CREDIT CARD FRAUD DETECTION

ALMAS FATHIN IRBAH

OVERVIEW

BACKGOUND

According to katadata, credit card transaction activities began to increase by 24.6% in March 2021. Several Islamic banks have begun to deploy Islamic credit cards that can reach a wider business sector for personal or MSMEs.

PROBLEM

While the threat of credit card fraud, especially for online buying and selling activities, we need to anticipate this.

OBJECTIVE

For that, we need to create a kind of system that can detect credit card fraud in the future. In this project, I created several models that we're able to predict credit card fraud and chose the best model. and find the business impact of the insight analysis that I did.

DATA INTRODUCTION

Source : Kaggle - **Credit Card Transactions Fraud Detection Dataset**
https://www.kaggle.com/kartik2112/fraud-detection

About the Dataset

- This is a simulated credit card transaction dataset containing legitimate and fraud transactions from the duration **1 Jan 2019 – 31 Dec 2020**.
- It covers credit cards of **1000 customers** doing transactions with a pool of **800 merchants**.
- **1.852.394 rows** & **22 columns**.

Data Related to Customer Account Information

- first name, last name, date of birth, trans date trans time, cc number, amount, trans number, unix time

Data Related to Customer Demographic Information

- gender, street, city, state, zip, lat, long, city pop, job

Data Related to Merchant Information

- merchant, category, merchant latitude, merchant longitude

Data Related to Fraud Information

- Is fraud

Data Preparation

Missing values(%): 0.0%

Address

- Create function to calculate the distance between two address
- Concanate the lat and longitude of client into one column and same for the merchant location

Residence

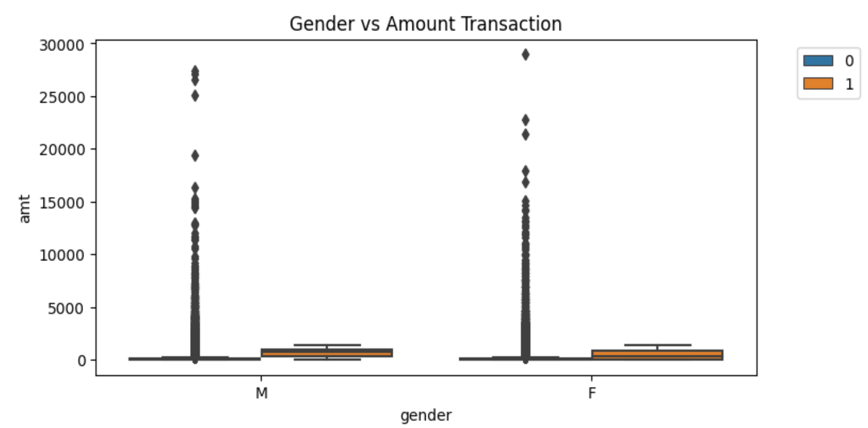
- Create the column where the if the population is less than 25 % to be rural, 25-50% semi urban, and more than 50% urban

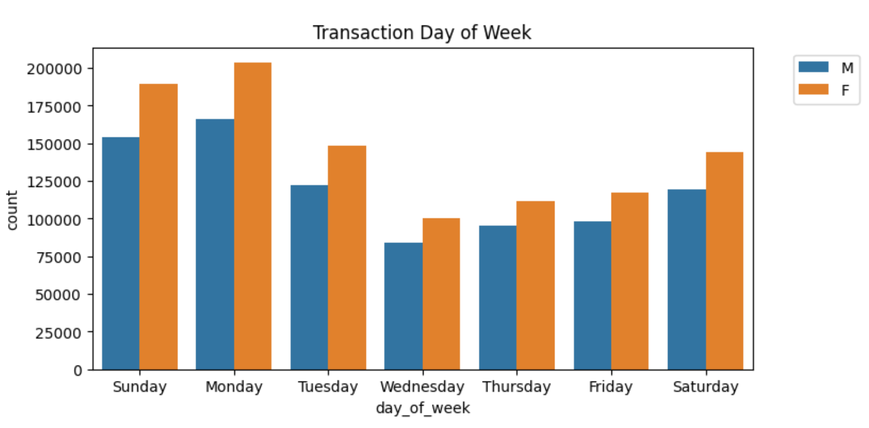
Time

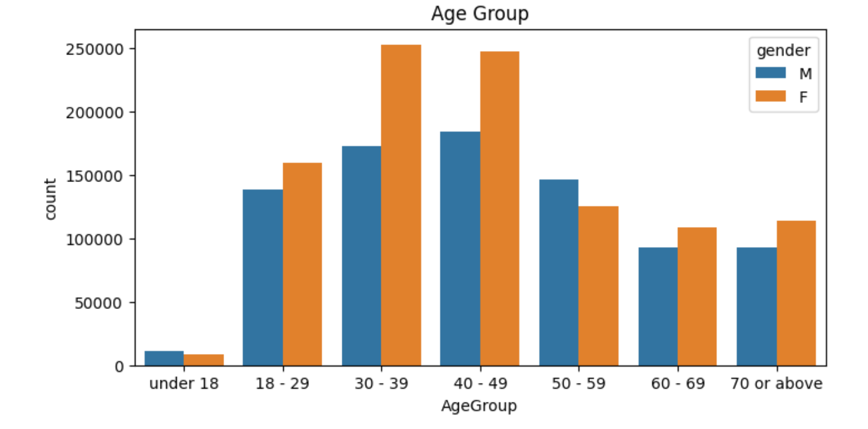
- Get hours from the transaction
- Get days when the transaction occured
- Get the age of the customer when the transaction occurred
- Create age group

Exploratory Data Analysis

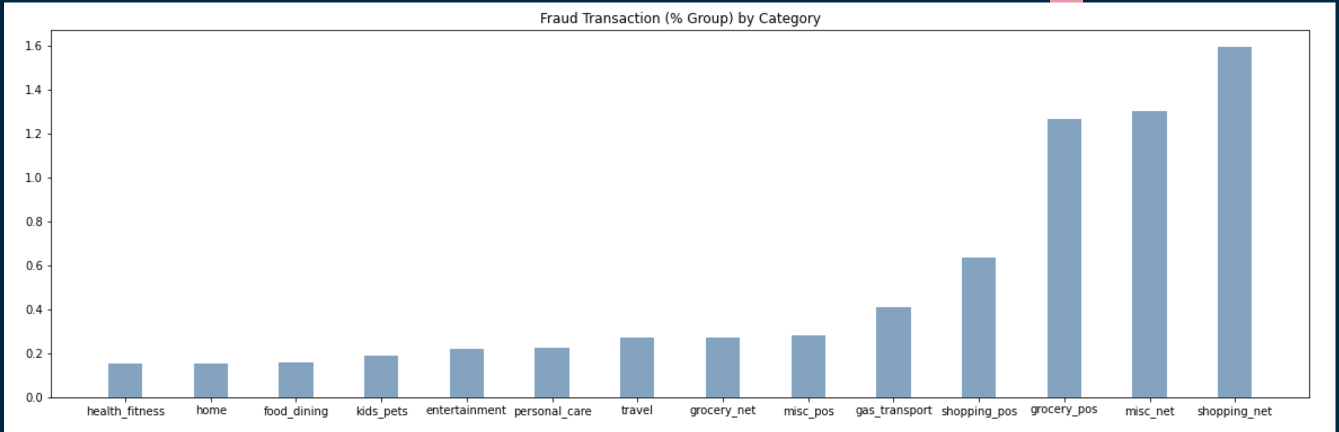
	Male	Female	
Fraud	4.752 (**0,256 %**)	4.899 (**0,264 %**)	9.651 (**0,521 %**)
Non Fraud	832.893 (44,963 %)	1.009.850 (54,515 %)	1.842.743 (99,478 %)
	837.645 (**45,219 %**)	1.014.749 (**54,78%**)	1.852.394 (100 %)

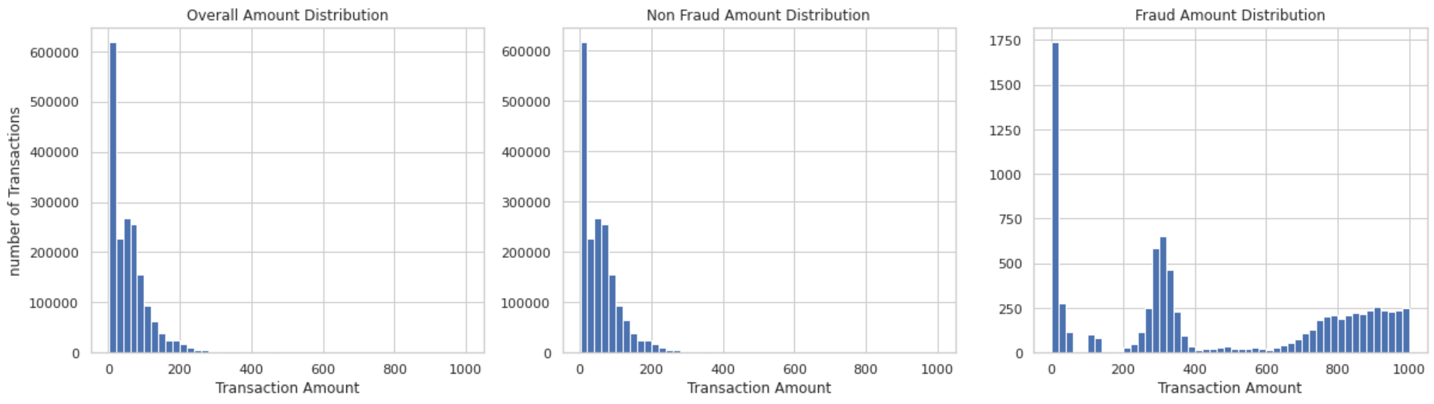






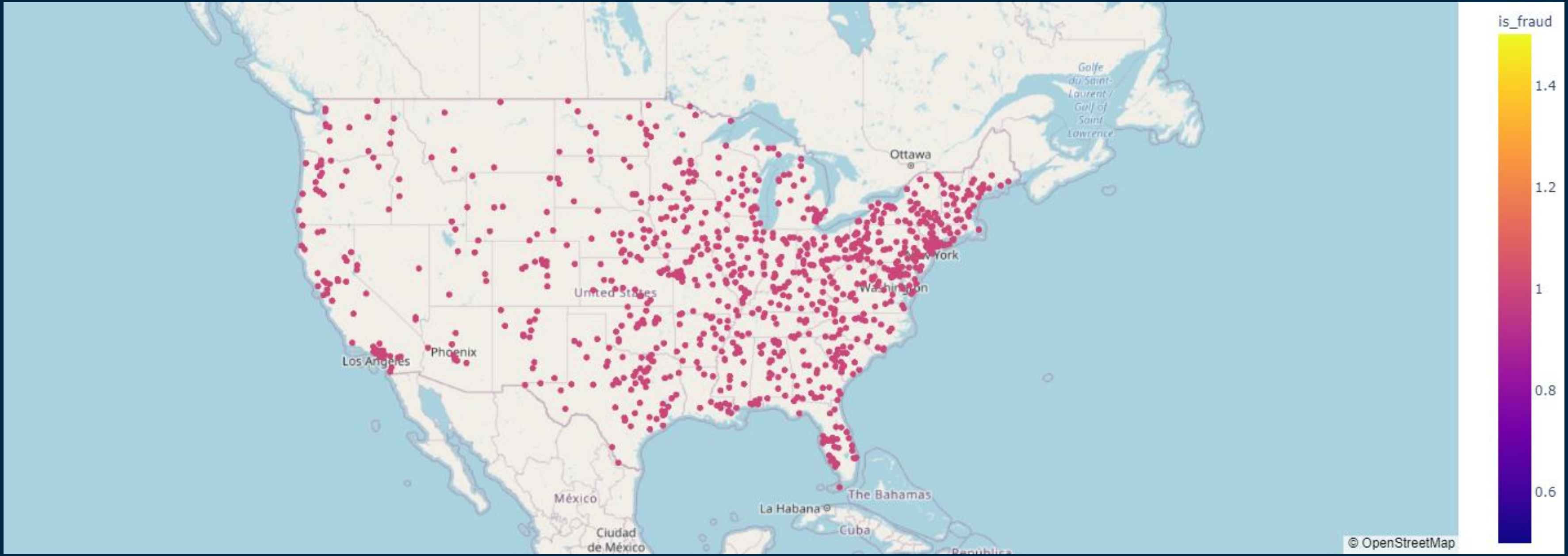
Exploratory Data Analysis





Exploratory Data Analysis

Transaction Distribution



MODELLING : DATA PREPARATION

- Seperating nominal from numeric (drop 20 columns).
- There are almost 2 million records in dataframe.
- In order to avoid the heavy calculation,only the first 500.000 rows were selected.

	category	amt	gender	unix_time	is_fraud	trans_hour	day_of_week	age	residence	distance
0	personal_care	2.86	M	1371816865	0	12	Sunday	52.0	urban	24.546041
1	personal_care	29.84	F	1371816873	0	12	Sunday	30.0	rural	104.859216
2	health_fitness	41.28	F	1371816893	0	12	Sunday	50.0	urban	59.042985
3	misc_pos	60.05	M	1371816915	0	12	Sunday	33.0	urban	27.681177
4	travel	3.19	M	1371816917	0	12	Sunday	65.0	semi_urban	104.269600
...	...	...	...	...	...	...	...	...	...	...
499995	gas_transport	76.65	M	1387500745	0	0	Sunday	40.0	urban	126.542495
499996	grocery_pos	66.25	M	1387500745	0	0	Sunday	44.0	urban	81.341516
499997	food_dining	72.42	M	1387500746	0	0	Sunday	92.0	semi_urban	44.997590
499998	shopping_pos	3.01	M	1387500791	0	0	Sunday	82.0	rural	96.965803
499999	grocery_net	35.39	F	1387500799	0	0	Sunday	23.0	urban	91.734236

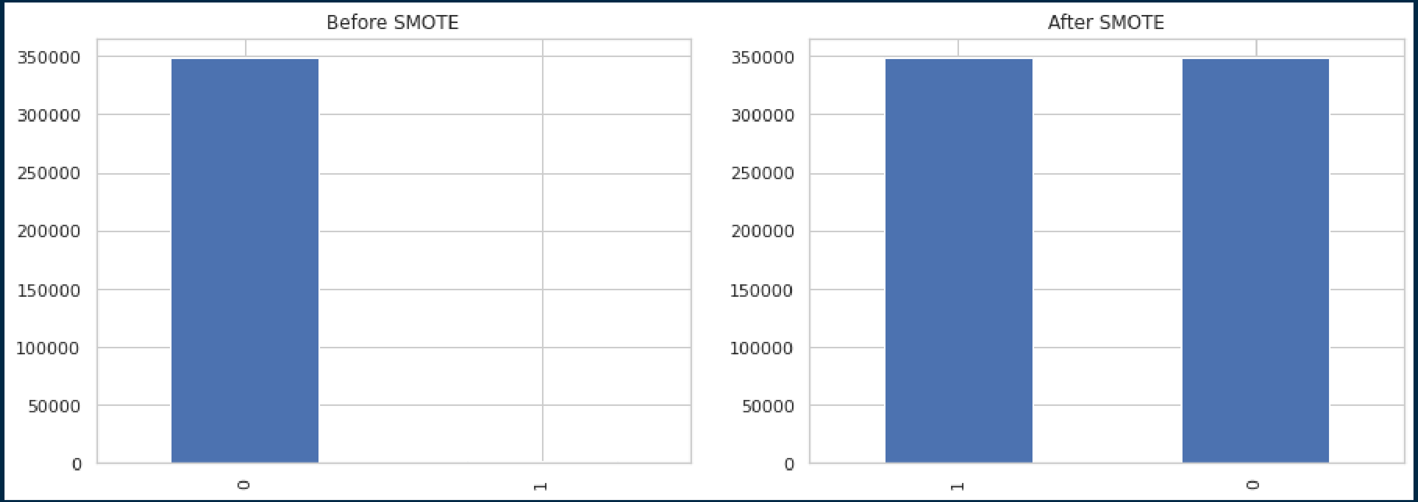
500000 rows × 10 columns

MODELLING : DATA PREPARATION

- Creating a dummy variable for one of the categorical variables ('category', ’day of week', 'gender', 'residence’) and drop the first ones.
- Adding the results to the master dataframe.
- Dropping the repeated variables.
- Since we have a huge amount of data, its better to normalize the dataset by using RobustScaler which scales the data according to the quantile range.
- Train data size is 80% of observation and Test data size is 20% of observation.

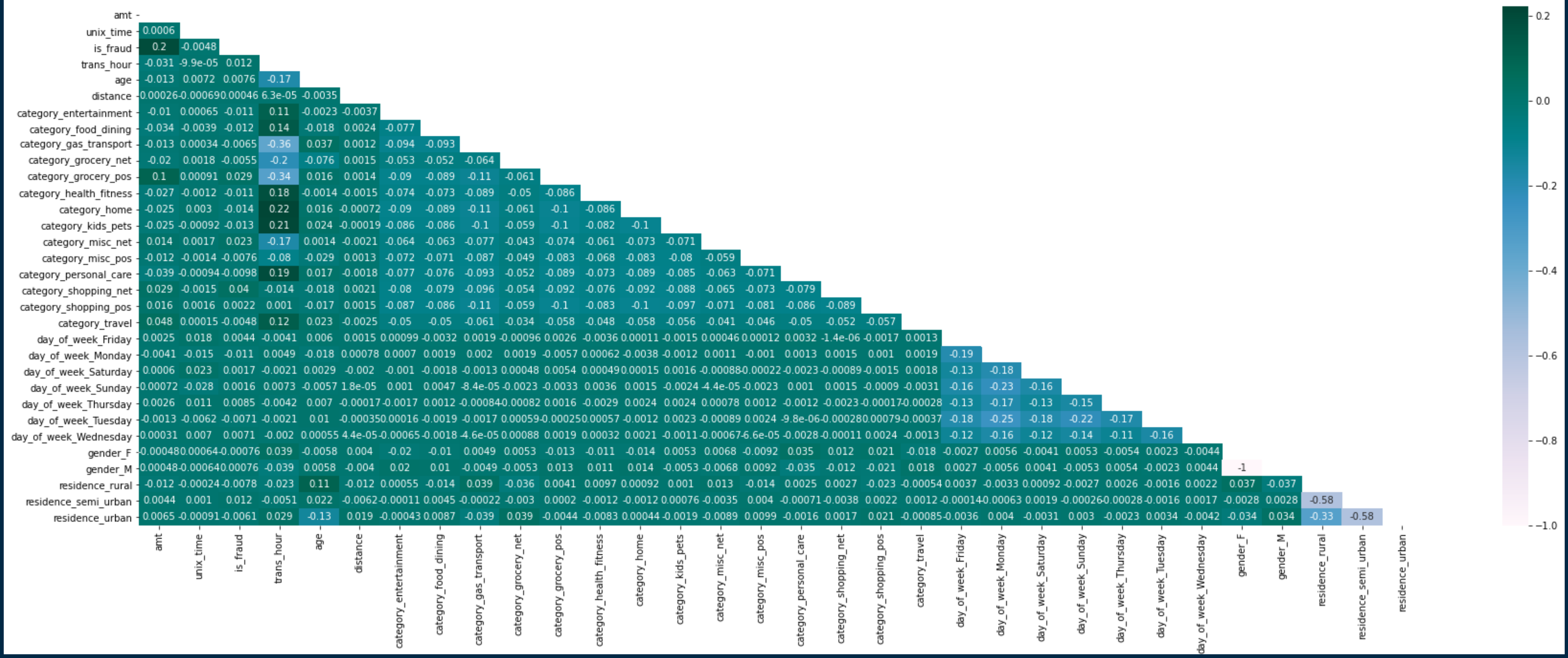
MODELLING : DATA PREPARATION

The dataset is heavily imbalanced. Through resampling, fraud transactions (Class = 1) are randomly increased to the same amount as non-fraud transactions (Class = 0) in order to avoid the bias results toward the non-fraudulent class.



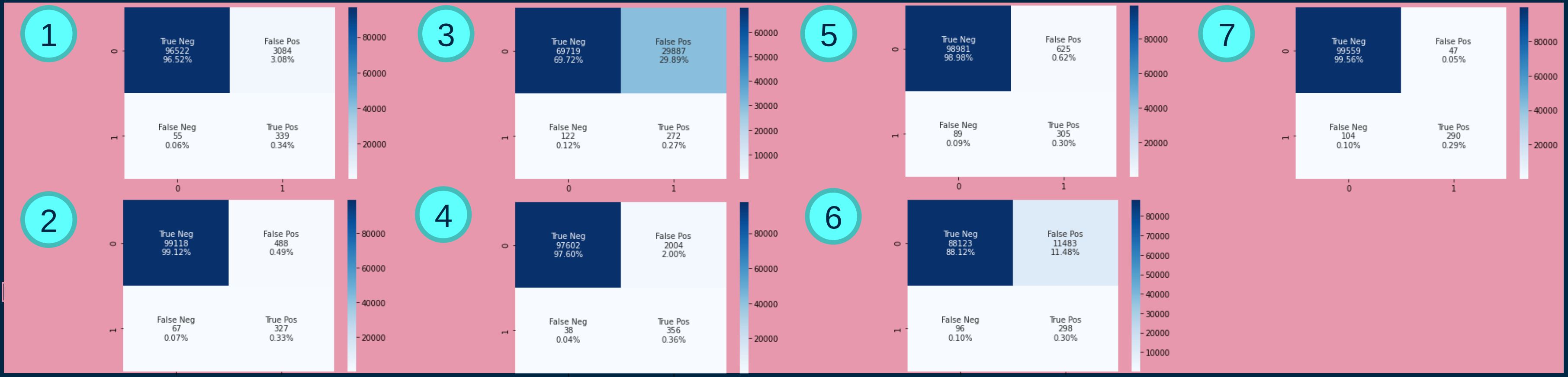
MODELLING : DATA PREPARATION

Let's see the correlation matrix

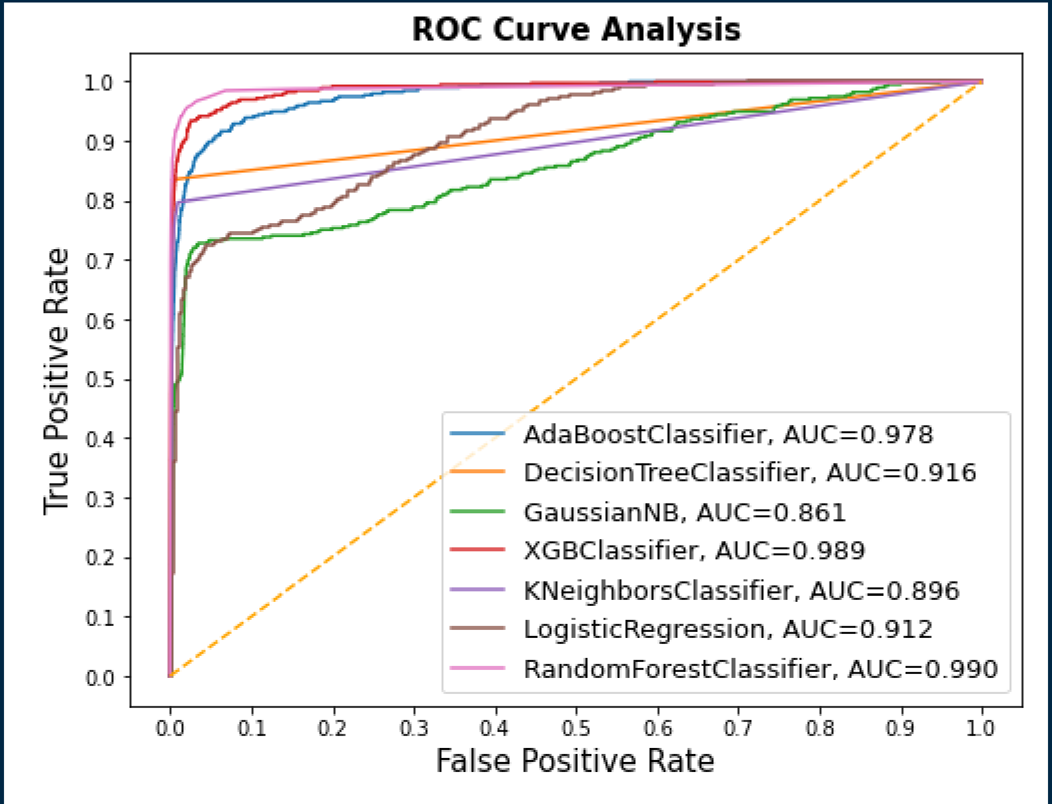


MODELLING : TRAINING-EVALUATION

	Method	Test								
		Precision		Recall		F1 Score		Accuracy	MAE	RMSE
		Fraud	Non-Fraud	Fraud	Non-Fraud	Fraud	Non-Fraud			
1	Ada Boost Clasifier	10%	100%	86%	97%	18%	98%	97%	0,031	0,176
2	Decision Tree Clasifier	40%	100%	83%	100%	54%	100%	99%	0,005	0,074
3	Gaussian NB	1%	100%	69%	70%	2%	82%	70%	0,300	0,547
4	XGB Classifier	15%	100%	90%	98%	26%	99%	98%	0,020	0,142
5	K Neighbor Classifier	54%	100%	68%	100%	60%	100%	100%	0,003	0,060
6	Logistic Regression	3%	100%	76%	88%	5%	94%	88%	0,116	0,341
7	Random Forest Classifier	86%	100%	74%	100%	79%	100%	100%	0,001	0,039



MODELLING : TRAINING-EVALUATION



MODELLING : HYPERPARAMETER TUNING

Grid Search CV & Randomized Search CV:

Estimator;
Random Forest
(N estimator = 100 & random state = 42)

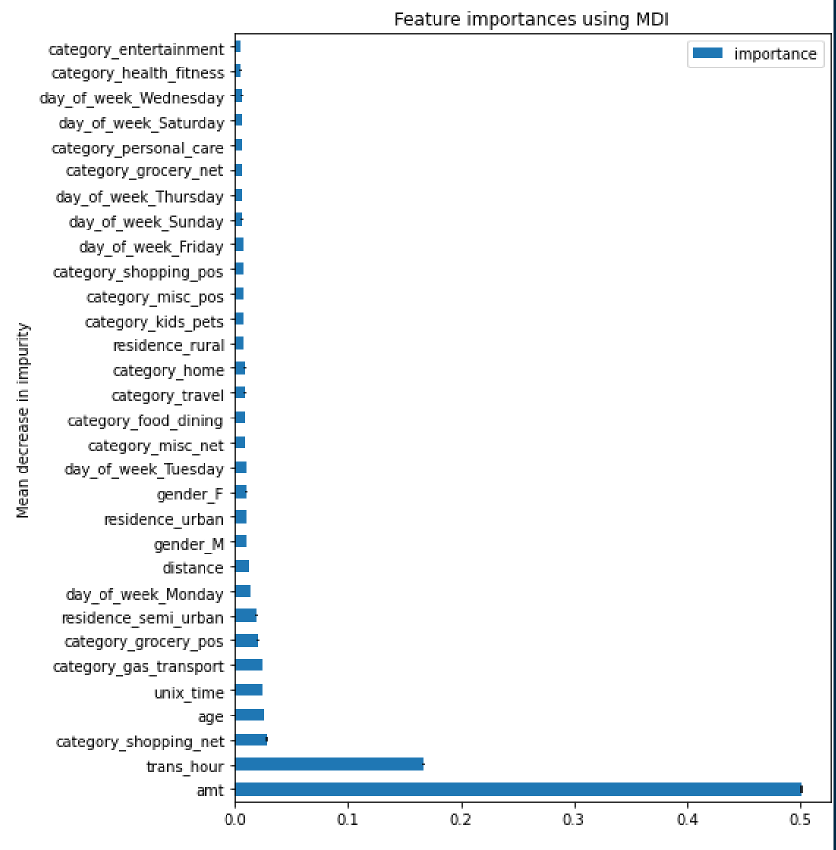
Parameter ;
N estimator: [500]
Max features: [sqrt]
Min samples split: [2]
Bootstrap: [False]

CV = 3
Verbose = 2
N jobs = -1

	Random Forest (N estimator = 100 & random state = 42)		**Randomized Search CV**(Max depth = 80 & Min samples leaf =15)		**Grid Search CV** (Max depth = 100 & Min samples leaf =30)	
	Fraud	Non Fraud	Fraud	Non Fraud	Fraud	Non Fraud
Precision	86%	100%	68%	100%	58%	100%
Recall	74%	100%	80%	100%	83%	100%
F1-score	79%	100%	73%	100%	69%	100%
Accuracy	100%		100%		100%	
RMSE	0,001		0,002		0,002	
MAE	0,038		0,047		0,054	

Fitting Duration: Grid Search CV = 44,6 min & Randomized Search CV = 46,8 min.

MODELLING : FEATURE IMPORTANT



INSIGHT & RECOMMENDATION

Here's what you'll get insight in this **project**:

1. The **recommended machine learning model** for detecting fraudulent transactions is the **random forest** because it has the best F1 score, RMSE, MAE, and ROC curve analysis.

2. **Hyperparameter tuning** in the random forest **does not affect** to increase the F1 score.

3. The **2 highest feature importances** from random forest are **amount transaction** and **transation hour**.

$ 530

Average amount per fraud transaction

402

Average number of fraudulent transaction per month

77,183

Average number of transactions per month

Cost Benefit Analysis

Do you have any questions?

almasfathinirbah@gmail.com
~
almasfathinirbah.github.io

THANKS

in Almas Fathin Irbah