CREDIT CARD

FRAUD DETECTION

GROUP 02

- 1753075 – Huynh Doan Minh Ngoc
- 1753074 – Nguyen Kim Ngan
- 1753086 – Tong Le Thien Phuc

CONTENTS

- Dataset
- Choosing model and data
- Demo and Result

1.

DATASET

We use Kaggle dataset in this project

KAGGLE DATASET

284,807

Instances

31

Attributes

100%

Non-null data of each attribute

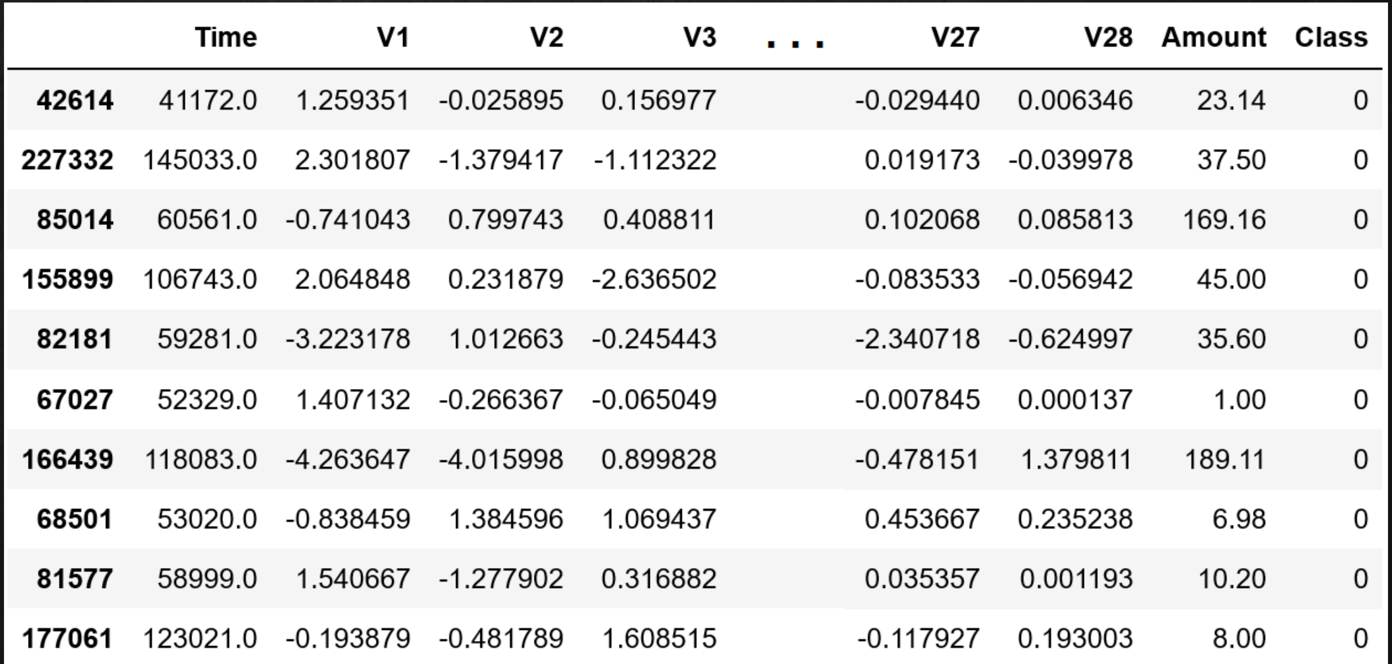
KAGGLE DATASET

31 columns

- Time: number of seconds elapsed
- Amount: transaction amount
- Class:
  - 1 : fraudulent transactions
  - 0 otherwise
- V1 – V28: Result of a PCA Dimensionality reduction

Maps

Random 10 samples from the dataset

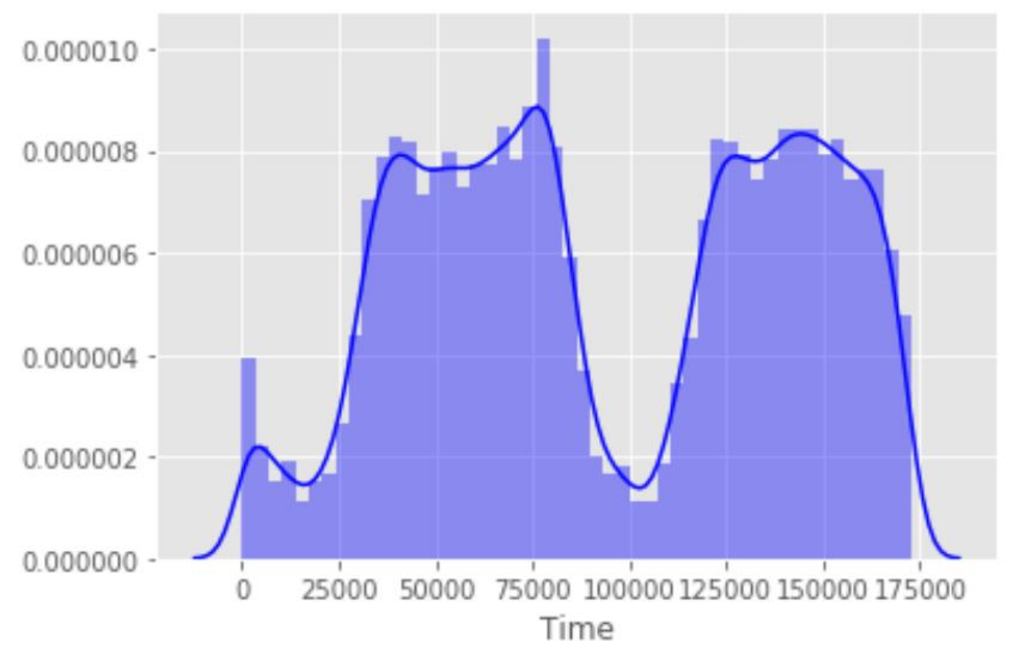


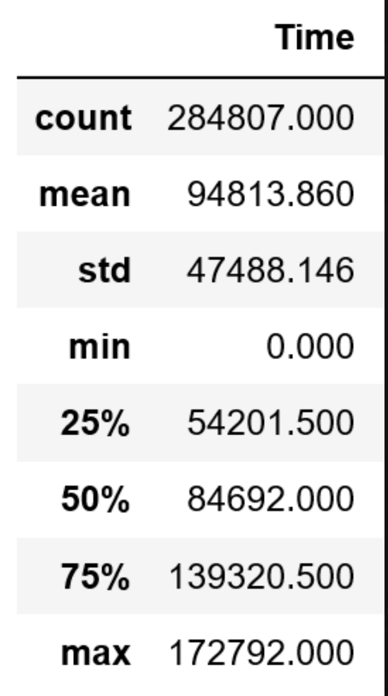
	Time	V1	V2	V3	...	V27	V28	Amount	Class
42614	41172.0	1.259351	-0.025895	0.156977		-0.029440	0.006346	23.14	0
227332	145033.0	2.301807	-1.379417	-1.112322		0.019173	-0.039978	37.50	0
85014	60561.0	-0.741043	0.799743	0.408811		0.102068	0.085813	169.16	0
155899	106743.0	2.064848	0.231879	-2.636502		-0.083533	-0.056942	45.00	0
82181	59281.0	-3.223178	1.012663	-0.245443		-2.340718	-0.624997	35.60	0
67027	52329.0	1.407132	-0.266367	-0.065049		-0.007845	0.000137	1.00	0
166439	118083.0	-4.263647	-4.015998	0.899828		-0.478151	1.379811	189.11	0
68501	53020.0	-0.838459	1.384596	1.069437		0.453667	0.235238	6.98	0
81577	58999.0	1.540667	-1.277902	0.316882		0.035357	0.001193	10.20	0
177061	123021.0	-0.193879	-0.481789	1.608515		-0.117927	0.193003	8.00	0

Exploratory Data Analysis

Few initial comparisons between three attributes: Time, Amount, Class

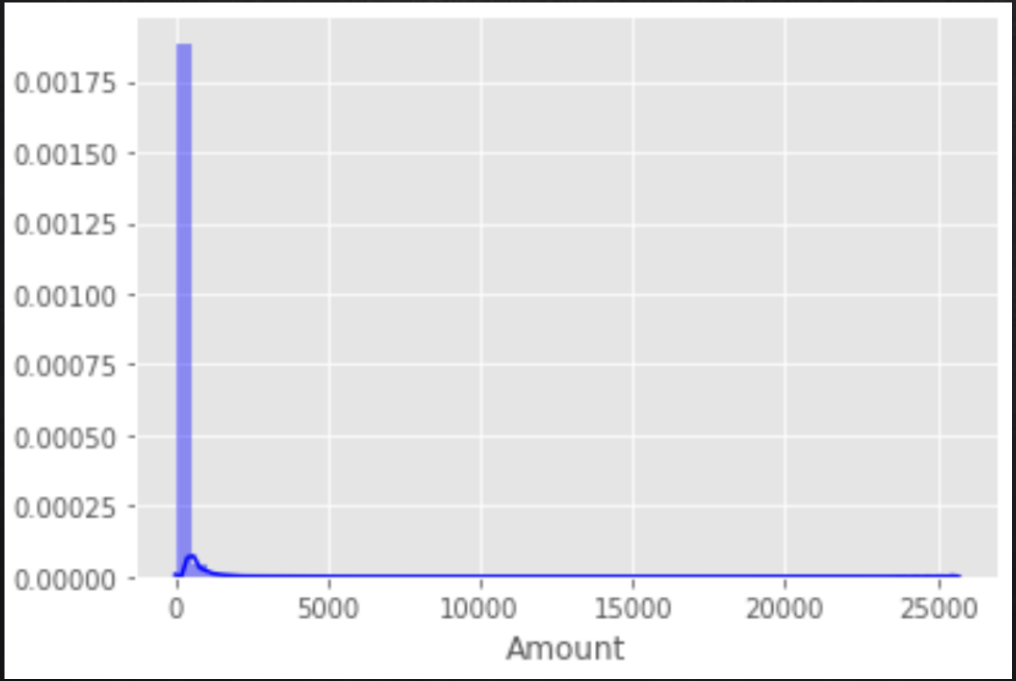
Time Distribution of Credit Card Data

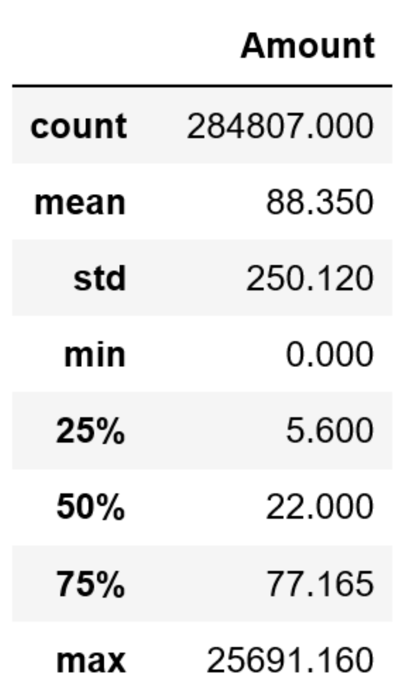




	Time
count	284807.000
mean	94813.860
std	47488.146
min	0.000
25%	54201.500
50%	84692.000
75%	139320.500
max	172792.000

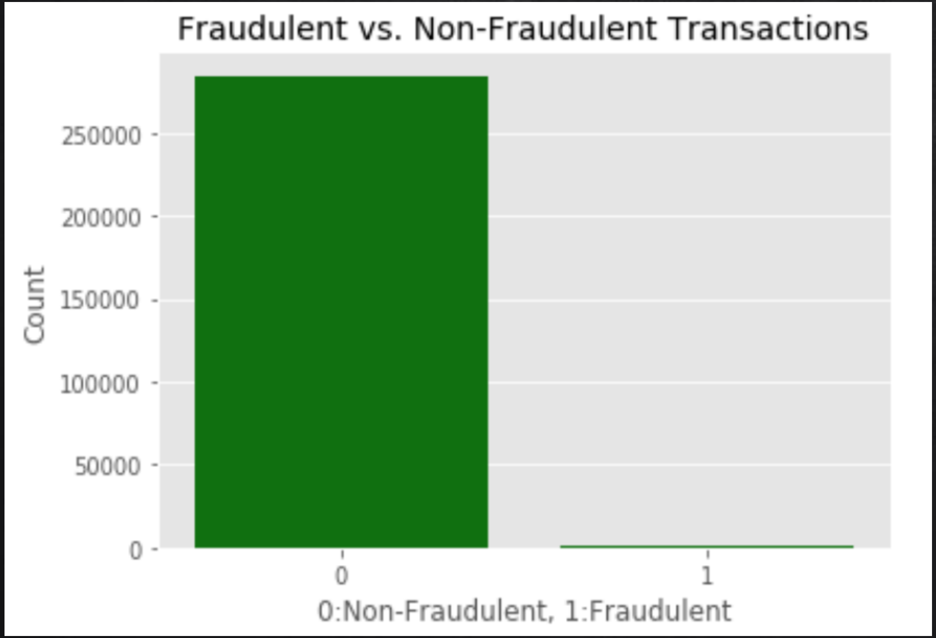
Amount Distribution of Credit Card Data





	Amount
count	284807.000
mean	88.350
std	250.120
min	0.000
25%	5.600
50%	22.000
75%	77.165
max	25691.160

Class Attribute



Ratio: 0.173%	Number of transactions
Fraudulent	✗ 492
Non-fraudulent	✓ 284,315

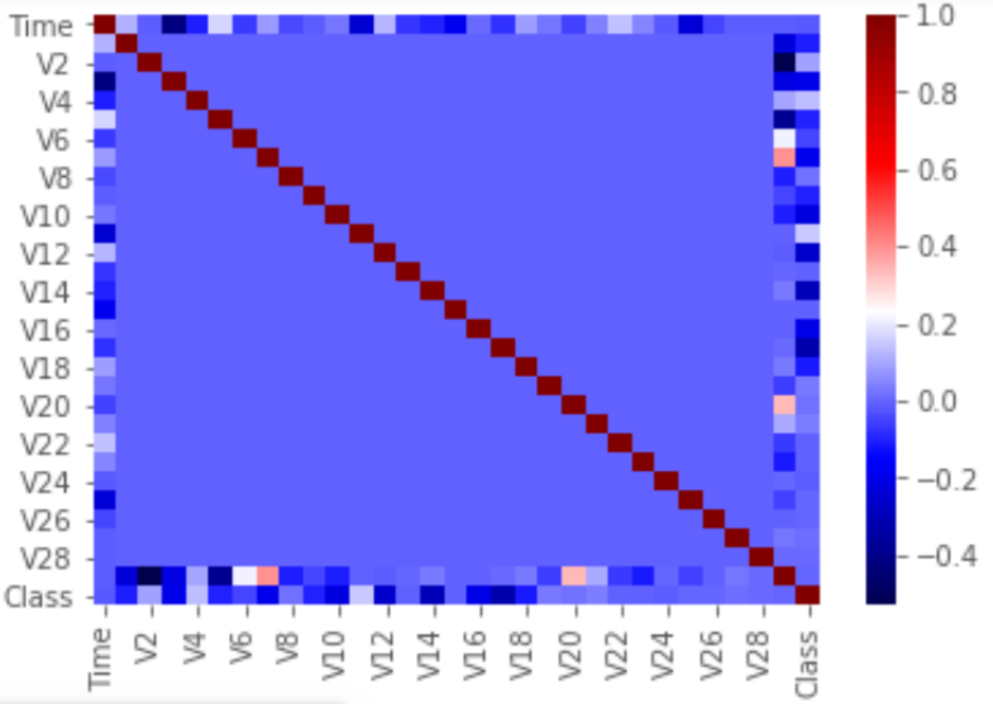
Only 2 days

$60,127.97

Total amount of fraud transactions

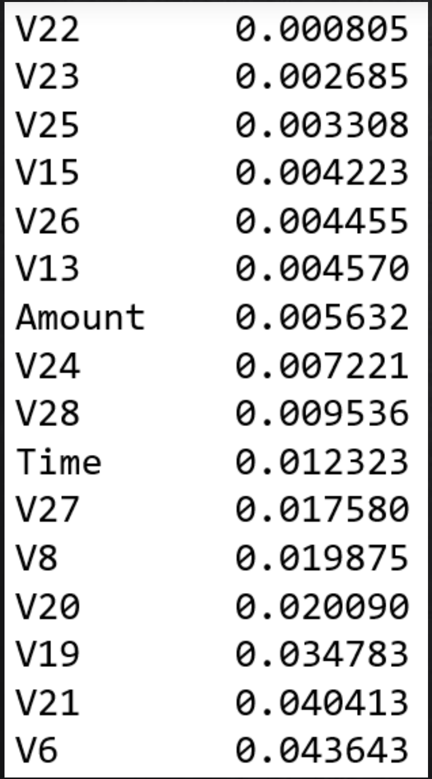
(Nearly 0.24%)

Find Highest Correlations

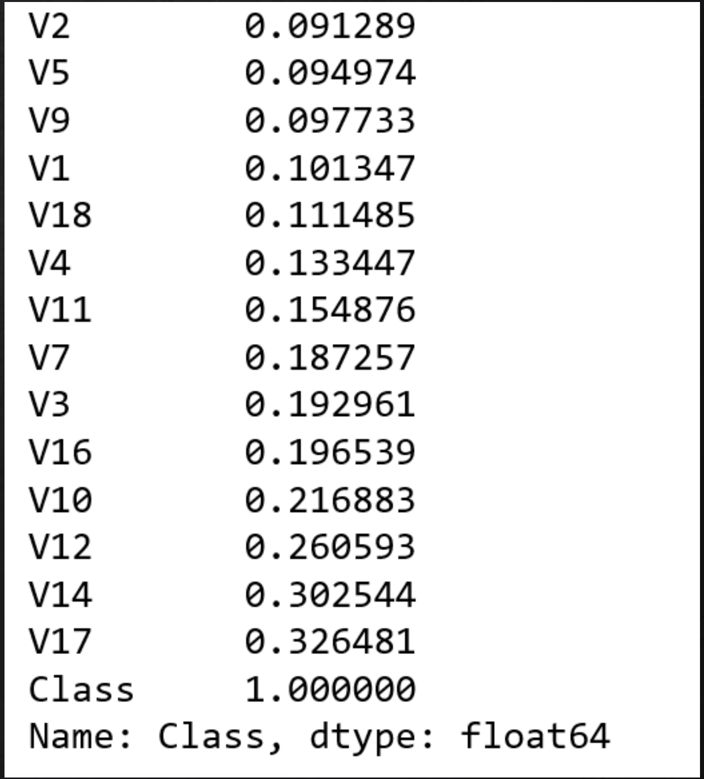


- ✗ Time & V3 (−0.42)
- ✗ Amount & V2 (−0.53)
- ✗ Amount & V4 (0.4)

Find Highest Correlations Of Class Attribute



```
V22       0.000805
V23       0.002685
V25       0.003308
V15       0.004223
V26       0.004455
V13       0.004570
Amount    0.005632
V24       0.007221
V28       0.009536
Time      0.012323
V27       0.017580
V8        0.019875
V20       0.020090
V19       0.034783
V21       0.040413
V6        0.043643
```

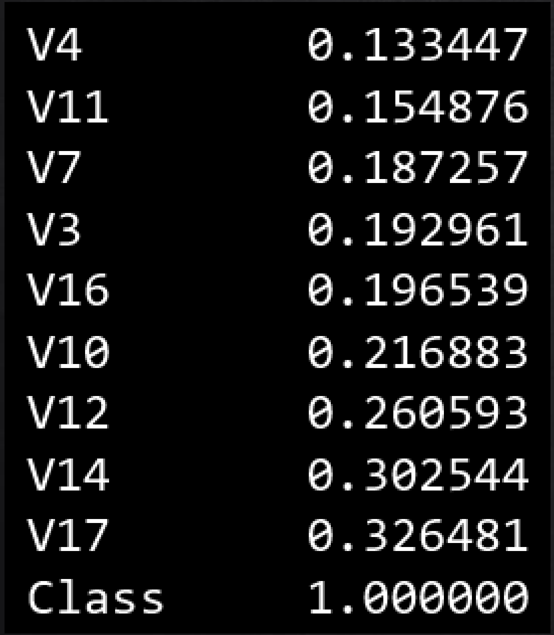


```
V2        0.091289
V5        0.094974
V9        0.097733
V1        0.101347
V18       0.111485
V4        0.133447
V11       0.154876
V7        0.187257
V3        0.192961
V16       0.196539
V10       0.216883
V12       0.260593
V14       0.302544
V17       0.326481
Class     1.000000
Name: Class, dtype: float64
```

2.

CHOSING MODEL AND DATA

Choose Attributes for Training



```
V4        0.133447
V11       0.154876
V7        0.187257
V3        0.192961
V16       0.196539
V10       0.216883
V12       0.260593
V14       0.302544
V17       0.326481
Class     1.000000
```

- 10 columns ⇔ 10 attributes
- Based on correlation values
- Avoiding overfiting

10 chosen columns are: **Class, V17, V14, V12, V10, V16, V3, V7, V11, V4.**

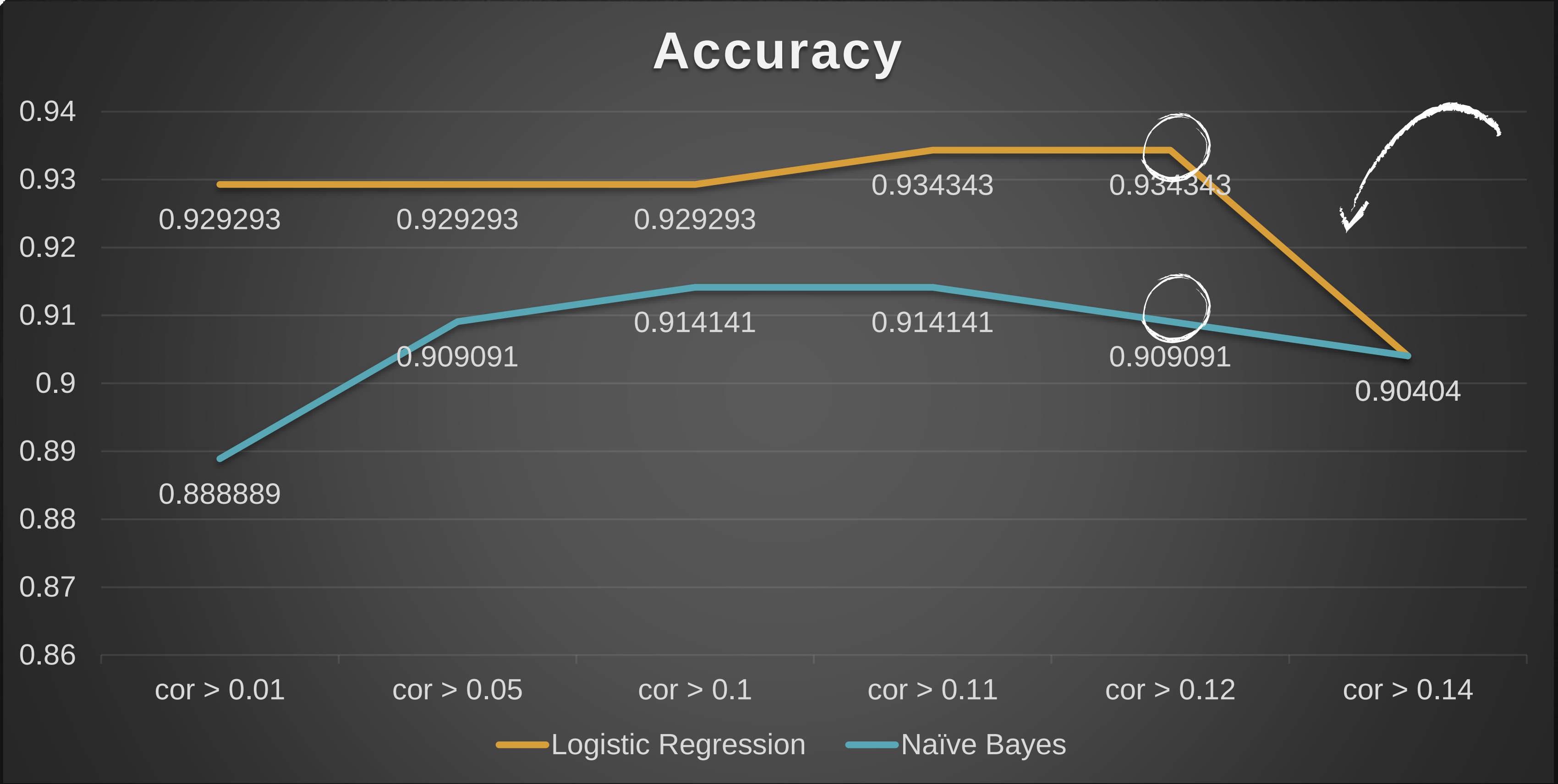
Before Training

- Splitting dataset to 2 paths: training, validation and testing
    - Test_size = 0.2
    - Train_size = 0.75 (of 0.8 dataset)
    - Val_size = 0.25 (of 0.8 dataset)
- Using z-score to normalize

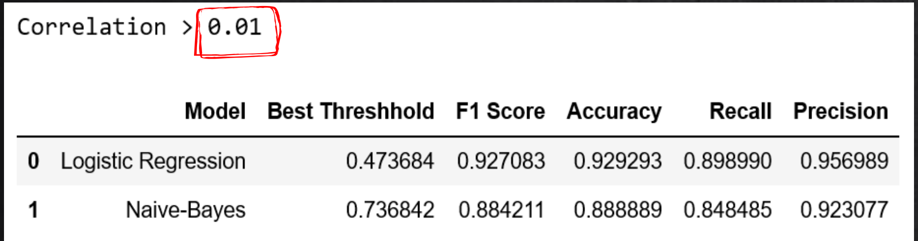
Models for Training

- Logistic Regression
- Naïve Bayes Model
- Training with 20 epochs to find best threshold (using validation dataset)

Accuracy Of 2 Models (validation set)

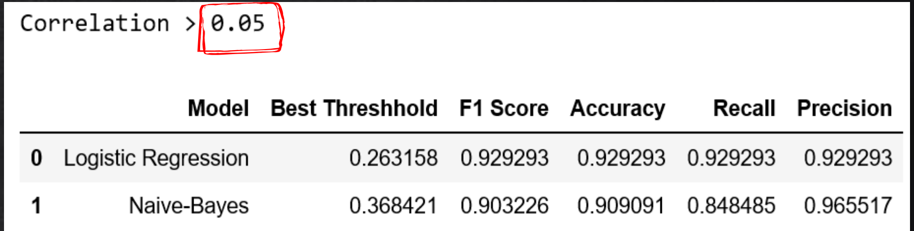


Compare results of Different Chosen Dataset



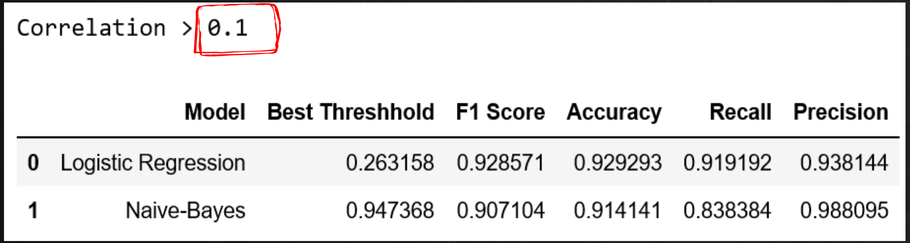
Correlation > 0.01

	Model	Best Threshhold	F1 Score	Accuracy	Recall	Precision
0	Logistic Regression	0.473684	0.927083	0.929293	0.898990	0.956989
1	Naive-Bayes	0.736842	0.884211	0.888889	0.848485	0.923077



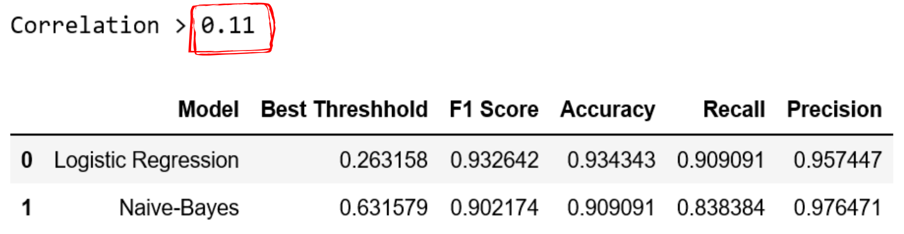
Correlation > 0.05

	Model	Best Threshhold	F1 Score	Accuracy	Recall	Precision
0	Logistic Regression	0.263158	0.929293	0.929293	0.929293	0.929293
1	Naive-Bayes	0.368421	0.903226	0.909091	0.848485	0.965517



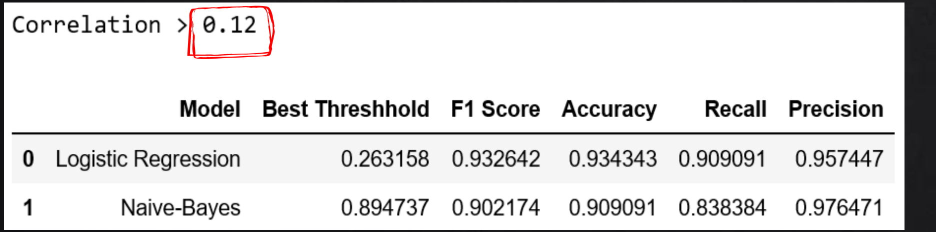
Correlation > 0.1

	Model	Best Threshhold	F1 Score	Accuracy	Recall	Precision
0	Logistic Regression	0.263158	0.928571	0.929293	0.919192	0.938144
1	Naive-Bayes	0.947368	0.907104	0.914141	0.838384	0.988095



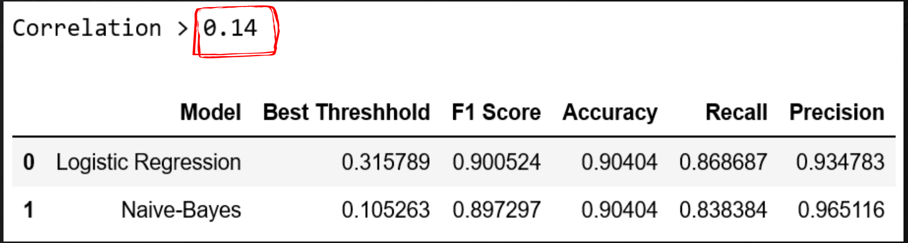
Correlation > 0.11

	Model	Best Threshhold	F1 Score	Accuracy	Recall	Precision
0	Logistic Regression	0.263158	0.932642	0.934343	0.909091	0.957447
1	Naive-Bayes	0.631579	0.902174	0.909091	0.838384	0.976471



Correlation > 0.12

	Model	Best Threshhold	F1 Score	Accuracy	Recall	Precision
0	Logistic Regression	0.263158	0.932642	0.934343	0.909091	0.957447
1	Naive-Bayes	0.894737	0.902174	0.909091	0.838384	0.976471



Correlation > 0.14

	Model	Best Threshhold	F1 Score	Accuracy	Recall	Precision
0	Logistic Regression	0.315789	0.900524	0.90404	0.868687	0.934783
1	Naive-Bayes	0.105263	0.897297	0.90404	0.838384	0.965116

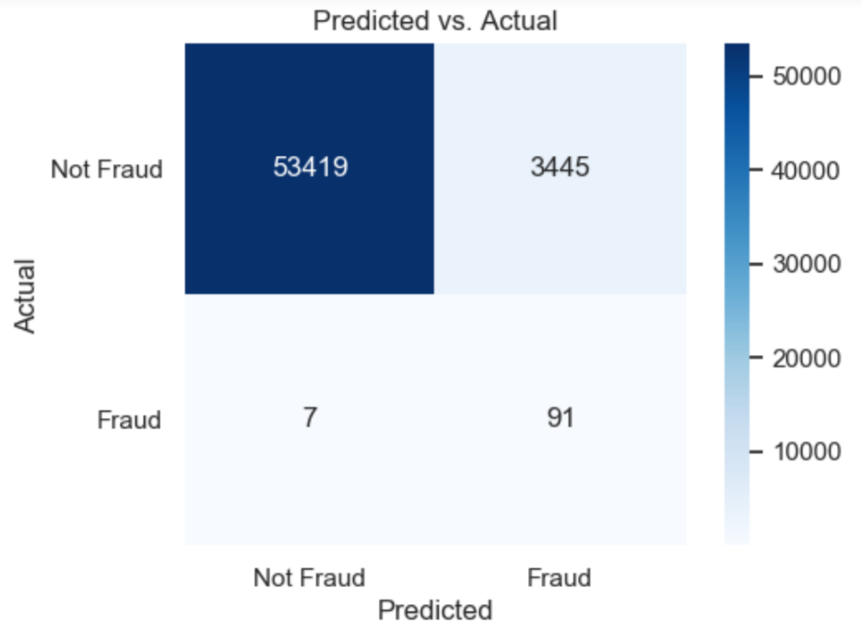
3.

DEMO AND RESULT

Compare Accuracy of Two Models (test set)

Accuracy of logistic regression model: 0.94

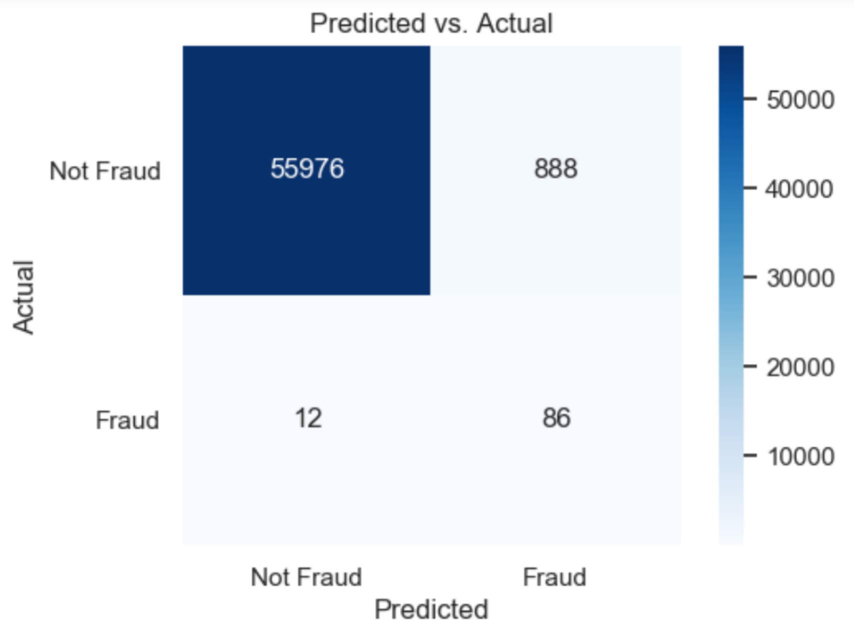
Accuracy of Naïve Bayes model: 0.98



Predicted vs. Actual

Actual \ Predicted	Not Fraud	Fraud
Not Fraud	53419	3445
Fraud	7	91

Logistic Regresison



Predicted vs. Actual

Actual \ Predicted	Not Fraud	Fraud
Not Fraud	55976	888
Fraud	12	86

Naïve Bayes

Thanks For Listening!

Any questions?

Group 02