DATA SCIENCE PORTOFOLIO

By : Adi Sutrisno

ADI SUTRISNO

I have completed bachelor's degree in mining engineering and currently looking for job in practical data. I've been learning data science to expand my knowledge and upgrade my skills. Currently active learning from web binary and data science bootcamp to develop my knowledge and get more expert

Skill and Proficiency

- Python (Programming Language)
- SQL Database
- Tableau (Data Visualization)
- Machine Learning

Credit Card Fraud Detection

By : Adi Sutrisno

Outline

01

Background

02

Data Perprocessing

03

Modeling And Evaluation

04

Insight

01

Background

Background

Credit card fraud is a form of identity theft that involves an unauthorized taking of another's credit card information for the purpose of charging purchases to the account or removing funds from it.

losses due to fraudulent cashless transactions aka cybercriminals worldwide have reached US$ 21 billion or more than Rp300 trillion in 2015.

The value of losses continues to increase and is expected to reach US$31 billion in 2020 (Nielson Report)

The dataset contains transactions made by credit cards in September 2013 by European cardholders.
Data from Kaggle Uploaded by Machine Learning Group

Bussines Question

- how to identify the new transaction is fraudulent or not?
- The goal is to detect 100% of fraudulent transactions while minimizing the classification of fraudulent frauds incorrectly.

02
Data Preprocessing

Dataset Infromation

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	149.62	0
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	2.69	0
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	378.66	0
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458	123.50	0
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	69.99	0

5 rows × 31 columns

284.807

rows

30

features

1

target

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

Target variable

- Class

1 = Fraud

0 = No Fraud

Missing value and duplicated data

No missing value found, but duplicated data exist

0

Missing Values

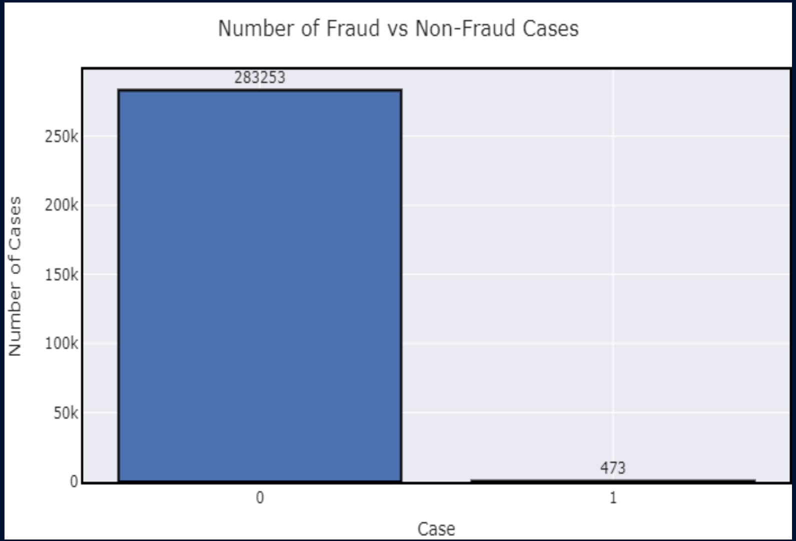
1081

Duplicated Data

After we drop duplicated data, shape the data to 283.726 rows

Data Pre-Processing & Analysis

Examine the class imbalance



Class	Transaction	Total Revenue ($)
No Fraud	283.253	25.043.410,29
Fraud	473	58.591,39

Data Pre-Processing & Analysis

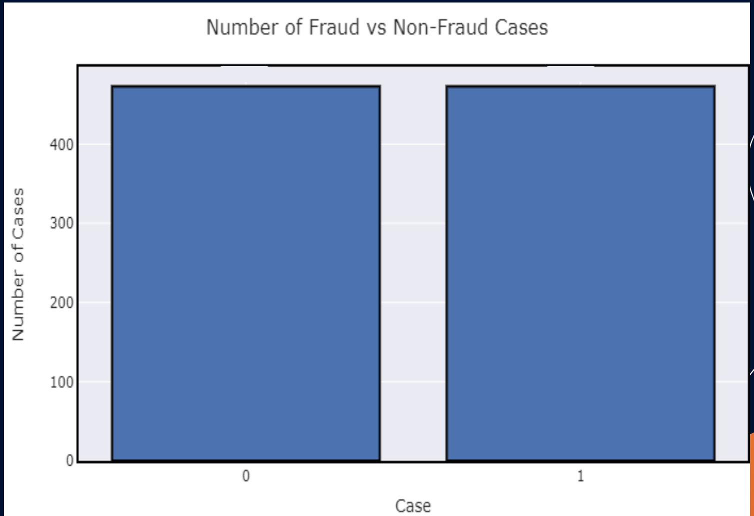
Feature Scaling

From the analysis of the dataset above we can see that the existing features have been scaled, except for the Amout and Time features. therefore we scale the feature

	scaled_amount	scaled_time
0	1.774718	-0.995290
1	-0.268530	-0.995290
2	4.959811	-0.995279
3	1.411487	-0.995279
4	0.667362	-0.995267

Sampling Data (Random Undersampling)

random undersampling is done to balance the data

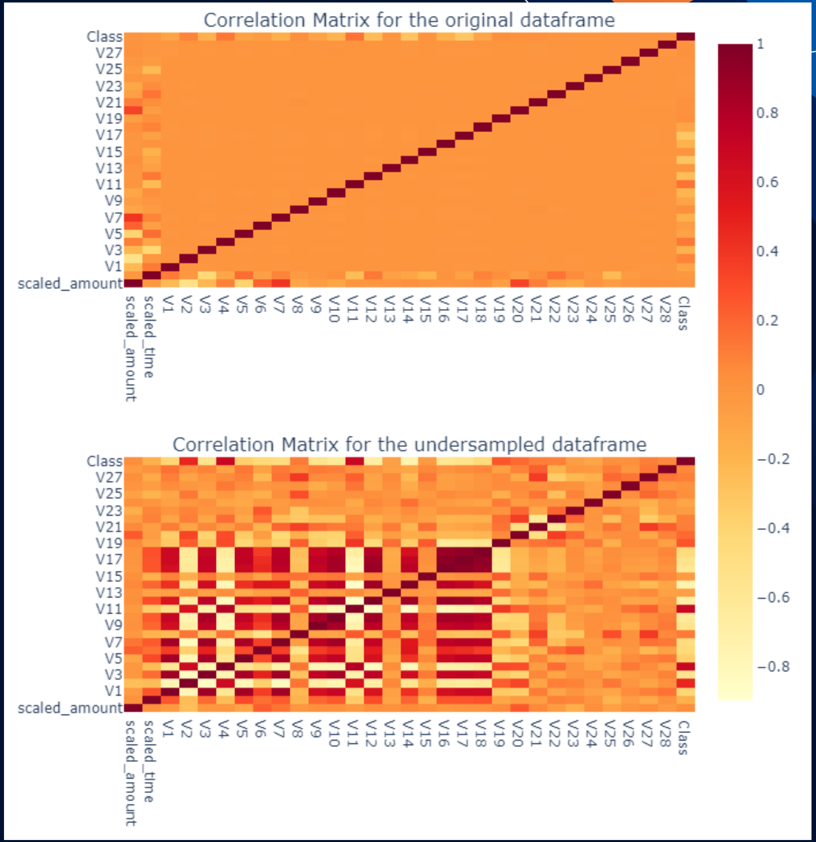


Data Pre-Processing & Analysis

Heatmap Correlation

Positive Correlation: Features V2, V4, V11 & V19 show a positive correlation with the class. The higher the value of these features, the higher the likelihood of the transaction becoming a fraudster.

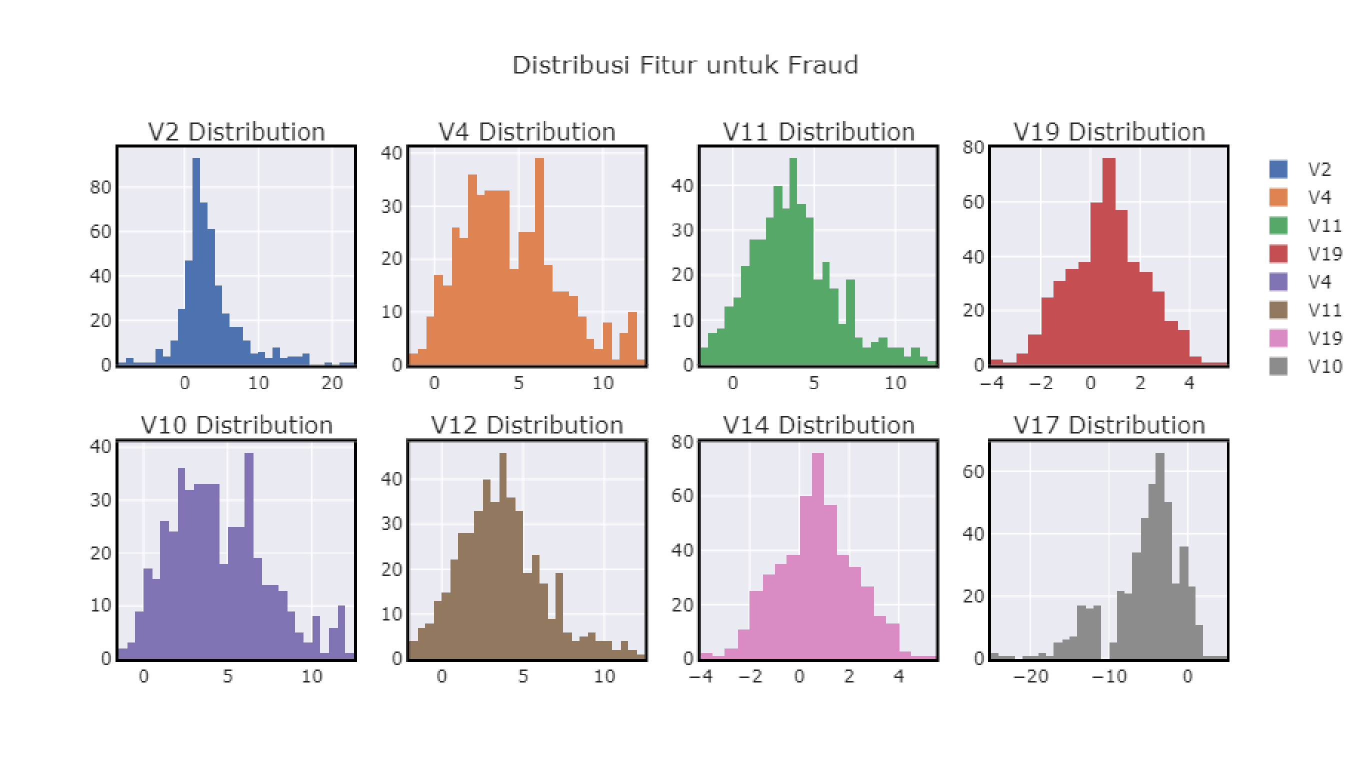
Negative Correlation: Features V10, V12, V14 & V17 show a negative correlation with classes. The lower the value of these features, the higher the likelihood of the transaction becoming a fraudster



Data Pre-Processing & Analysis

Histogram

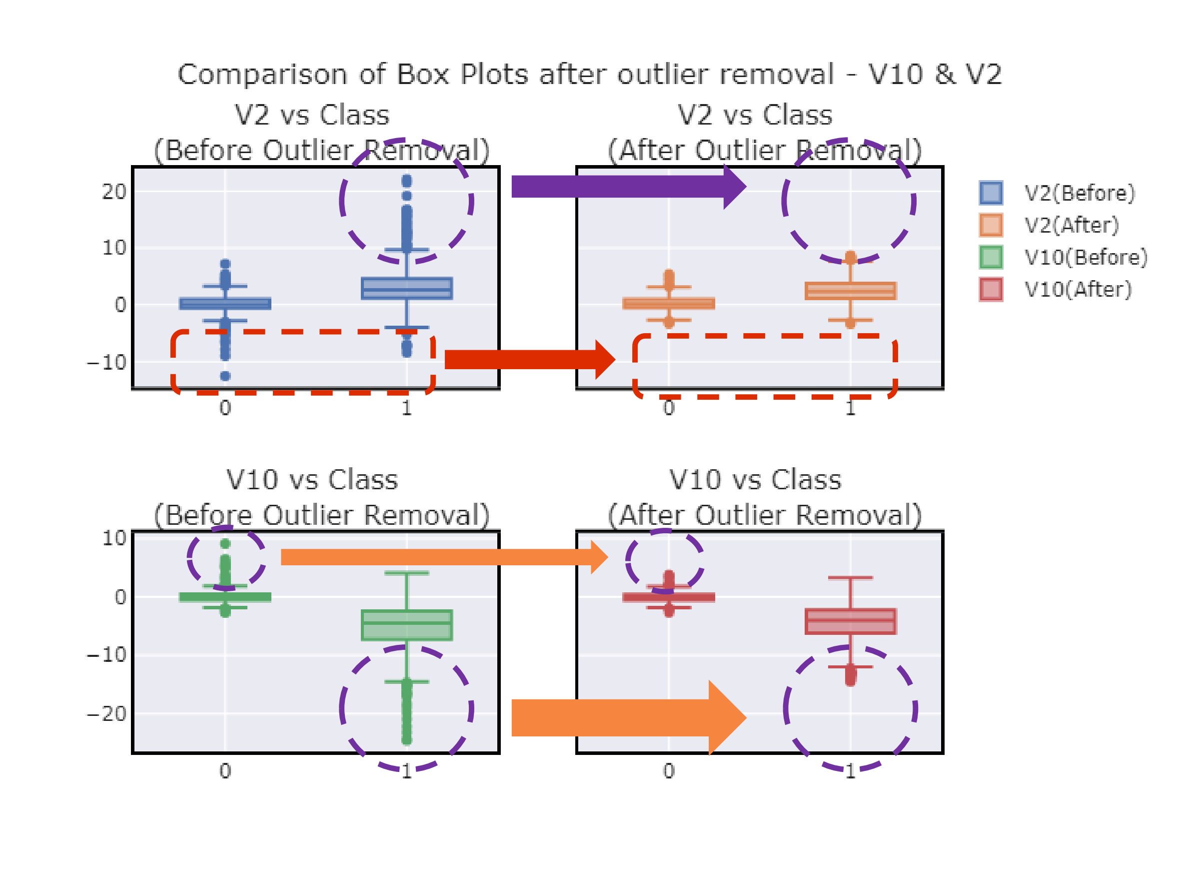
You can see the distribution of each feature that has a correlation



Data Pre-Processing & Analysis

Box Plot

I have plotted boxplot for all the features that show both positive and negative correlations with the Class. The boxplot representation for each feature is displayed separately for each class (0 &1).



Outlier Removal

using the inter quartile method. we can identify and discard outliers with a cutoff of 1.5 x IQR

we can see the outlier is missing, by using the lower limit and the upper limit.

03 Modeling and Evaluation

Base Model Comparison - Classifier

Model	tp	tn	fp	fn	correct	incorrect	accuracy	precision	recall	f1	roc_auc
GaussianNB	76	92	4	18	168	22	0.884	0.950	0.809	0.874	0.883
DummyClassifier	94	0	96	0	94	96	0.495	0.495	1.000	0.662	0.500
LogisticRegression	85	94	2	9	179	11	0.942	0.977	0.904	0.939	0.942
LGBMClassifier	83	93	3	11	176	14	0.926	0.965	0.883	0.922	0.926
XGBClassifier	82	93	3	12	175	15	0.921	0.965	0.872	0.916	0.921
DecisionTreeClassifier	82	87	9	12	169	21	0.889	0.901	0.872	0.886	0.889
RandomForestClassifier	81	94	2	13	175	15	0.921	0.976	0.862	0.915	0.920
AdaBoostClassifier	84	92	4	10	176	14	0.926	0.955	0.894	0.923	0.926
GradientBoostingClassifier	82	92	4	12	174	16	0.916	0.953	0.872	0.911	0.915
GaussianNB	76	92	4	18	168	22	0.884	0.950	0.809	0.874	0.883

The dataset is splitted into train (80%) and test data (20%)

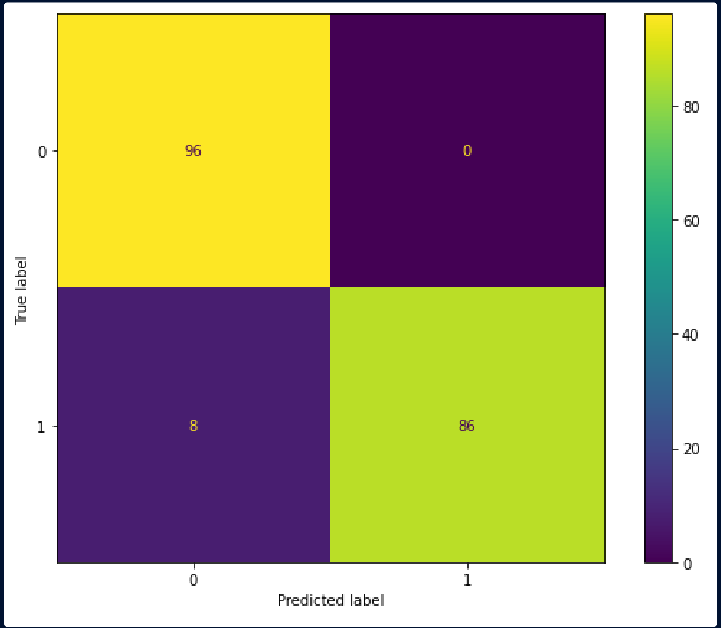
From the table the Top 2 Model is Logistic Regression, and Random Forest

Hyperparameter Tuning

Logistic Regression

From the results of hyperparameter tuning we can see several things, including:

1. Precision increased from the original 97.7% to 99%
2. The f1-score value increased from 93.9 % to 95.7%



Model	tp	tn	fp	fn	precision	F1-score
Base LR	83	93	3	11	0.977	0.939
Tuned LR	96	86	0	8	0.99	0.957

Insight 04

Insight

- Fraudulent transactions occur for 2 days or 172.792 second
- Fraudulent transactions occurred 473 times with a amount 58,591.39 dollars
- Feature V2, V4, V11 & V19 have a positive correlation with Fraud.
  Featured V10, V12, V14, V17 have a negative correlation with Fraud
- Top 5 features Importance V14, V10, V16, V11, V3 most influential against fraud
- Our best model logistic regression (undersampling) has a presicion of 99.9%.
- correctly predicted 96 of the 104 frauds in the 284,795 data set. and it's not wrong to mark fraud when people aren't cheating.
- Based on the average fraud transaction value of $123.87 per transaction.
  Our model can save the bank **$5,945.76** per day, and **$2,170,202.4** per year.

Thanks!

Do you have any questions?

adisutrisno22@gmail.com
+62822 1782 3863