SAMSUNG

Samsung Innovation Campus

Artificial Intelligence Course

Together for Tomorrow!
Enabling People
Education for Future Generations

Credit Card Fraud Detection

Presenters:

DALIA HESHAM
Mechatronics Engineering Student

MANAR KANDIL
Student at Dep. of Chemistry, Faculty of Science

YASSMEN YOUSSEF
Electronics & Communication Engineer

Group Facilitator: Eng. Haneen

Kaggle: https://www.kaggle.com/code/yassmenyoussef/semi-project-credit-card-detection
GitHub: https://github.com/Dalia-Hesham/Credit-Card-Fraud-Detection

Supervised By:

Dr. Doaa Mahmoud

Agenda

1.
2.

The Problem
Objective
Dataset Overview
Data Exploration
Exploratory Data Analysis
Data Preprocessing
Data Sampling
Machine Learning Modeling
Models Evaluation
Business Solutions
and Recommendations
Find us

The Problem

€1.8 billion

Card Fraud Losses in 2018 in SEPA

10 %

American Victims of Credit Card Fraud

Worldwide financial losses caused by credit card fraudulent activities are worth tens of billions of dollars.

Sources: Statistic Brain Research Institute. Credit card fraud statistics April 2018, European Central Bank. 6th report on card fraud August 2020

Problems in Egypt

01 Reduced adoption of digital payment methods which inhibits attainment of financial inclusion.

02 Losses in sectors that rely extensively on e-payments, such as e-commerce.

03 BNPL companies suffer from having a shortage of data inputs for their proprietary algorithms.

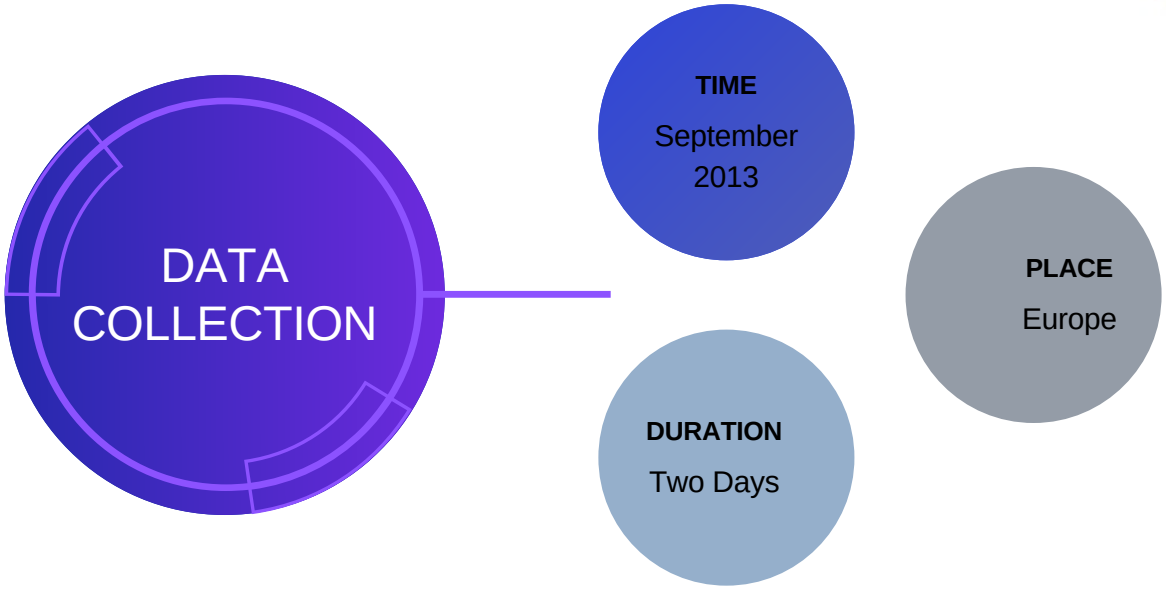
Objectives

Increase customer confidence in using credit cards.

Assist the banks in reaching out to more customers

Promote BNPL internal credit scoring system

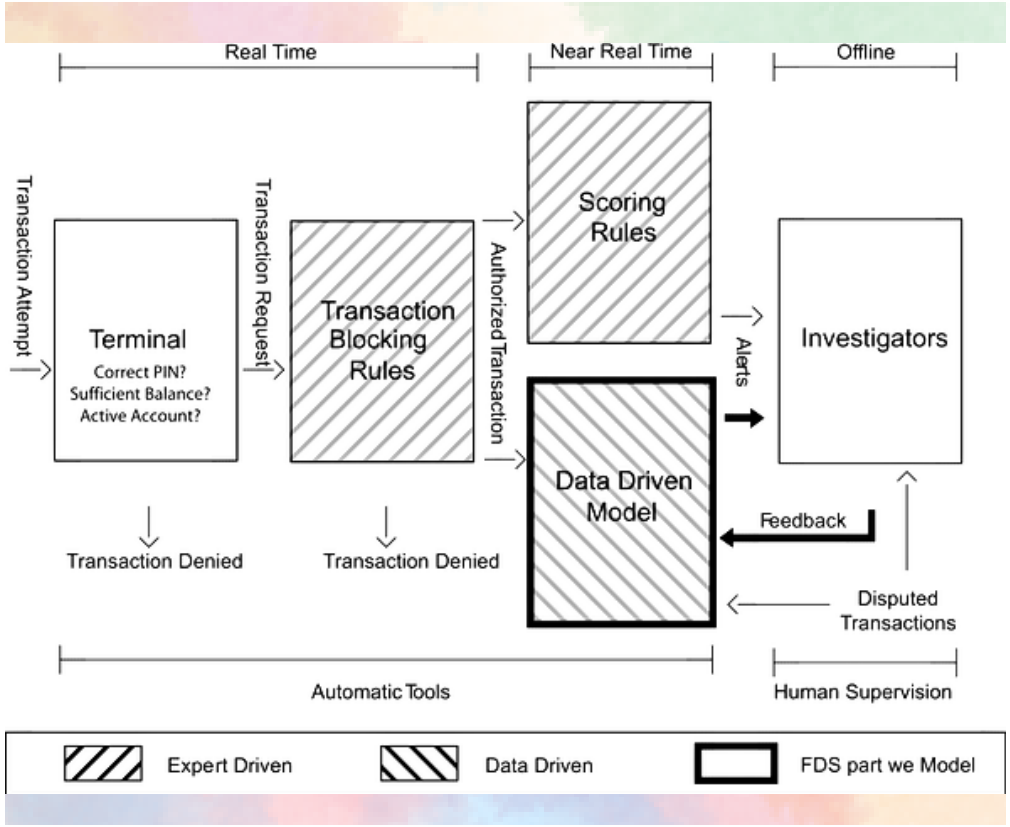
Data Information



What is considered as a fraud

Data Source: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Fraud Detection System (FDS)



(FDS) requires finding, out of millions of daily transactions, which ones are fraudulent. Due to the ever-increasing amount of data.

Terms

Transaction scenarios:

1. Card-present transactions

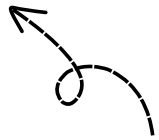
(CP) scenarios, refer to scenarios where a physical card is needed, such as transactions at a store (also referred to as a point-of-sale - POS) or transactions at a cashpoint (for instance at an ATM).

2. Card-not-present transactions

(CNP) scenarios, refers to scenarios where a physical card does not need to be used, which encompasses payments performed on the Internet, by phone, or by mail.

Data Information

```
 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
```



Our Data contains only numerical input variables which are the result of a PCA transformation.

Any Missing Data?

```
Time      0
V1        0
V2        0
V3        0
V4        0
V5        0
V6        0
V7        0
V8        0
V9        0
V10       0
V11       0
V12       0
V13       0
V14       0
V15       0
V16       0
V17       0
V18       0
V19       0
V20       0
V21       0
V22       0
V23       0
V24       0
V25       0
V26       0
V27       0
V28       0
Amount    0
Class     0
dtype: int64
```

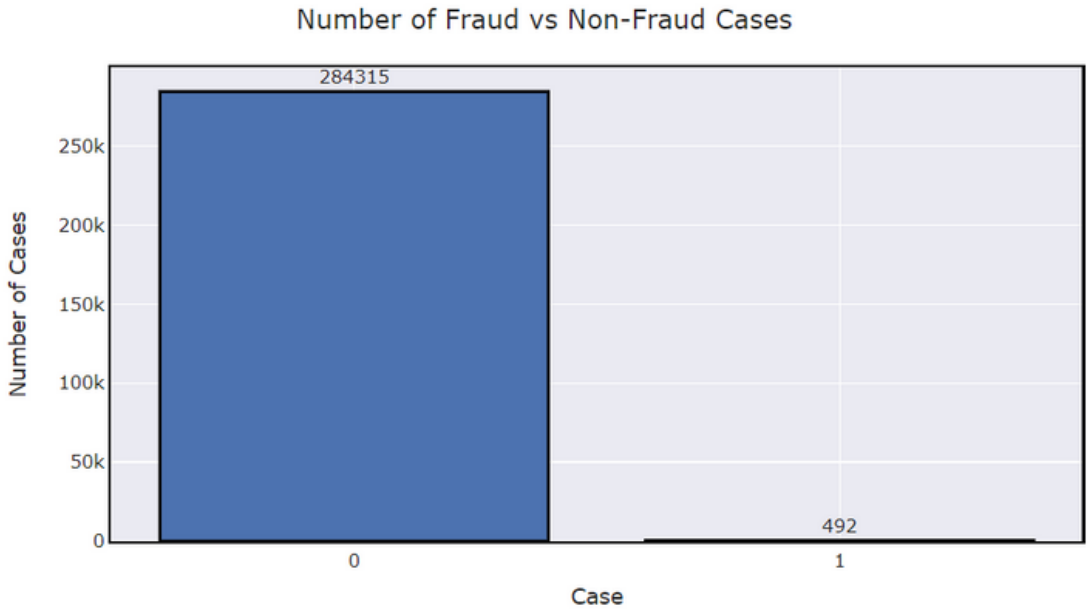
Data Description

	count	mean	std	min	25%	50%	75%	max
Time	284807.0	9.481386e+04	47488.145955	0.000000	54201.500000	84692.000000	139320.500000	172792.000000
V1	284807.0	3.918649e-15	1.958696	-56.407510	-0.920373	0.018109	1.315642	2.454930
V2	284807.0	5.682686e-16	1.651309	-72.715728	-0.598550	0.065486	0.803724	22.057729
V3	284807.0	-8.761736e-15	1.516255	-48.325589	-0.890365	0.179846	1.027196	9.382558
V4	284807.0	2.811118e-15	1.415869	-5.683171	-0.848640	-0.019847	0.743341	16.875344
V5	284807.0	-1.552103e-15	1.380247	-113.743307	-0.691597	-0.054336	0.611926	34.801666
V6	284807.0	2.040130e-15	1.332271	-26.160506	-0.768296	-0.274187	0.398565	73.301626
V7	284807.0	-1.698953e-15	1.237094	-43.557242	-0.554076	0.040103	0.570436	120.589494
V8	284807.0	-1.893285e-16	1.194353	-73.216718	-0.208630	0.022358	0.327346	20.007208
V9	284807.0	-3.147640e-15	1.098632	-13.434066	-0.643098	-0.051429	0.597139	15.594995
V10	284807.0	1.772925e-15	1.088850	-24.588262	-0.535426	-0.092917	0.453923	23.745136
V11	284807.0	9.289524e-16	1.020713	-4.797473	-0.762494	-0.032757	0.739593	12.018913
V12	284807.0	-1.803266e-15	0.999201	-18.683715	-0.405571	0.140033	0.618238	7.848392
V13	284807.0	1.674888e-15	0.995274	-5.791881	-0.648539	-0.013568	0.662505	7.126883
V14	284807.0	1.475621e-15	0.958596	-19.214325	-0.425574	0.050601	0.493150	10.526766
V15	284807.0	3.501098e-15	0.915316	-4.498945	-0.582884	0.048072	0.648821	8.877742
V16	284807.0	1.392460e-15	0.876253	-14.129855	-0.468037	0.066413	0.523296	17.315112
V17	284807.0	-7.466538e-16	0.849337	-25.162799	-0.483748	-0.065676	0.399675	9.253526
V18	284807.0	4.258754e-16	0.838176	-9.498746	-0.498850	-0.003636	0.500807	5.041069
V19	284807.0	9.019919e-16	0.814041	-7.213527	-0.456299	0.003735	0.458949	5.591971
V20	284807.0	5.126845e-16	0.770925	-54.497720	-0.211721	-0.062481	0.133041	39.420904
V21	284807.0	1.473120e-16	0.734524	-34.830382	-0.228395	-0.029450	0.186377	27.202839
V22	284807.0	8.042109e-16	0.725702	-10.933144	-0.542350	0.006782	0.528554	10.503090
V23	284807.0	5.282512e-16	0.624460	-44.807735	-0.161846	-0.011193	0.147642	22.528412
V24	284807.0	4.456271e-15	0.605647	-2.836627	-0.354586	0.040976	0.439527	4.584549
V25	284807.0	1.426896e-15	0.521278	-10.295397	-0.317145	0.016594	0.350716	7.519589
V26	284807.0	1.701640e-15	0.482227	-2.604551	-0.326984	-0.052139	0.240952	3.517346
V27	284807.0	-3.662252e-16	0.403632	-22.565679	-0.070840	0.001342	0.091045	31.612198
V28	284807.0	-1.217809e-16	0.330083	-15.430084	-0.052960	0.011244	0.078280	33.847808
Amount	284807.0	8.834962e+01	250.120109	0.000000	5.600000	22.000000	77.165000	25691.160000
Class	284807.0	1.727486e-03	0.041527	0.000000	0.000000	0.000000	0.000000	1.000000

Features V1, V2, … V28 are the principal components obtained with PCA. the only features which have not been transformed with PCA are 'Time' and 'Amount'.

Exploratory Data Analysis

- Is our data balanced?

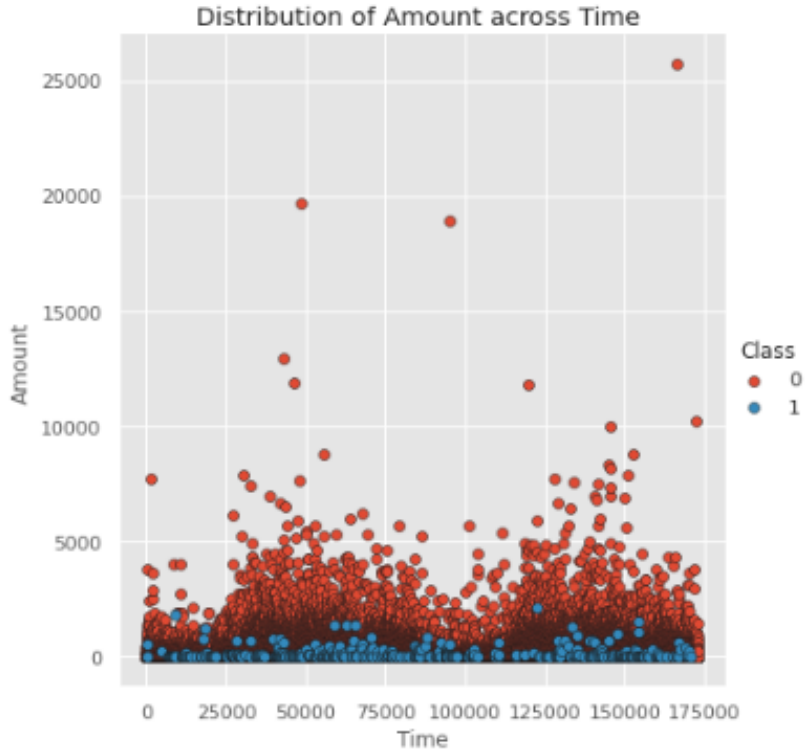


99.83%
No Fraud

0.17%
Frauded

EDA

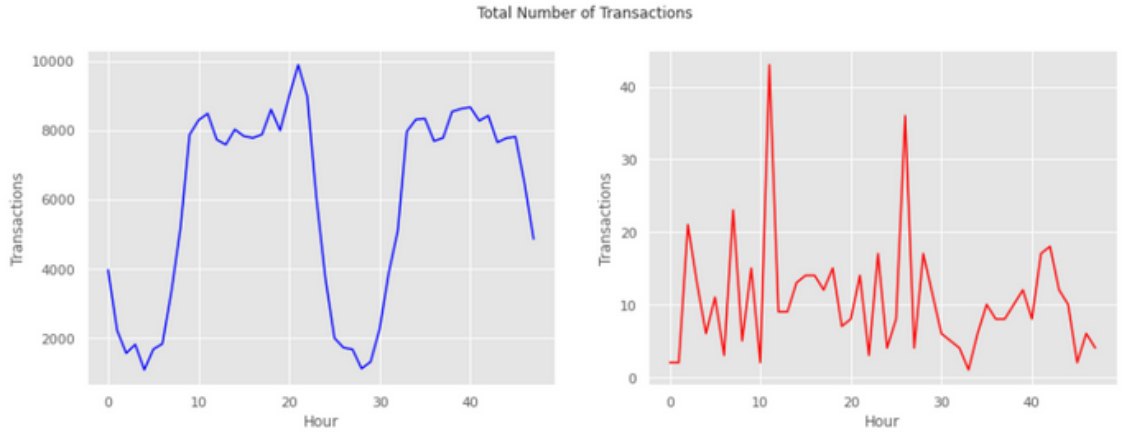
Money Amount VS. Transactions Time



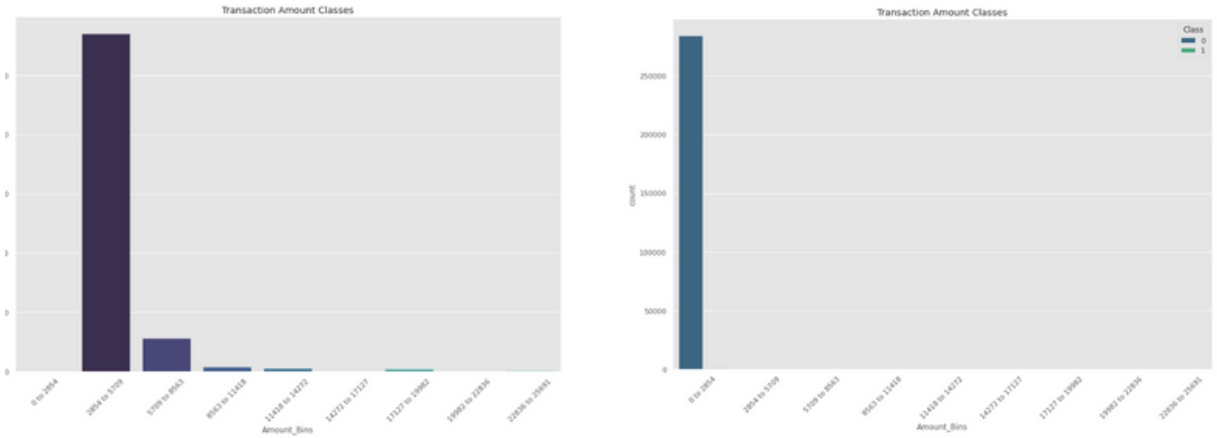
Frauds only on the transactions which have transaction amount less than 2500 (approx.)

EDA

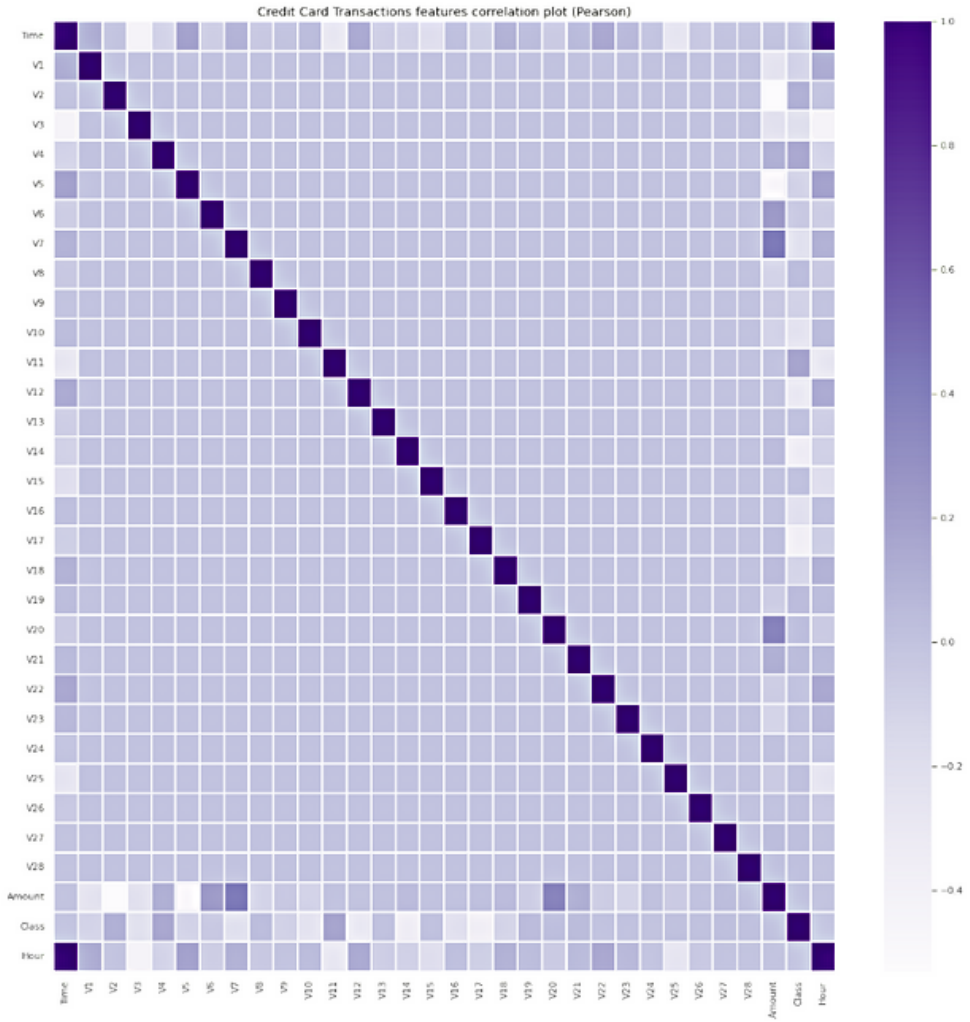
Distribution of the data:



Withdrawal amount with time:



Correlation Correlation



- There are certain correlations between some of these features and Time (inverse correlation with V3) and Amount (direct correlation with V7 and V20, inverse correlation with V2 and V5).

EDA Conclusion

We concluded that we have 2 classes:

Class A

People who deal with large amount of transaction

Class B

People who deal with small amount of transaction

Preprocessing

- Since, the dataset contains no null or missing values and all the features seem to be scaled properly, there is no need of any preprocessing.

- Only the features Time & Amount are required to be scaled.

Metrics Of Supporting Dataset

Model Name	Train Accuracy	Test Accuracy	Precision	Recall	F1 score
KNeighborsClassifier	0.51	0.91	0.10	0.02	0.04
DecisionTreeClassifier	0.59	0.54	0.10	0.61	0.18
LogisticRegression	0.58	0.53	0.10	0.63	0.18

This dataset metrics were too bad to model on it

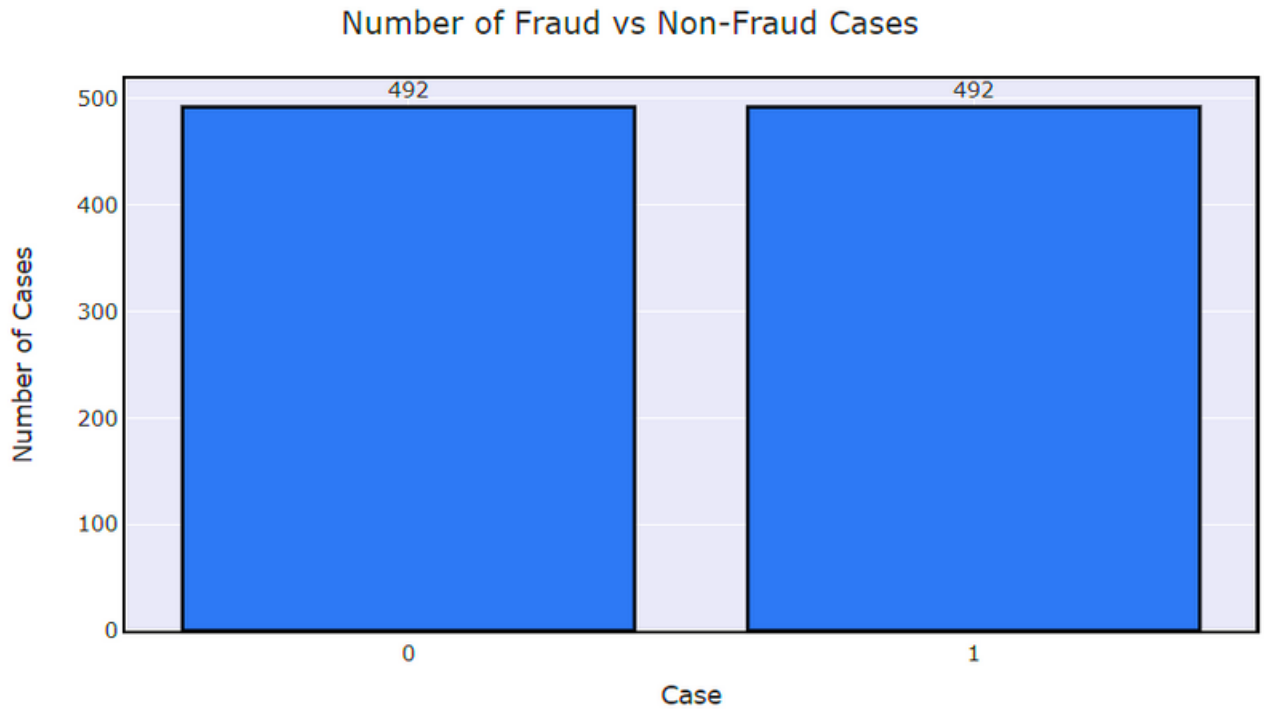
Imbalanced Data

How Can We Deal With This?

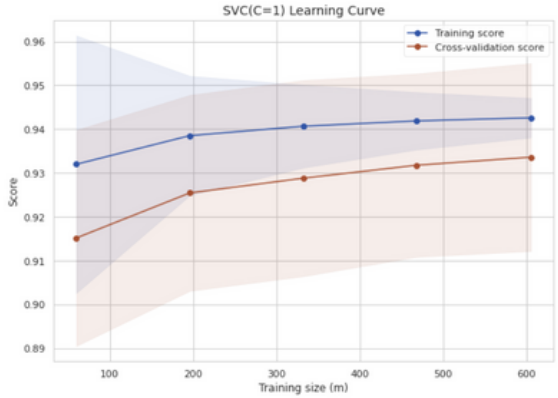
Should We Collect More Data???

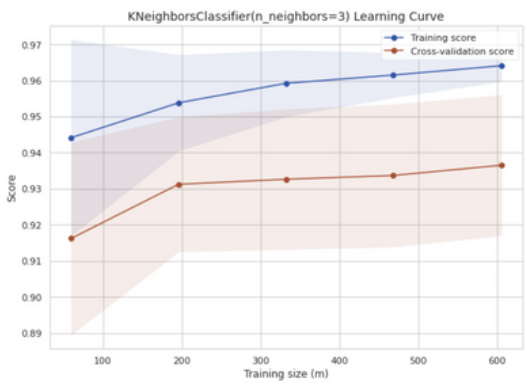
Undersampling

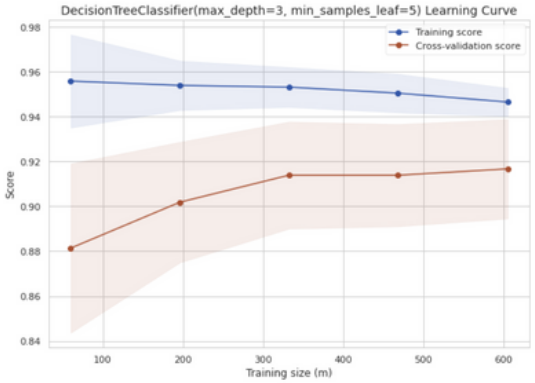
Introducing Balanced Data Sample:

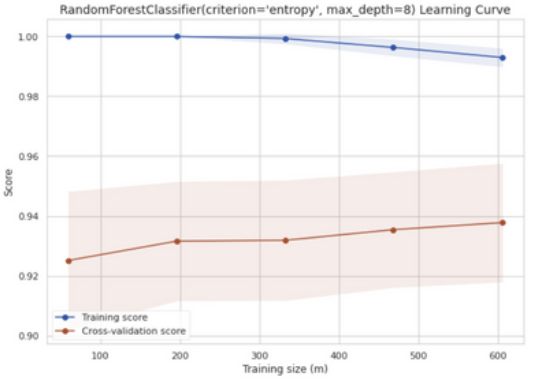


Learning Curves



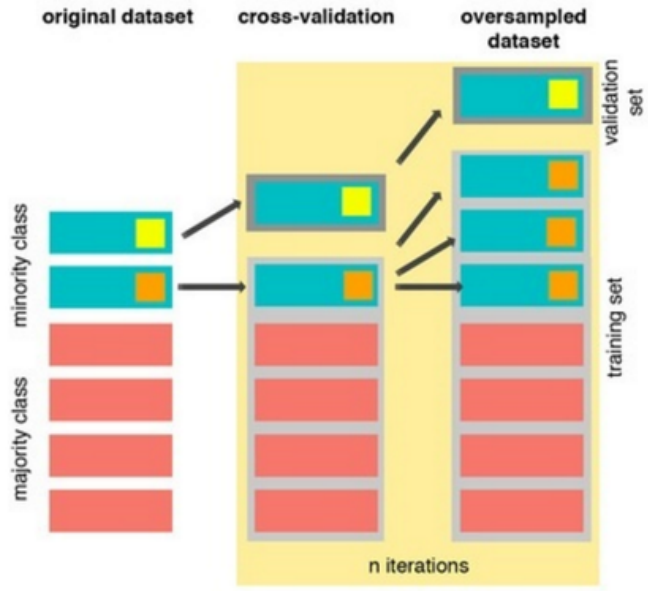




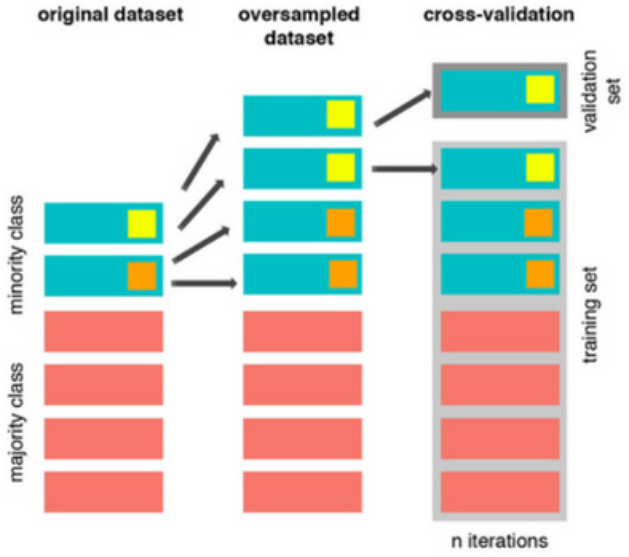


Some models were over fitted

Resampling & Cross Validation



Right Sequence



Wrong Sequence

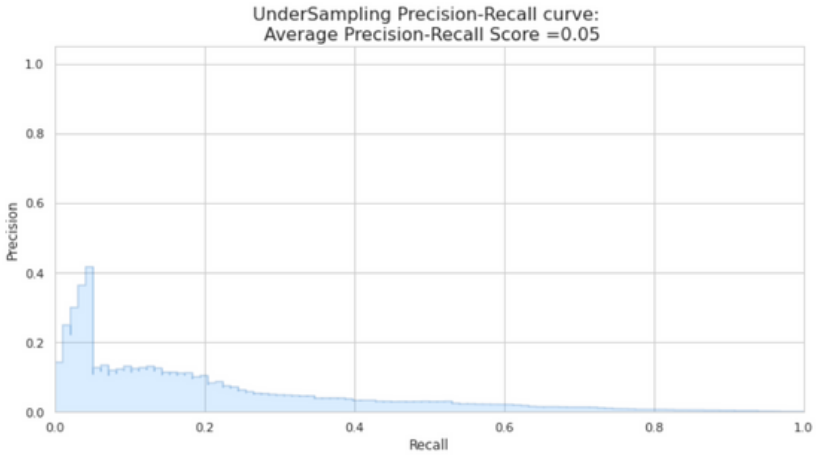
NearMiss Under Sampling

--> Near Miss refers to a collection of undersampling methods that select examples based on the distance of majority class examples to minority class examples.

	precision	recall	f1-score	support
No Fraud	1.00	0.96	0.98	56863
Fraud	0.02	0.54	0.05	98
accuracy			0.96	56961
macro avg	0.51	0.75	0.51	56961
weighted avg	1.00	0.96	0.98	56961

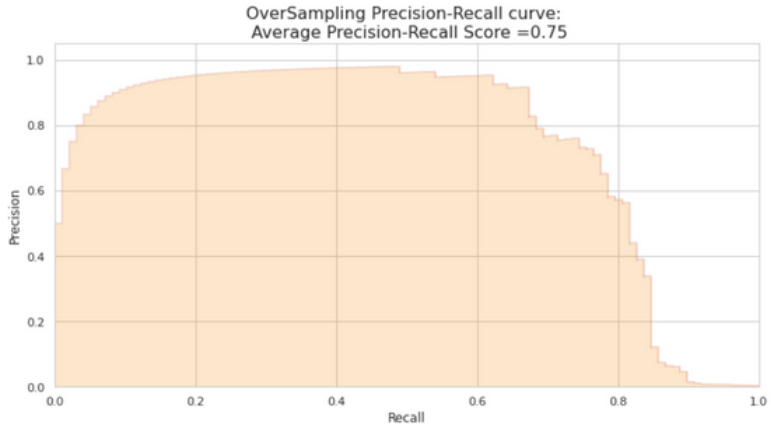
- Results was unexpected

Oversampling VS. Undersampling



	Technique	Score
0	Undersampling	0.961202
1	Oversampling (SMOTE)	0.987799

- Models Score



Evaluation Report

Logistic Regression:

	precision	recall	f1-score	support
No Fraud	0.92	0.96	0.94	100
Fraud	0.95	0.91	0.93	90
accuracy			0.94	190
macro avg	0.94	0.94	0.94	190
weighted avg	0.94	0.94	0.94	190

KNearest Neighbours:

	precision	recall	f1-score	support
No Fraud	0.92	0.97	0.95	100
Fraud	0.96	0.91	0.94	90
accuracy			0.94	190
macro avg	0.94	0.94	0.94	190
weighted avg	0.94	0.94	0.94	190

Support Vector Classifier:

	precision	recall	f1-score	support
No Fraud	0.90	0.97	0.93	100
Fraud	0.96	0.88	0.92	90
accuracy			0.93	190
macro avg	0.93	0.92	0.93	190
weighted avg	0.93	0.93	0.93	190

Logistic regression and KNN models are the best estimators which have the highest recall and f1_score.

Decision Tree:

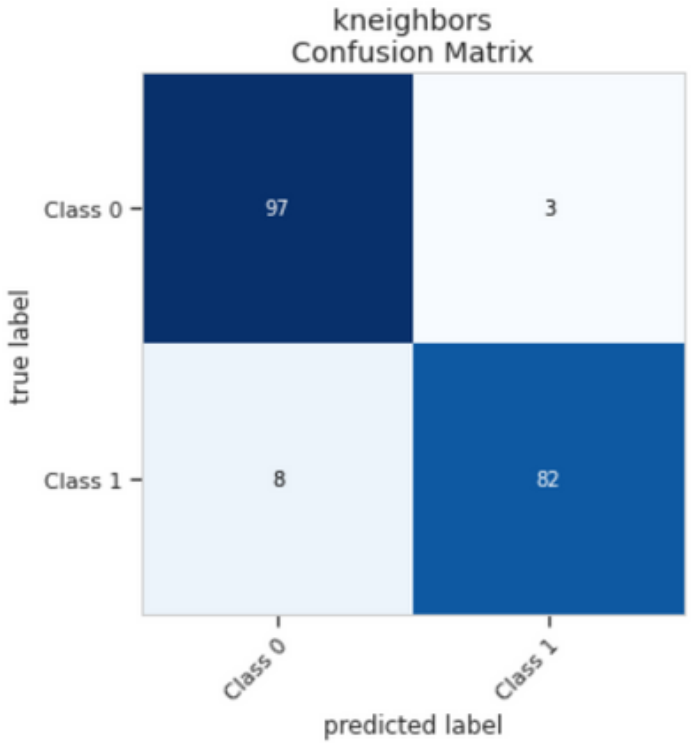
	precision	recall	f1-score	support
No Fraud	0.91	0.97	0.94	100
Fraud	0.96	0.89	0.92	90
accuracy			0.93	190
macro avg	0.94	0.93	0.93	190
weighted avg	0.93	0.93	0.93	190

Random Forest:

	precision	recall	f1-score	support
No Fraud	0.92	0.97	0.94	100
Fraud	0.96	0.90	0.93	90
accuracy			0.94	190
macro avg	0.94	0.94	0.94	190
weighted avg	0.94	0.94	0.94	190

Confusion Matrix

The Best Model:



Modeling Conclusion

CONCLUSION 1

After taking five different classifiers and tested them on the randomly under sampled dataset and evaluated several metrics on it:

1. Accuracy
2. Precision
3. Recall
4. F1-score.
5. ROC Curve

And due to the imbalacing of the data, many observations could be predicted as False Negatives, being, that we predict a normal transaction, but it is in fact a fraudulent one. Recall captures this.

CONCLUSION 2

Implementing SMOTE on our imbalanced dataset helped us with the imbalance of our labels (more no fraud than fraud transactions).

CONCLUSION 3

By under sampling our data, the model was unable to detect for a large number of cases non fraud transactions correctly and instead, misclassifies those non fraud transactions as fraud cases.

Business Solutions

01

Limitations over transactions with large amounts that happens at night.

02

Partnerships with retailers and merchants to offer promotions when using credit card as their payment method

03

Offering free fees cards for youth to encourage them adopt credit cards .

Business Solutions

04

Provide different cards type to meet all customers types needs

05

Partnership with corporates to provide their employees with financial services

06

Increase their awareness about types of fraud they can face and what to do in these cases.

Find Us:

https://www.linkedin.com/in/daliahesham

https://github.com/dalia-hesham

https://www.linkedin.com/in/manarkandil-/

https://github.com/manarkandeel

https://www.linkedin.com/in/yassmen-youssef-48439a166

https://github.com/YASsMeN1997

SAMSUNG