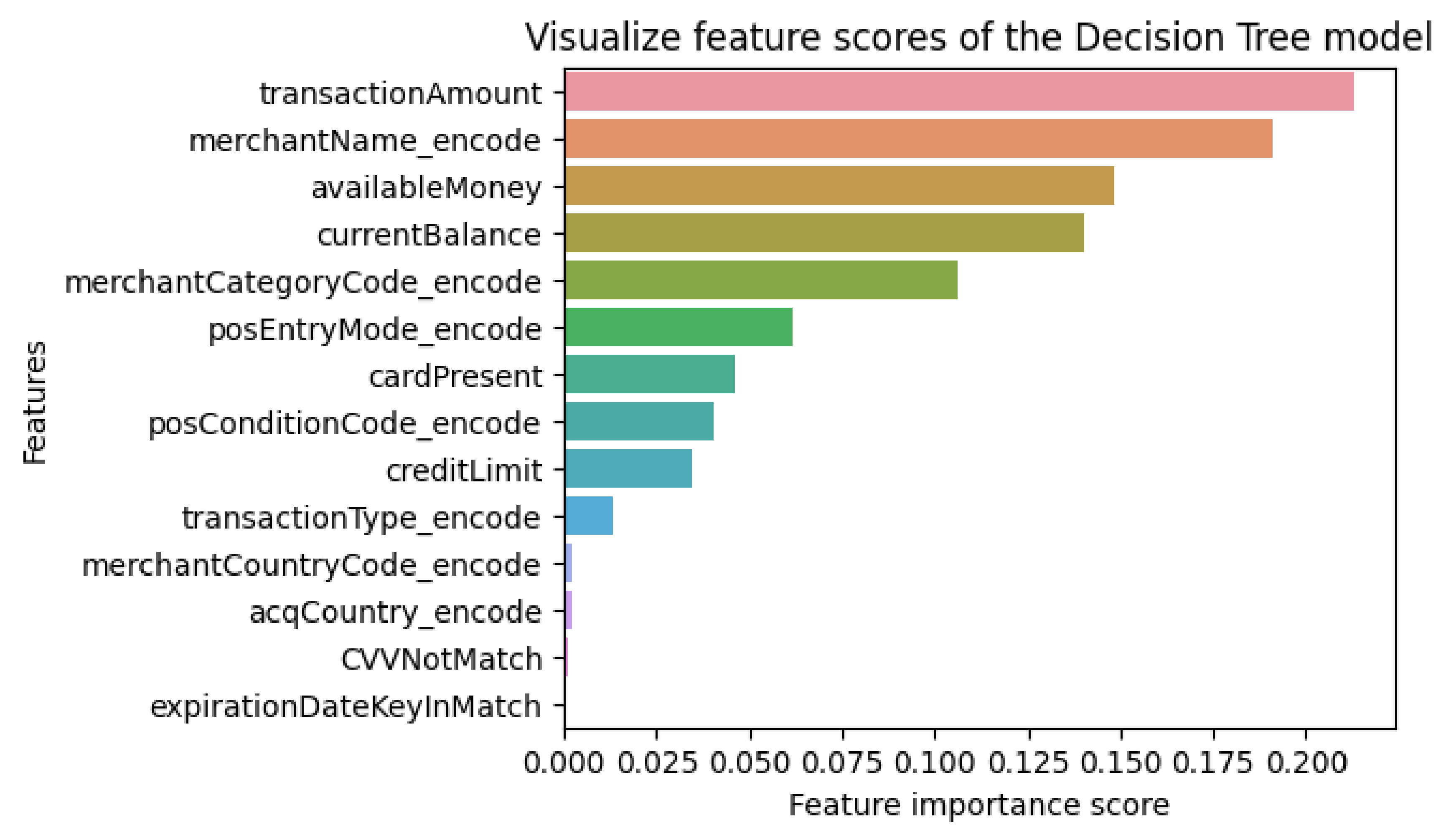
Specificity is important for imbalanced dataset since it’s focused on the accuracy of negative values. In this project, specificity shows us how accurate it is to predict the fraud transaction one.

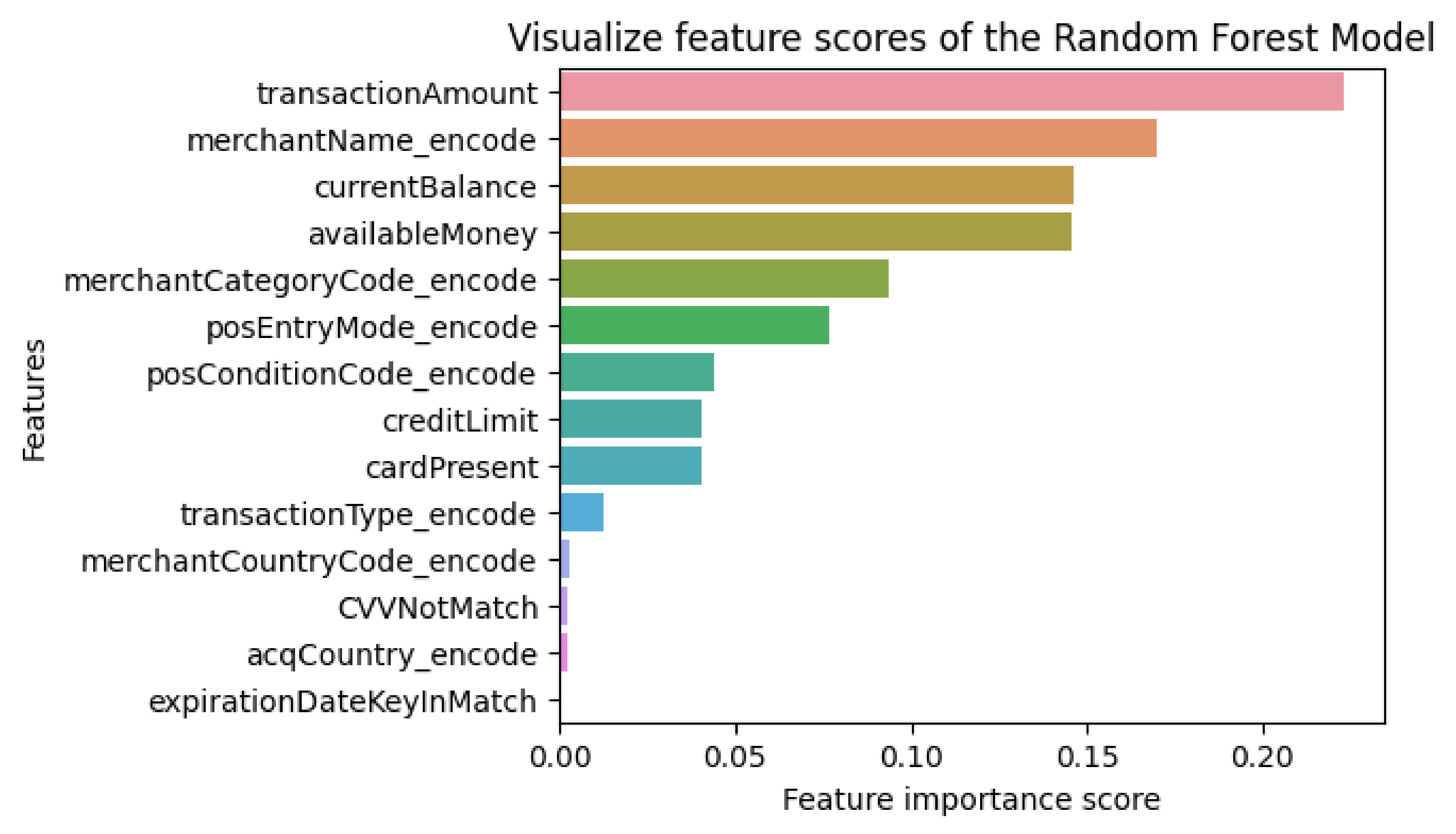
In other side, accuracy score shows us how accurate it is in predicting the fraud label for all of the transaction.

in linkedin.com/in/shellyanra/

FEATURE IMPORTANCE SCORE

DECISION TREE


Visualize feature scores of the Decision Tree model


Visualize feature scores of the Random Forest Model

RANDOM FOREST

linkedin.com/in/shellyanra/

linkedin.com/in/shellyanra/

CONCLUSION

VARIABLE THAT HAS THE MOST INFLUENCE TO THE TARGET VARIABLE

Transaction Amount

The result from decision tree and random forest modelling pointed out that Transaction Amount has the most influence for the target variable, isFraud. This means that transaction amount is a strong predictor of fraud and plays a crucial role in fraud detection models. It could also mean that higher transaction amounts might be more likely to be fraudulent.

THANK YOU!

SHELLYA NUR ATQIYA

+6283876583628

shellyanura@gmail.com

linkedin.com/in/shellyanra/

shellyanat.github.io