Trustly – Data Scientist Challenge

By Tomas Mosconi de Gouvêa

Summary ##arrumar

1- Challenge instructions:

•Develop a model from the received base (view the file): dataset_test_ds.csv

•There is no right or wrong approach; there are different ways to achieve the same result, although the data tells us what techniques we can use.

•The interpretation of the base is part of the evaluation.

•Be simple in approach. The goal is to assess the reasoning process behind the work.

•A presentation in English with the results of the model (in .pdf or .ppt), selected technique, metrics, analyzes used and recommendations must be returned.

•**Results must be sent to Jobs-br@trustly.com .**

2- Challenge context

Challenge from Trustly to perform a model on dataset_test_ds.csv. No contextualization of the problem or description of the data was provide. Also, there was no given successful metrics to achieve.

The work consists in analyzing the dataset, identifying which algorithms can have better results.
Due to the lack of metrics and contextualization, only one round of CRISP-DM will be execute. The need for further training can be confirmed only after the business evaluation.

The steps Business understanding, Evaluation and Deployment from the CRISP-DM plan can't be done, because the lack of information.

The code can be found on github
https://github.com/tmosconi/Challenge_trustly

3- Data understanding

In this phase will be analyze the data set provided.

- Data types
- Amount
- Missings
- Range
- Outliers
- Correlation

3.1 - Data understanding - Dataset

The data set has 11169 records and 12 variables.

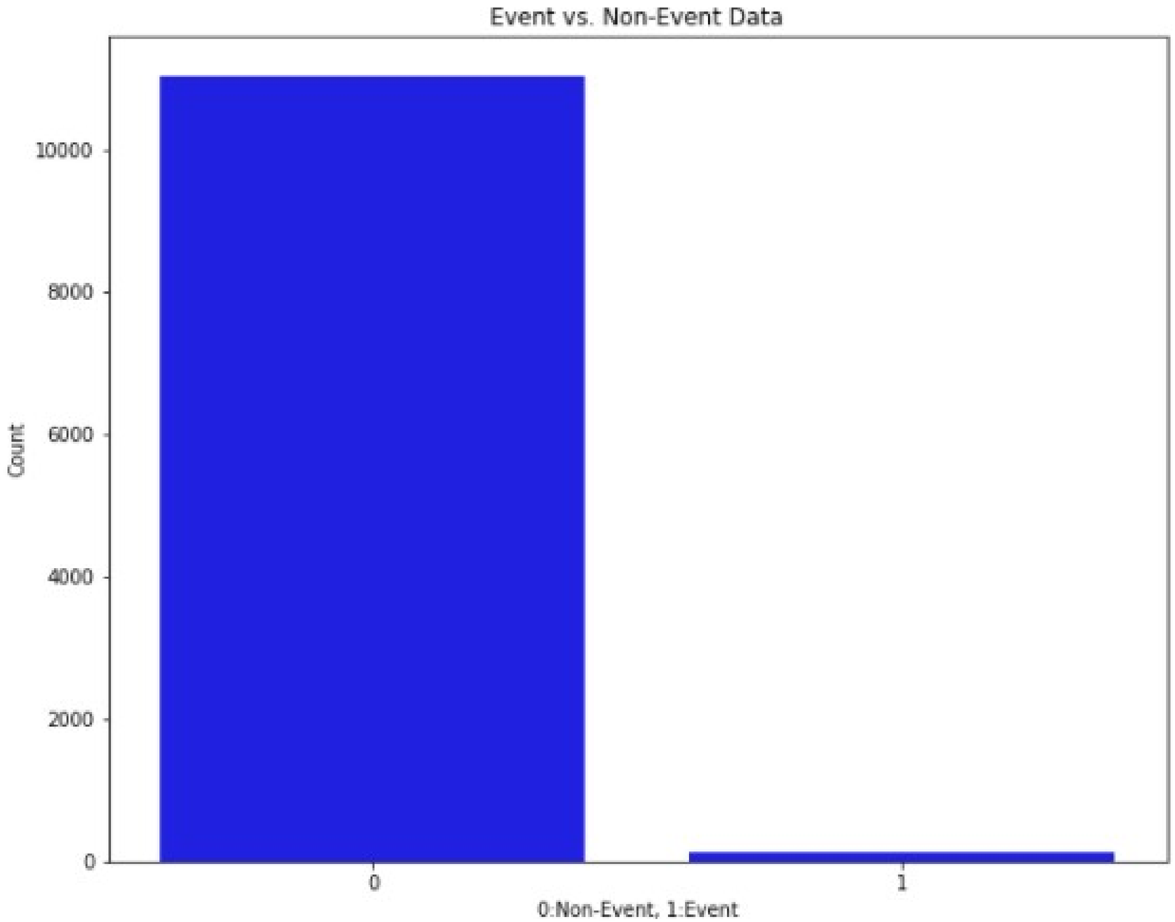
•10 variables identified as V1, V2 ... V10,
•A time variable identified as “Safra” in the Ym format
•A target variable, Boolean.
•There no null values

As a first step is to analyze the Target and Safra variables.

```
RangeIndex: 11169 entries, 0 to 11168
Data columns (total 12 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   V1      11169 non-null  int64  
 1   V2      11169 non-null  float64
 2   V3      11169 non-null  float64
 3   TARGET  11169 non-null  int64  
 4   V4      11169 non-null  int64  
 5   V5      11169 non-null  int64  
 6   V6      11169 non-null  int64  
 7   V7      11169 non-null  float64
 8   V8      11169 non-null  int64  
 9   V9      11169 non-null  int64  
 10  V10     11169 non-null  int64  
 11  Safra   11169 non-null  int64  
dtypes: float64(3), int64(9)
```

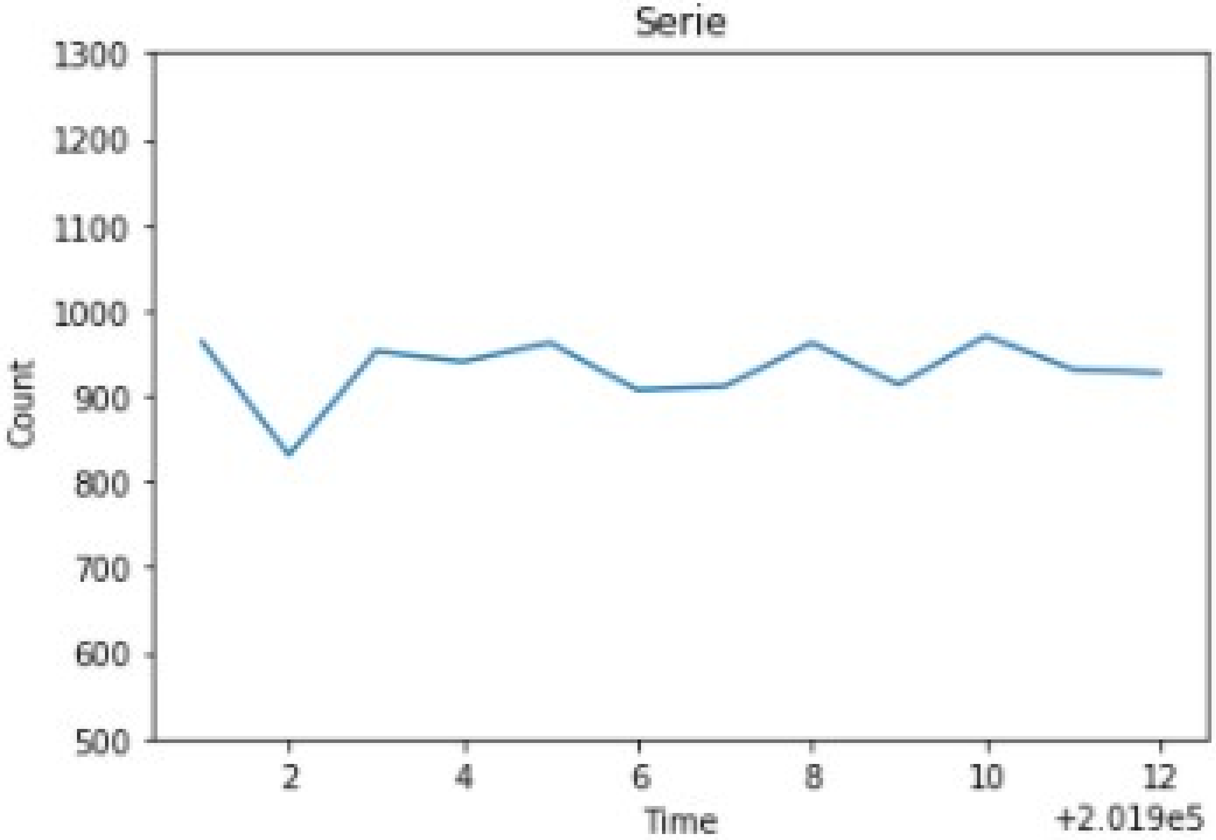
3.2 - Data understanding - Target

The Target variable, is a boolean variable and has 120 true (1) events, which represents 1% of the dataset. That may indicates the dataset refers to a rare events, like fraud.



3.3 - Data understanding - Safra

The variable Safra, refer to the year 2019 separated by month. The data are well distributed over time, noting only a slight drop in 2019/02



3.4 - Data understanding – Other Variables

Follow the study of the other variables, V1, V2 ... V10

DATASET'S DESCRIPTIVE STATISTICS:

	V1	V2	V3	V4	V5
count	11169.000000	11169.000000	11169.000000	11169.000000	11169.000000
mean	0.106366	19.726368	531.046901	1396.048438	0.189990
std	0.308319	25.438201	906.626021	1736.590512	0.656058
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	2.800000	37.520000	30.000000	0.000000
50%	0.000000	10.000000	135.000000	1321.000000	0.000000
75%	0.000000	25.200000	520.000000	1988.000000	0.000000
max	1.000000	100.000000	8540.000000	15616.000000	11.000000

	V6	V7	V8	V9	V10
count	11169.000000	11169.000000	11169.000000	11169.000000	11169.000000
mean	0.177903	4346.085975	0.397529	0.008506	0.030531
std	0.382448	11542.516550	0.489409	0.091837	0.172051
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	77.420000	0.000000	0.000000	0.000000
50%	0.000000	414.070000	0.000000	0.000000	0.000000
75%	0.000000	2799.060000	1.000000	0.000000	0.000000
max	1.000000	143268.550000	1.000000	1.000000	1.000000

3.4 - Data understanding – Other Variables

- There are no null values

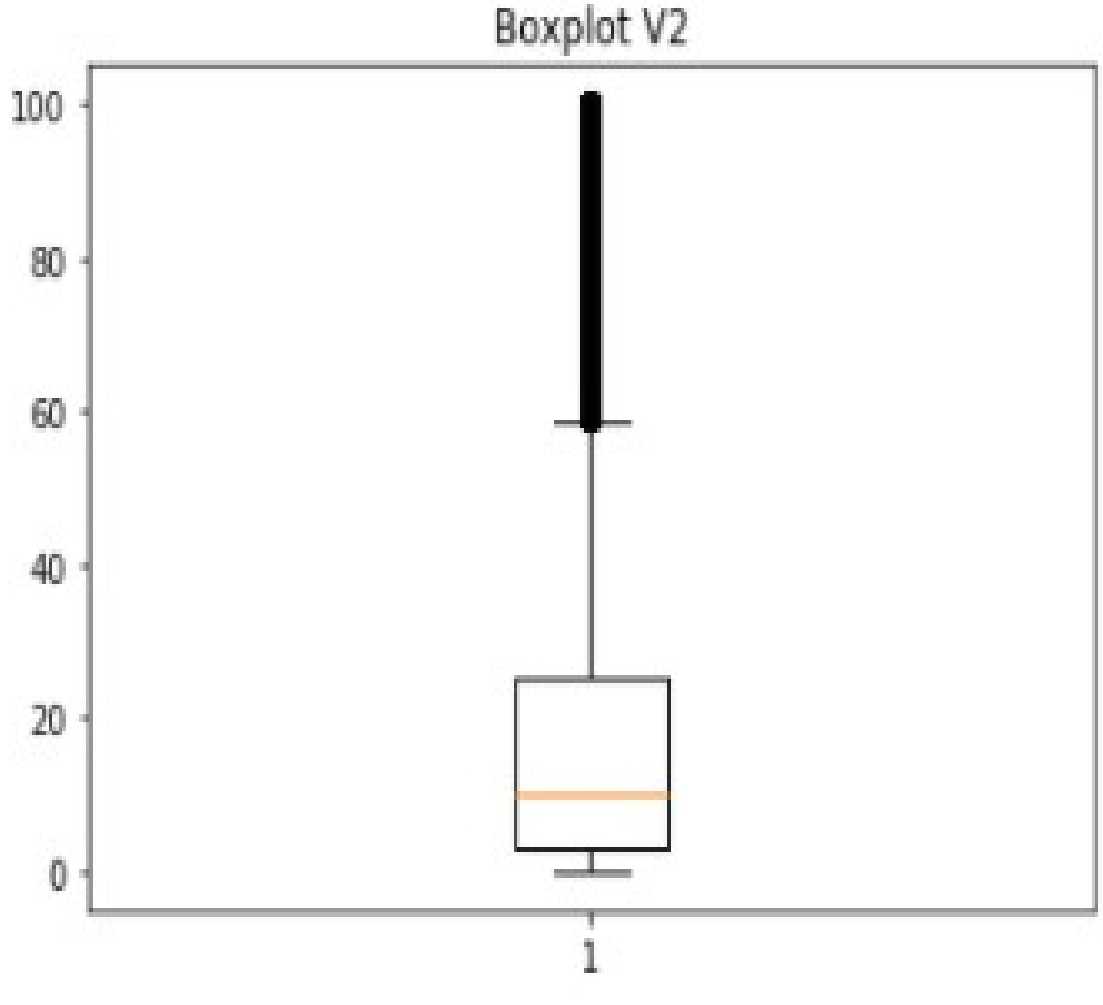
Variable types

- V1 - Boolean 0, 1
- V2 - Numeric, values between 0-100 (can be a percent variable)
- V3 - Numeric
- V4 - Integer
- V5 - Integer
- V6 - Boolean 0, 1
- V7 - Numeric
- V8 - Boolean 0, 1
- V9 - Boolean 0, 1
- V10 - Boolean 0, 1

The are a significant variation in the variables V2, V3, V4, V5 and V7 between the 3 quartile and the 4 quartile, which may indicate the presence of an outlier.

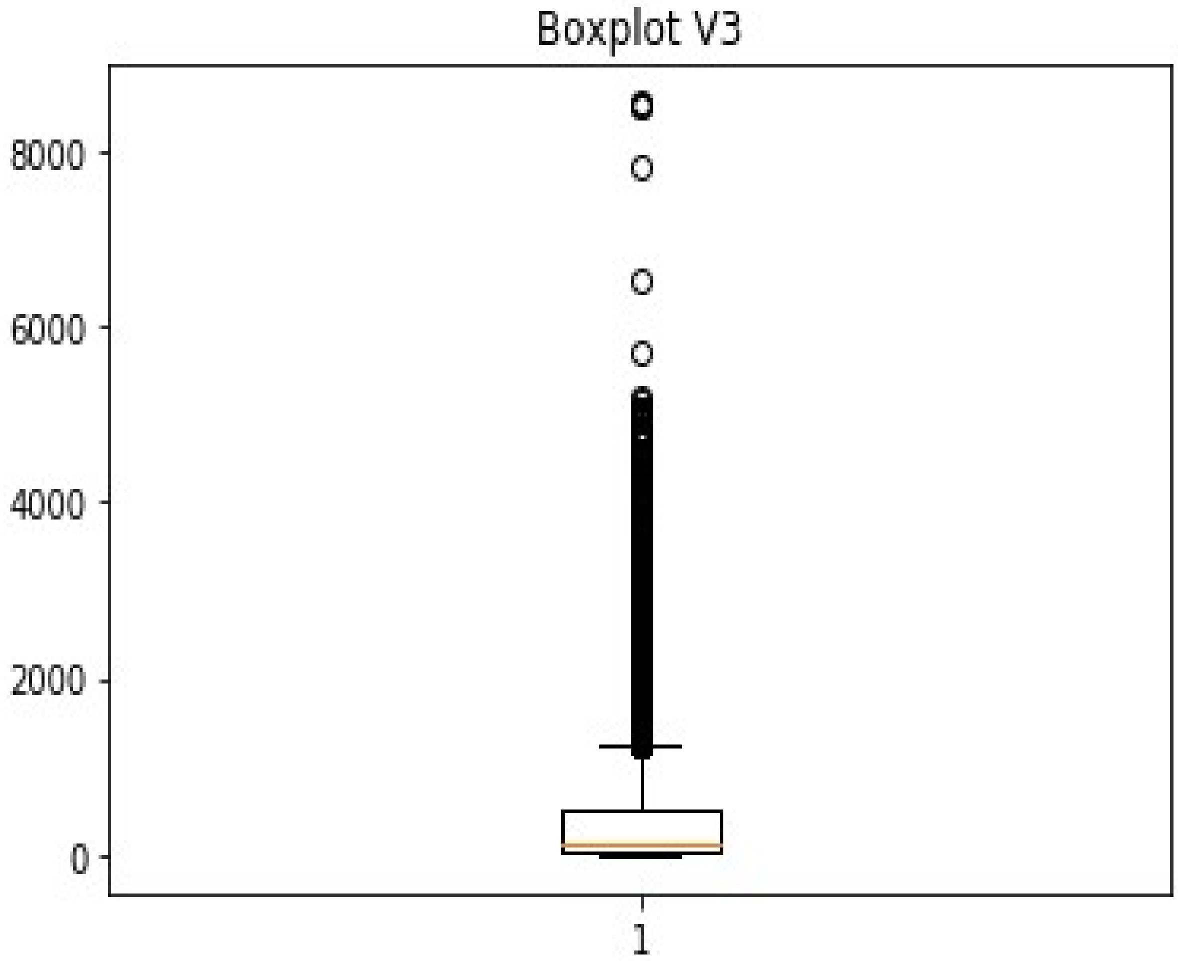
3.4 - Data understanding – Other Variables

Outlier study – V2



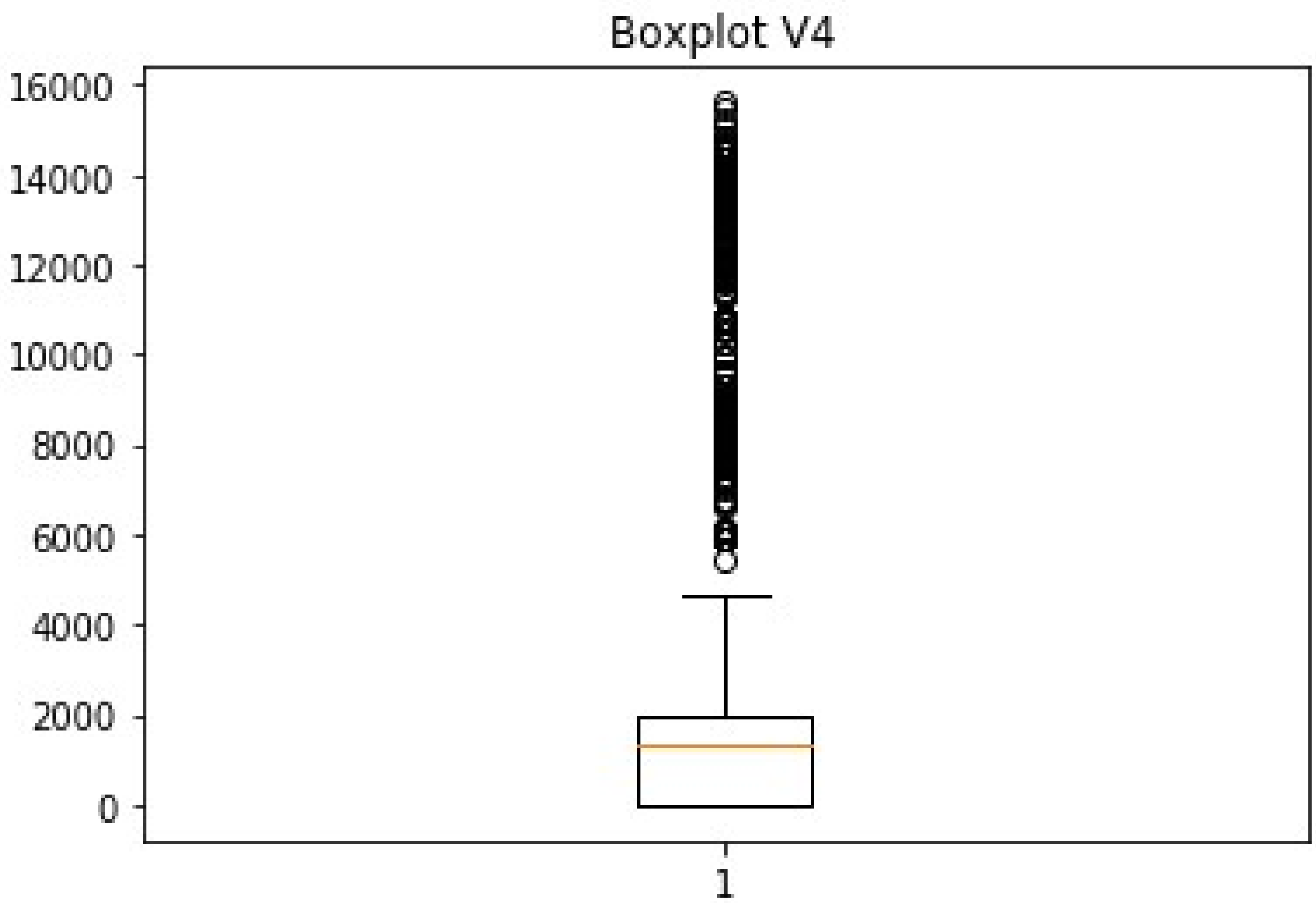
3.4 - Data understanding – Other Variables

Outlier study – V3



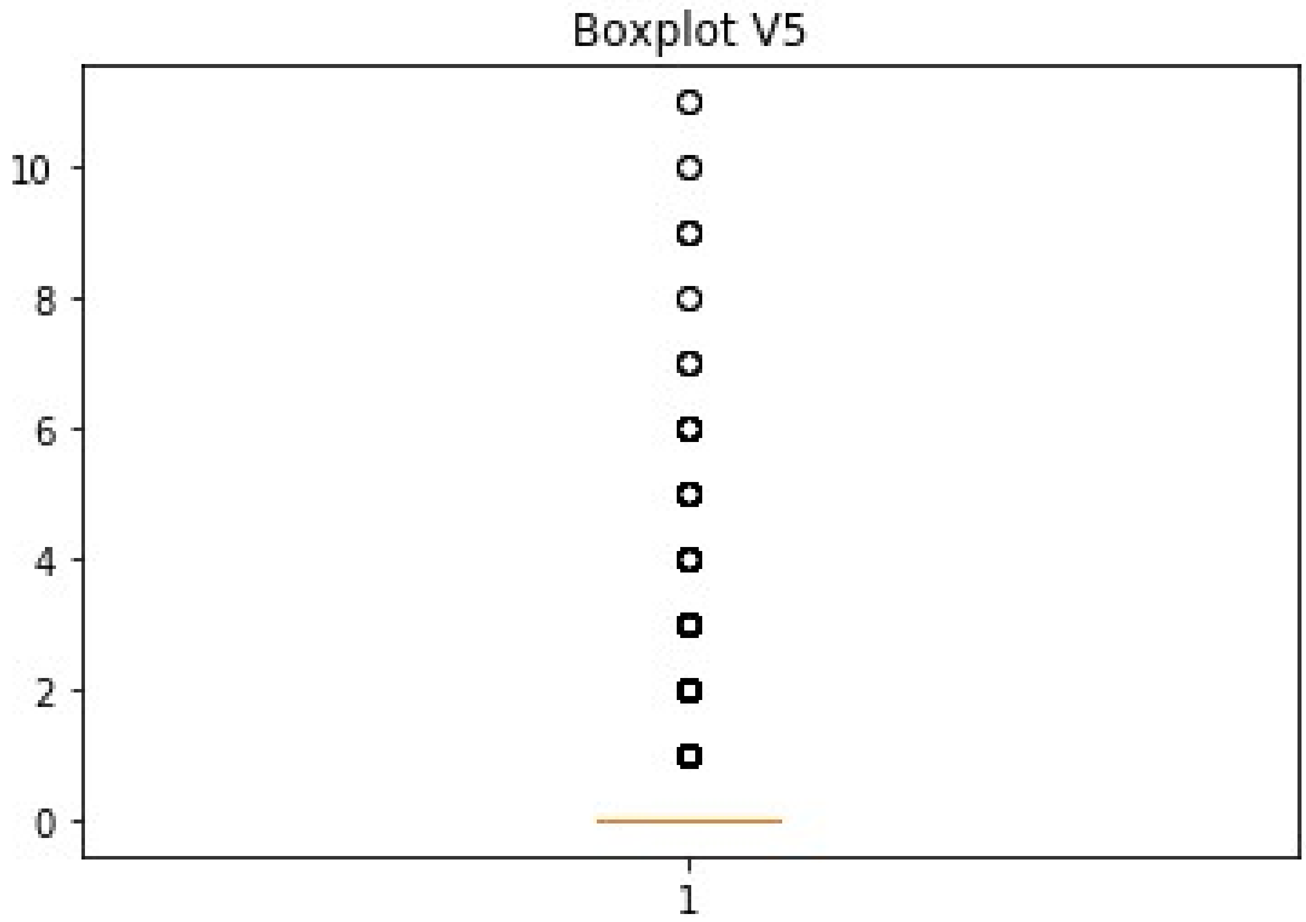
3.4 - Data understanding – Other Variables

Outlier study – V4



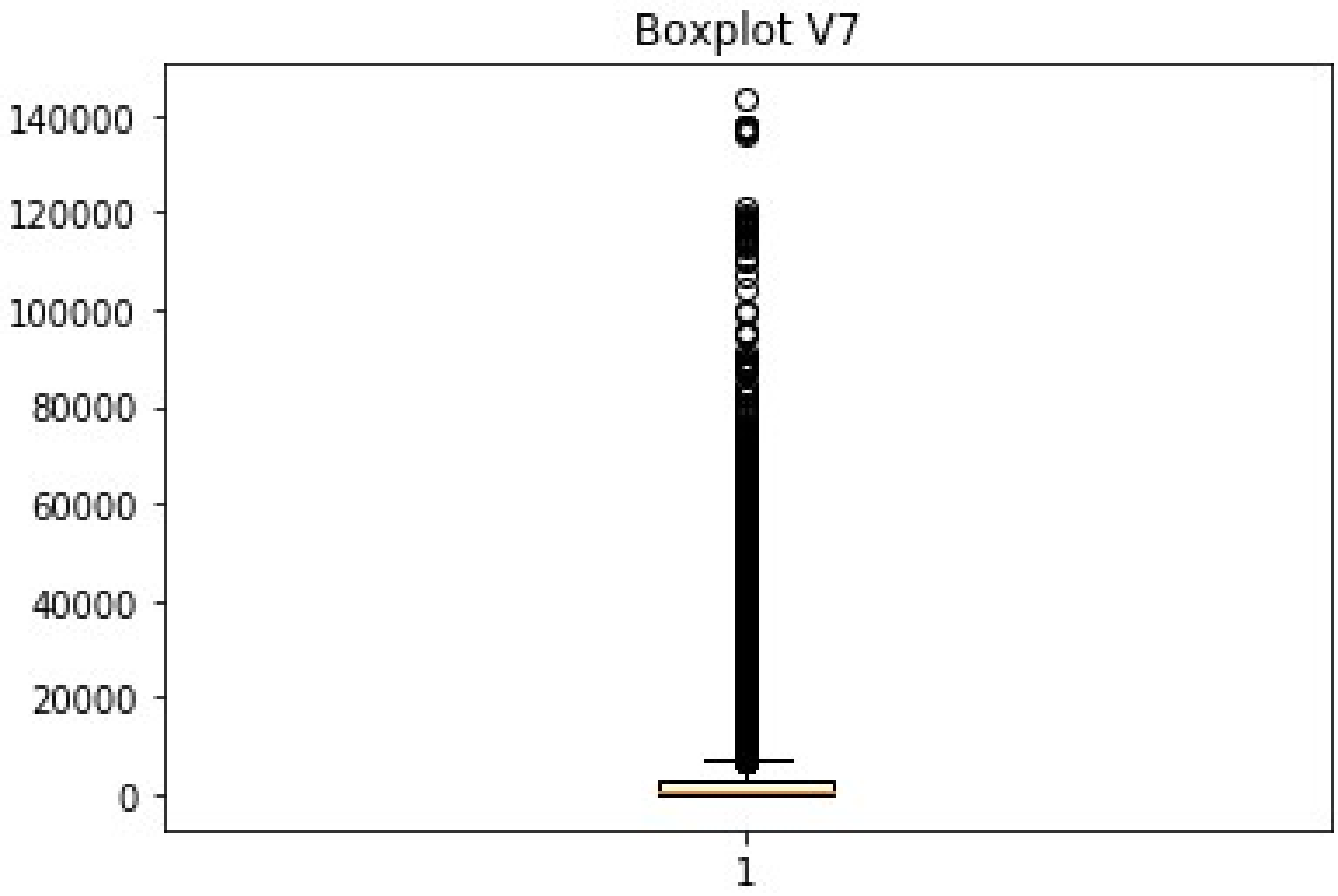
3.4 - Data understanding – Other Variables

Outlier study – V5



3.4 - Data understanding – Other Variables

Outlier study – V7



3.4 - Data understanding – Other Variables

The boxplot indicate that there are outliers, the Zscore teste confirm the outliers and identifies them.

Follow the Results

```
Amount of outlier in V2: 312
Amount of outlier in V3: 196
Amount of outlier in V4: 233
Amount of outlier in V5: 173
Amount of outlier in V7: 256
```

Due to the fact that the problem is to detect a rare events, outliers can be a clue.
In this scenario the outliers will not be remove form the dataset in the first CRISP-DM run, until the confirmation of how the outlier affects the models.

3.5 - Data understanding – Correlation

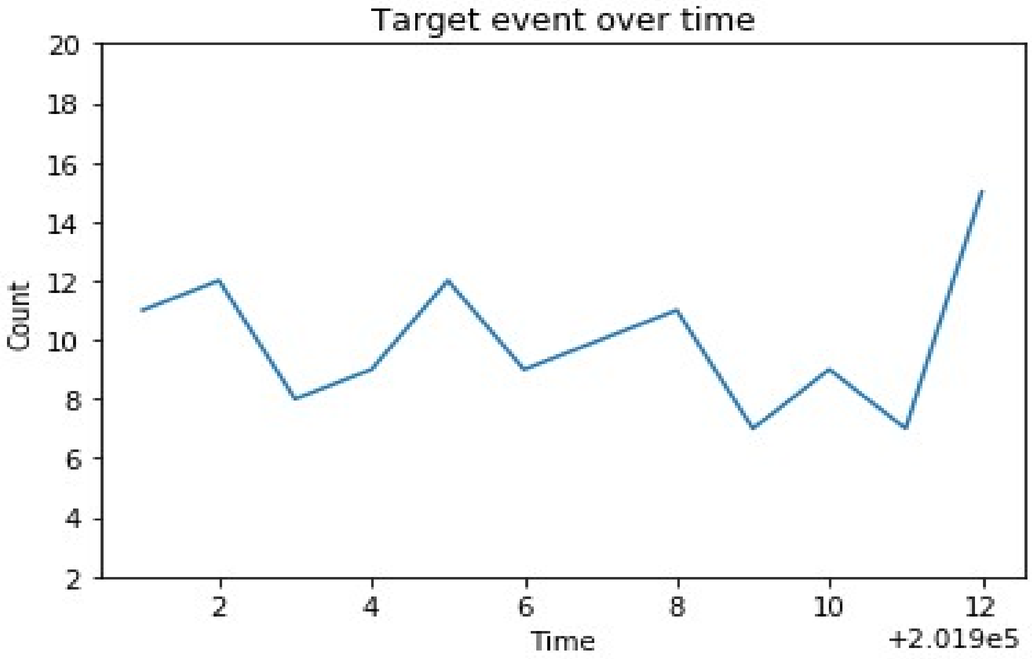
The Pearson correlation confirm there aren't a very strong correlation between the variables.

The variables in general have a weak or depressible correlation with the exception of variables V3 and V7 that has a strong correlation.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
V1	1.0	-0.13	0.11	0.012	0.073	0.052	-0.023	-0.14	0.028	0.0063
V2	-0.13	1.0	0.29	0.043	0.013	0.015	0.49	-0.12	0.011	0.051
V3	0.11	0.29	1.0	0.11	0.071	-0.033	0.81	-0.26	0.023	0.068
V4	0.012	0.043	0.11	1.0	0.011	-0.37	0.075	-0.049	0.021	0.08
V5	0.073	0.013	0.071	0.011	1.0	0.042	0.025	-0.19	0.034	0.073
V6	0.052	0.015	-0.033	-0.37	0.042	1.0	-0.018	0.019	-0.0023	-0.013
V7	-0.023	0.49	0.81	0.075	0.025	-0.018	1.0	-0.18	0.022	0.08
V8	-0.14	-0.12	-0.26	-0.049	-0.19	0.019	-0.18	1.0	-0.039	-0.039
V9	0.028	0.011	0.023	0.021	0.034	-0.0023	0.022	-0.039	1.0	0.018
V10	0.0063	0.051	0.068	0.08	0.073	-0.013	0.08	-0.039	0.018	1.0

3.6 - Data understanding – Target events over time

There a variation in the months - December has the twice amount of the targets events then September and November.
However it is not possible to affirm whether this is a phenomenon of seasonality or not, since there is only one year under analysis. In this case seasonality will not be considered as it may skew the models for futures datasets.



3.7 - Data understanding – Target x Other Variables

The following tests aim to verify the correlation of the Target variable X other variable and the relationship between outliers X Target events.

- The variables have a depressible correlation with the Target

	TARGET
TARGET	1.0
V1	0.04
V2	0.06
V3	0.058
V4	-0.01
V5	0.28
V6	0.063
V7	0.043
V8	-0.056
V9	0.057
V10	0.052

3.7 - Data understanding – Target x Other Variables

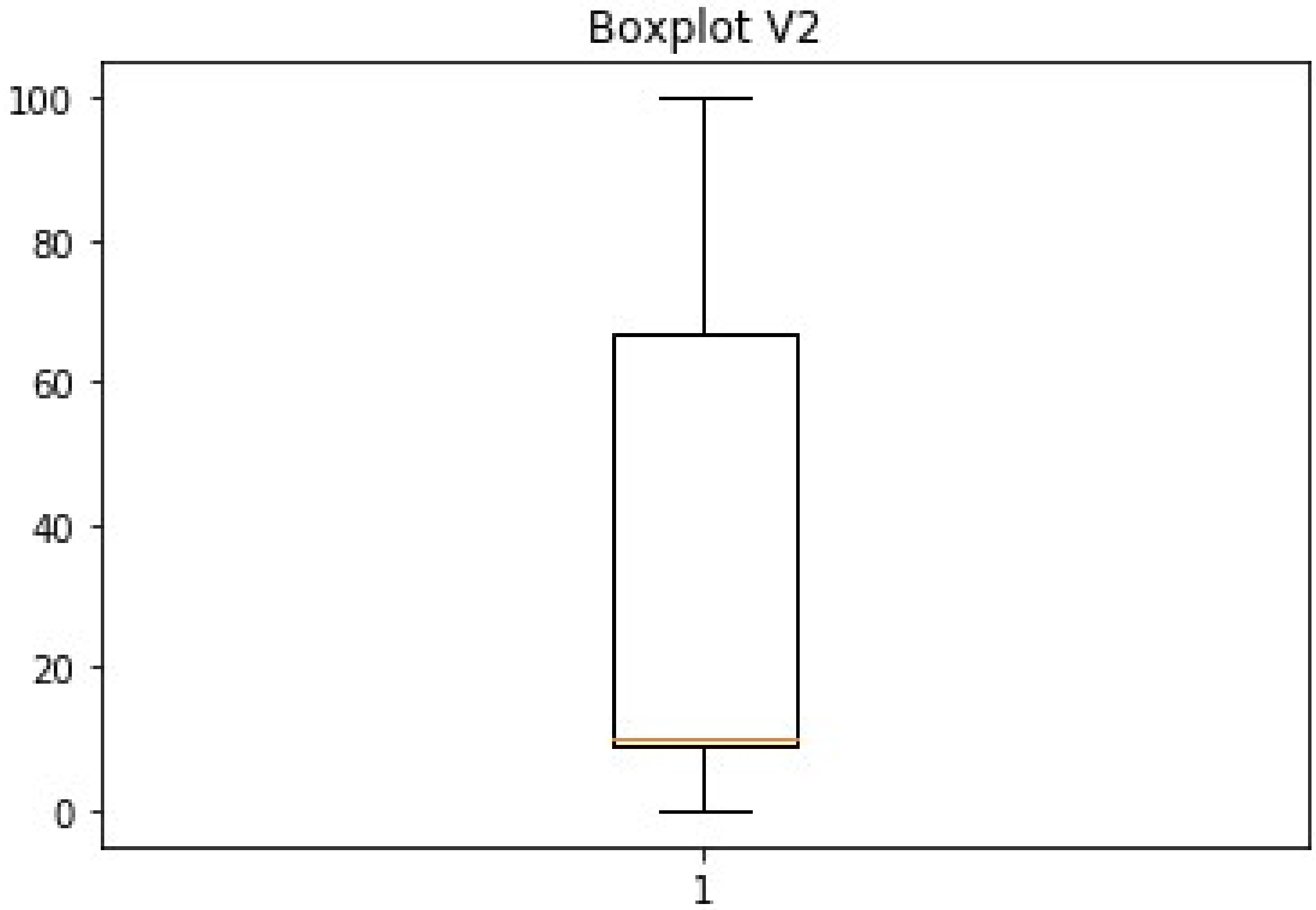
Behavior of the variables when the TARGET event is true:

	V1	V2	V3	V4	V5
count	120.000000	120.000000	120.000000	120.000000	120.000000
mean	0.225000	34.290833	1032.153167	1223.258333	1.933333
std	0.419333	36.252643	1269.059984	1917.909469	2.643315
min	0.000000	0.000000	6.250000	0.000000	0.000000
25%	0.000000	8.950000	181.247500	0.000000	0.000000
50%	0.000000	10.000000	505.000000	1069.500000	1.000000
75%	0.000000	66.850000	1038.937500	1929.500000	3.000000
max	1.000000	100.000000	5007.000000	12967.000000	11.000000

	V6	V7	V8	V9	V10
count	120.000000	120.000000	120.000000	120.000000	120.000000
mean	0.408333	9065.989167	0.133333	0.058333	0.116667
std	0.493586	16550.588803	0.341360	0.235355	0.322369
min	0.000000	7.560000	0.000000	0.000000	0.000000
25%	0.000000	636.490000	0.000000	0.000000	0.000000
50%	0.000000	3214.945000	0.000000	0.000000	0.000000
75%	1.000000	13146.530000	0.000000	0.000000	0.000000
max	1.000000	114814.190000	1.000000	1.000000	1.000000

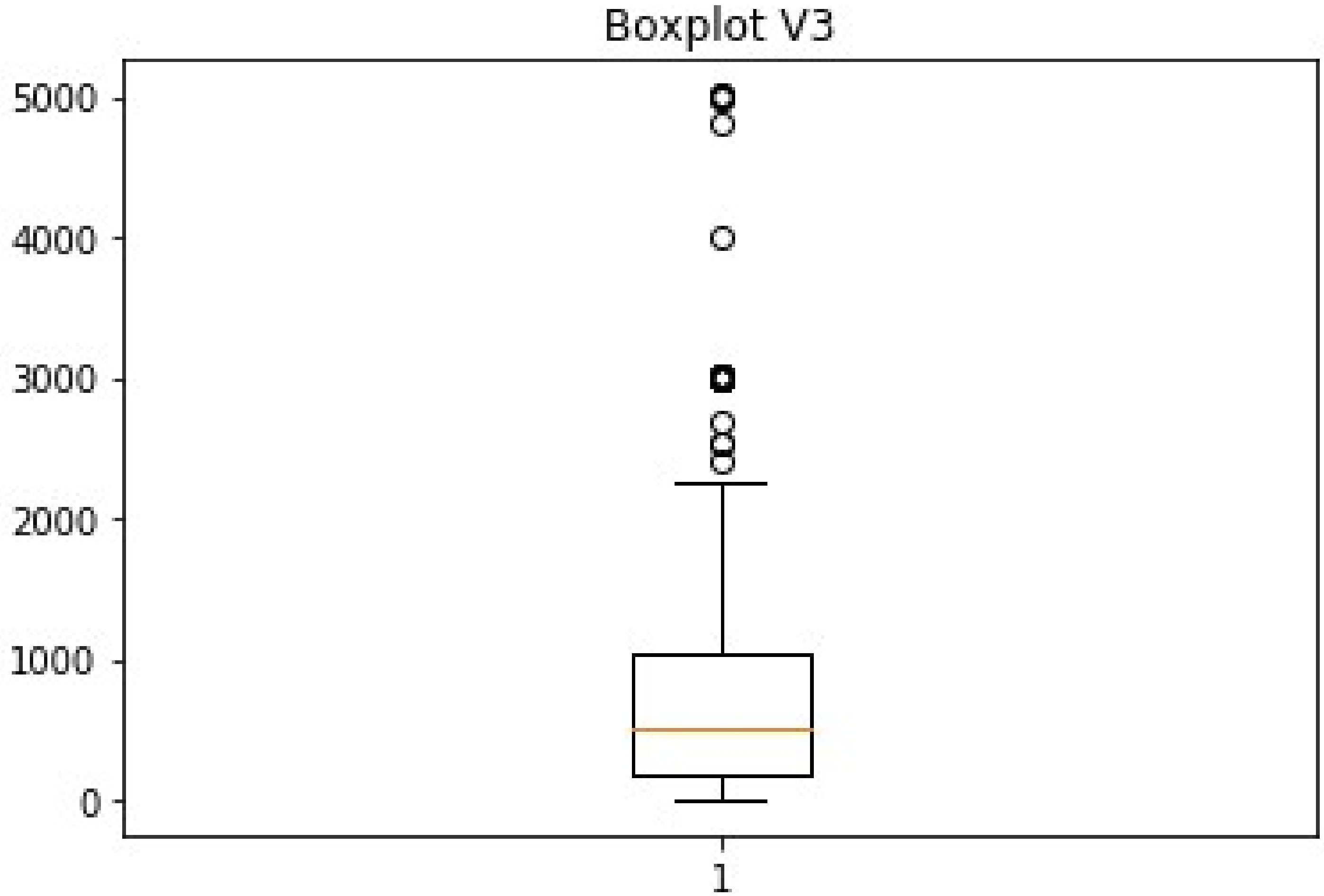
3.7 - Data understanding – Target x Other Variables

Outlier study – V2



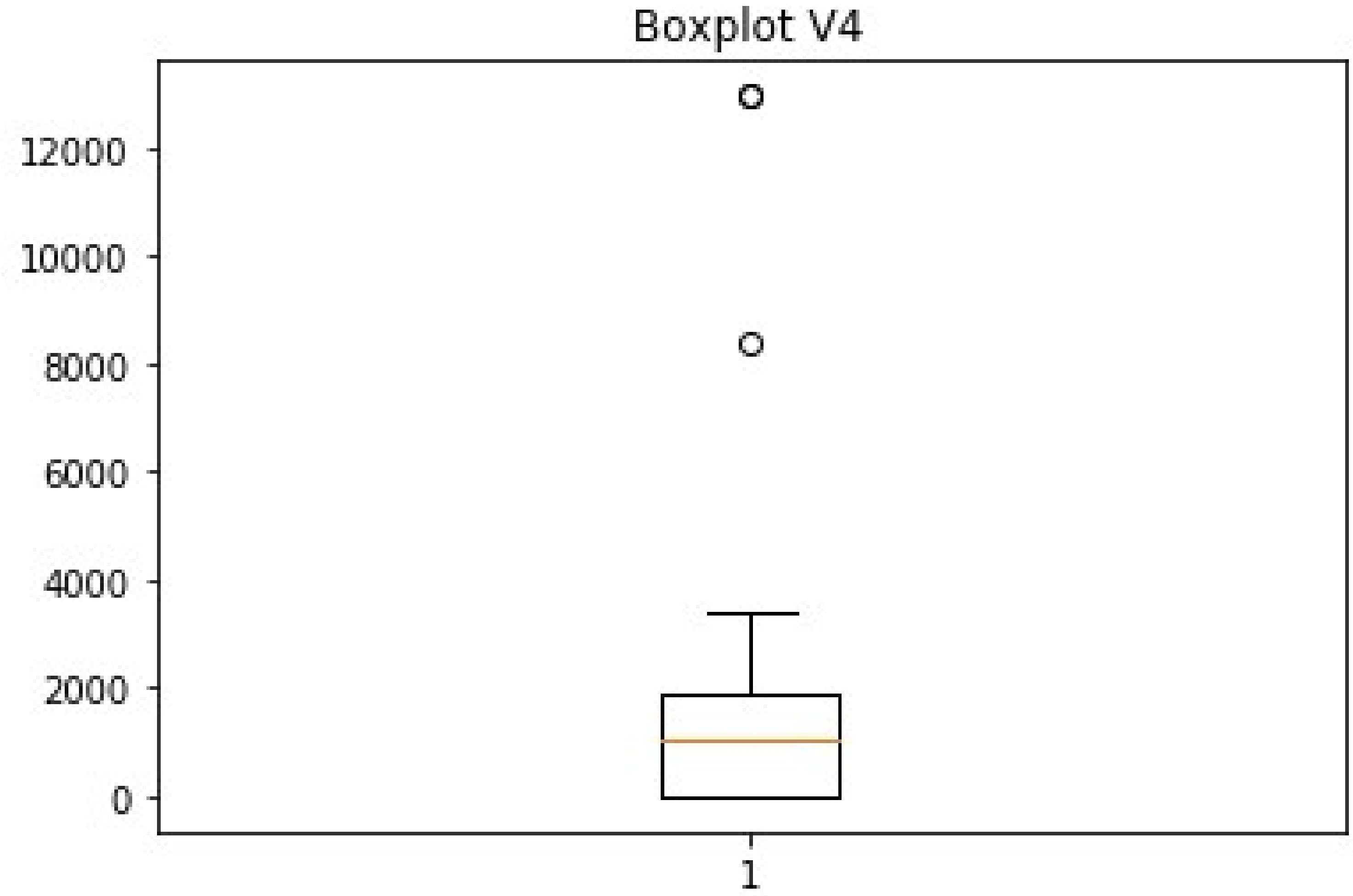
3.7 - Data understanding – Target x Other Variables

Outlier study – V3



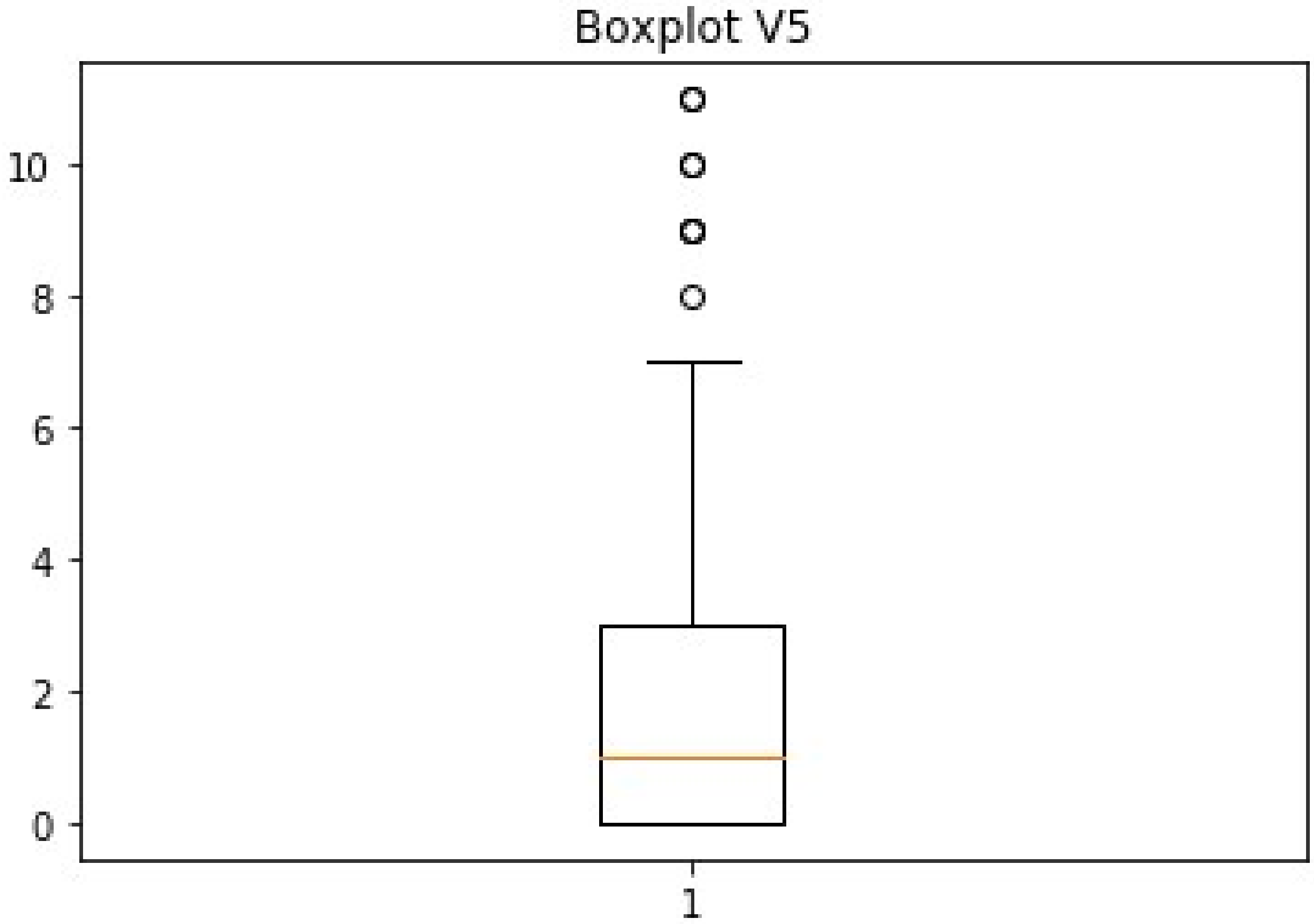
3.7 - Data understanding – Target x Other Variables

Outlier study – V4



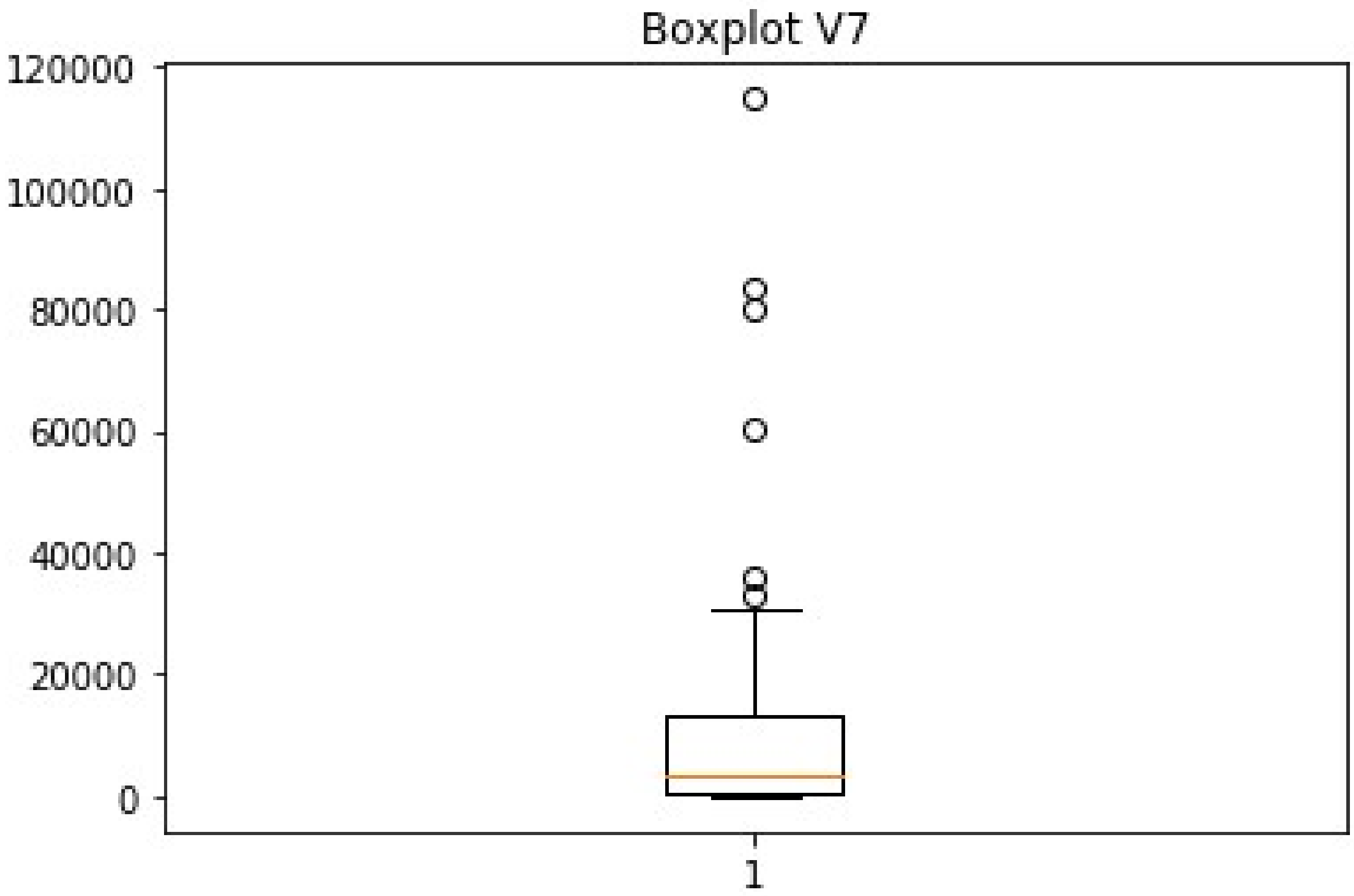
3.7 - Data understanding – Target x Other Variables

Outlier study – V5



3.7 - Data understanding – Target x Other Variables

Outlier study – V7



3.7 - Data understanding – Target x Other Variables

The boxplot indicate the number of outliers was affected and the Zscore teste confirm the outliers and identifies them.

Follow the results:

```
Amount of outlier in V2: 0
Amount of outlier in V3: 4
Amount of outlier in V4: 3
Amount of outlier in V5: 4
Amount of outlier in V7: 4
```

The significant reduction in the number of outlier indicates the outlier relationship with the Target event is not highlighted.

However, the ouliers will not be removed form the dataset since they represent 12.5% of the dataset of the target events. An expressive number of Target events to be discarded in the first run.

4 - Dataset – overview

Due to the lack of business successful metrics, the F1 score metric will be used as successful metric, with the formula:

$$F1 = 2TP / (2TP + FP + FN)$$

Where:
TP = True positive
TN = True negative
FP = False positive
FN = False negative

4 - Dataset – overview

Other metrics for comparation and study effects.

$$\text{Precision} = TP / (TP + FP)$$
$$\text{Recall} = TP / (TP + FN)$$
$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

For the model algorithms, the chosen techniques are:

- Logistic Regression
- Neural Networks
- Random Forest

Because they are recognized techniques in the academic literature, to treat cases of rare events.

5 - Data preparation

Data split

The data set will be separated into 3 groups: training, validation and testing.

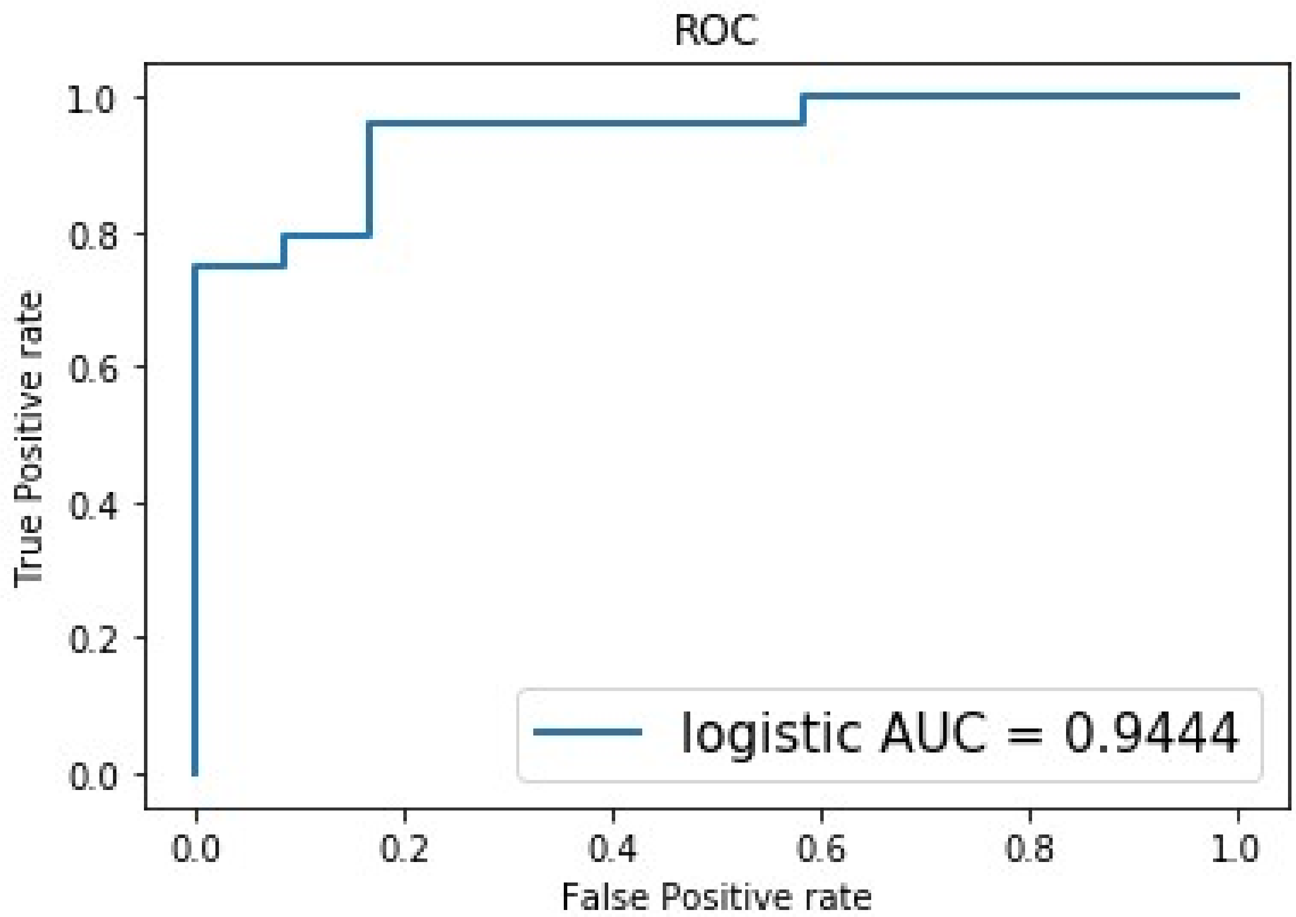
•60% Training
•20% Validation
•20% Test

Unbalanced data

The Undersampling technique are a recognized technique in the academic literature to treat cases of rare events.
However this will make the training set small with help in the training time, but can affect the useful to create a robust model.

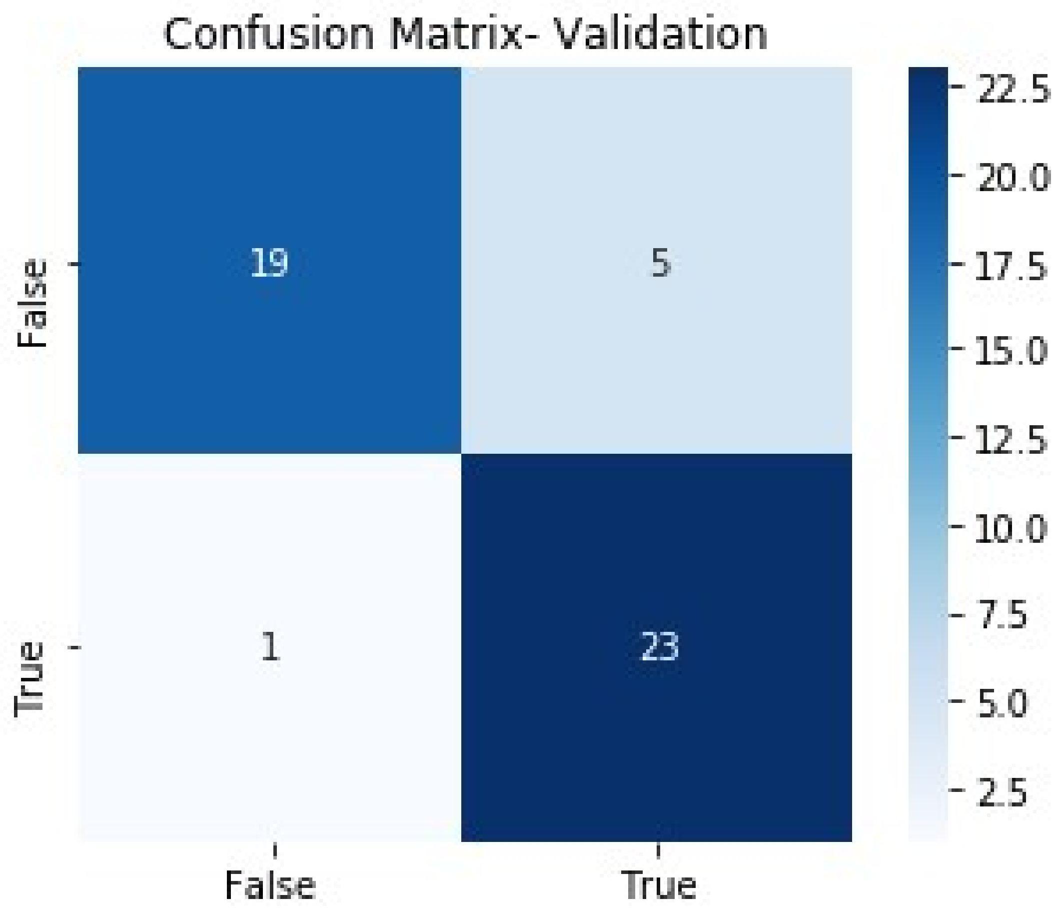
6.1 - Modeling – Logistic Regression

Logistic Regression training results.



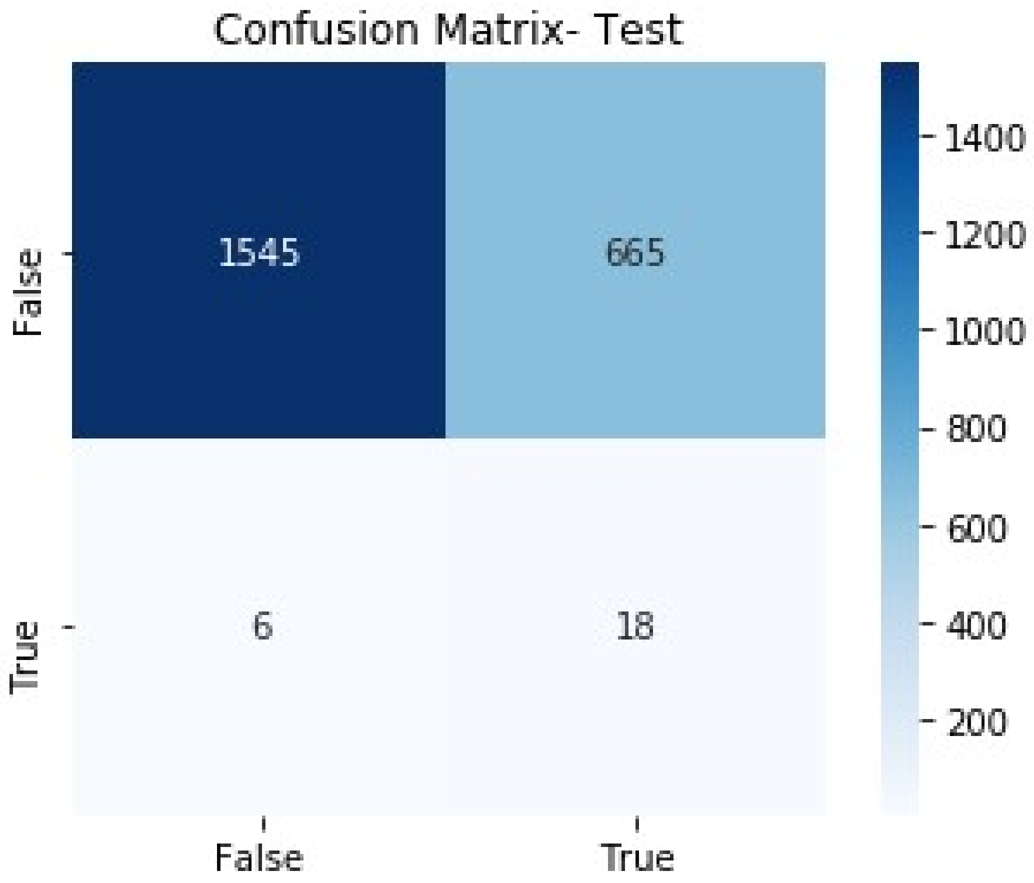
6.1 - Modeling – Logistic Regression

Logistic Regression training results.



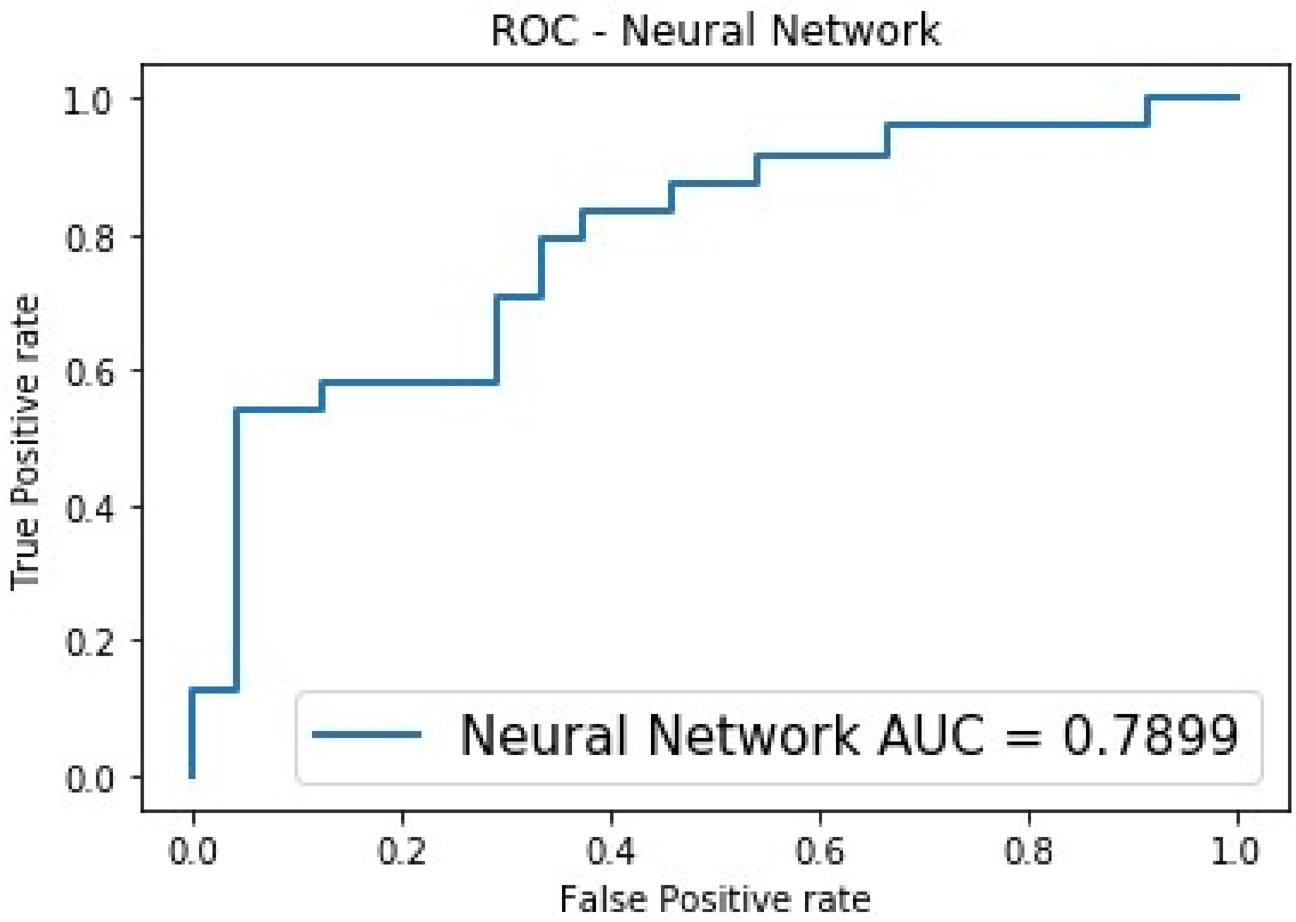
6.1 - Modeling – Logistic Regression

Logistic Regression training results.



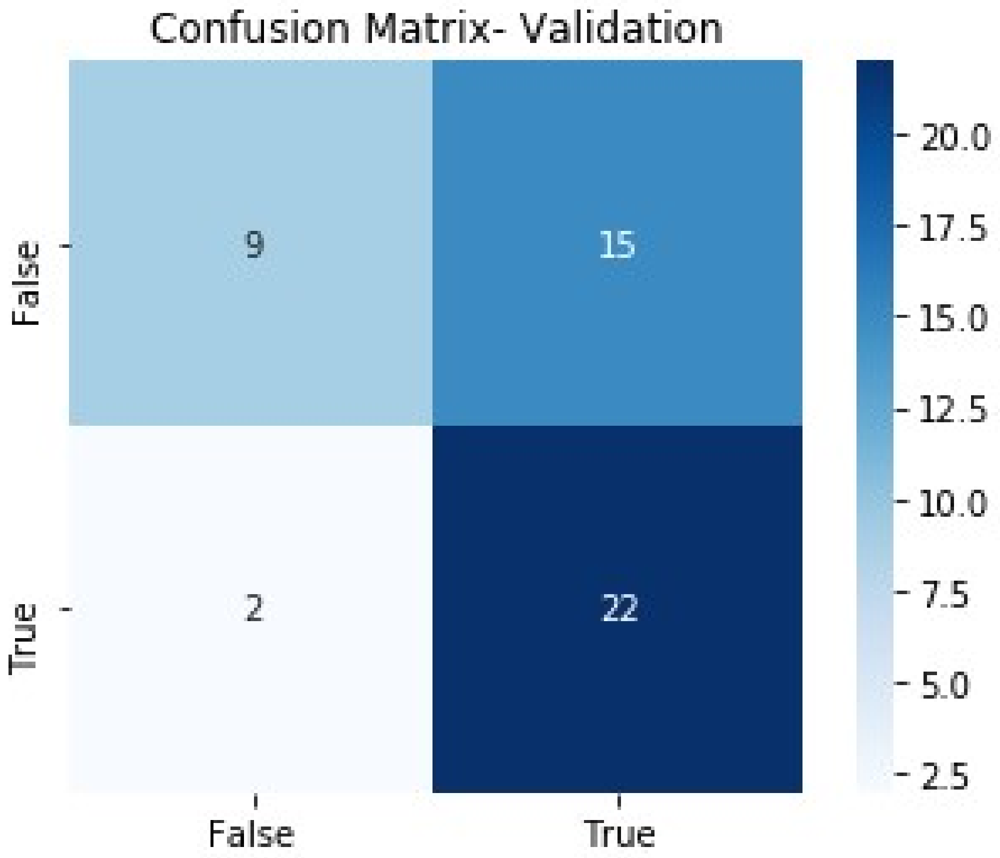
6.2 - Modeling – Neural Network

Neural network training results.



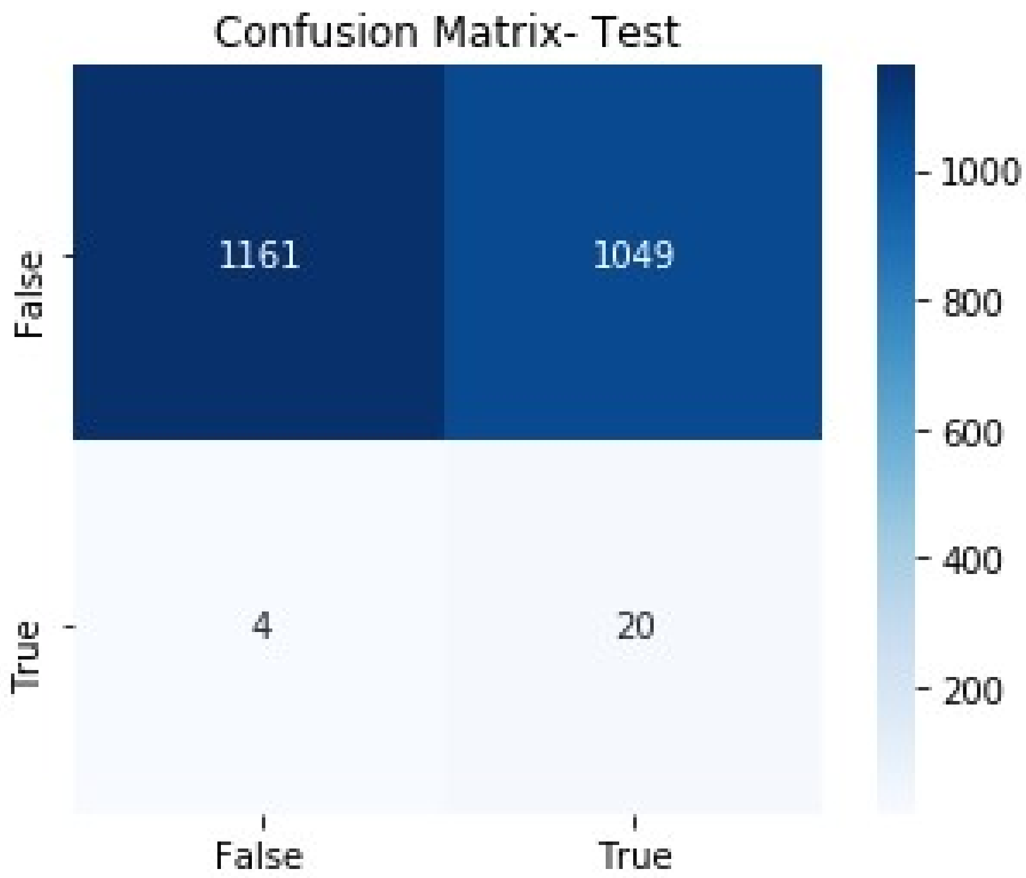
6.2 - Modeling –   Neural Network

Neural network training results.



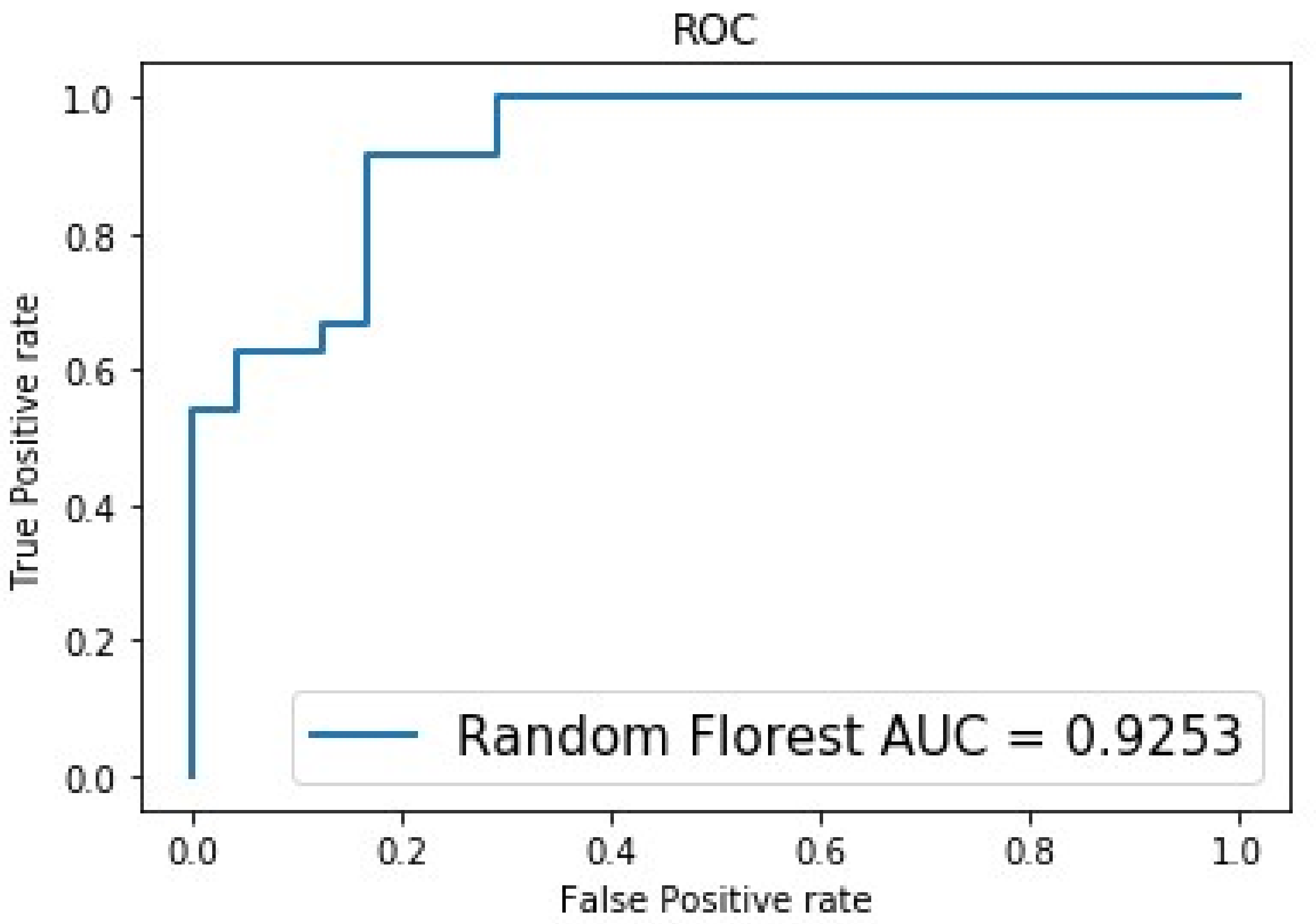
6.2 - Modeling – Neural Network

Neural network training results.



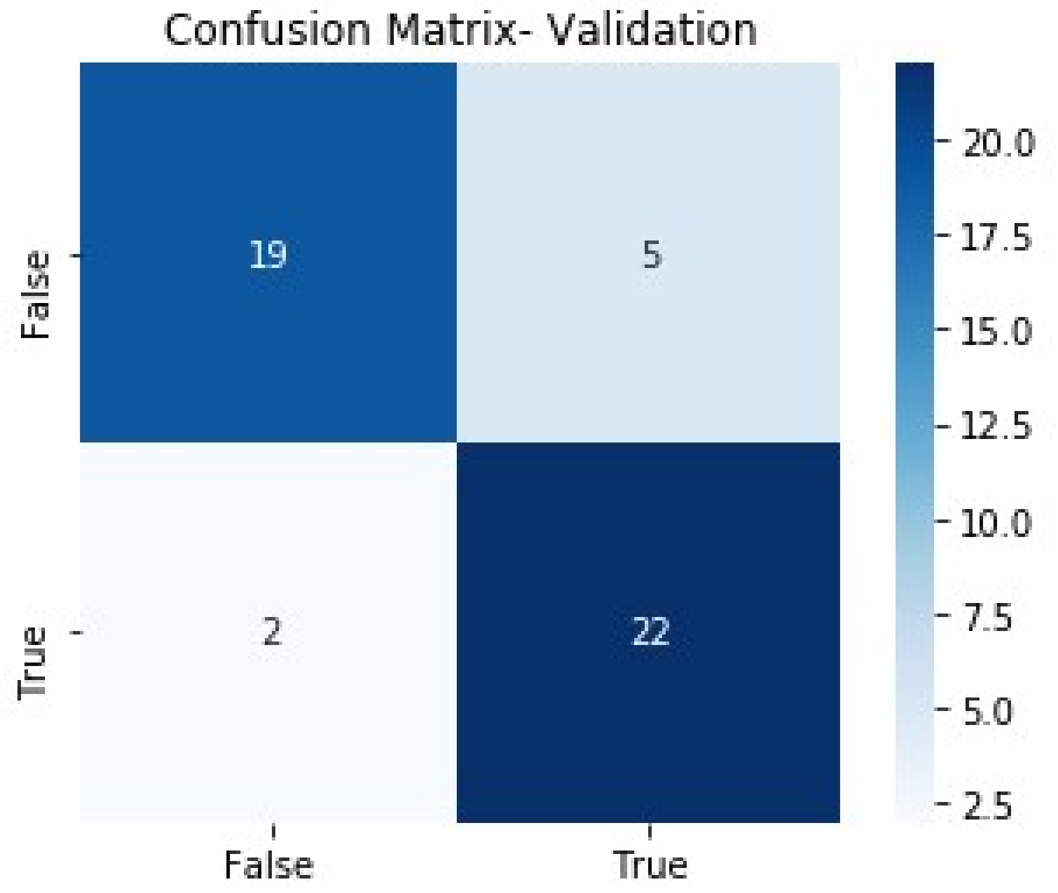
6.3 - Modeling – Random Forest

Random forest training results.



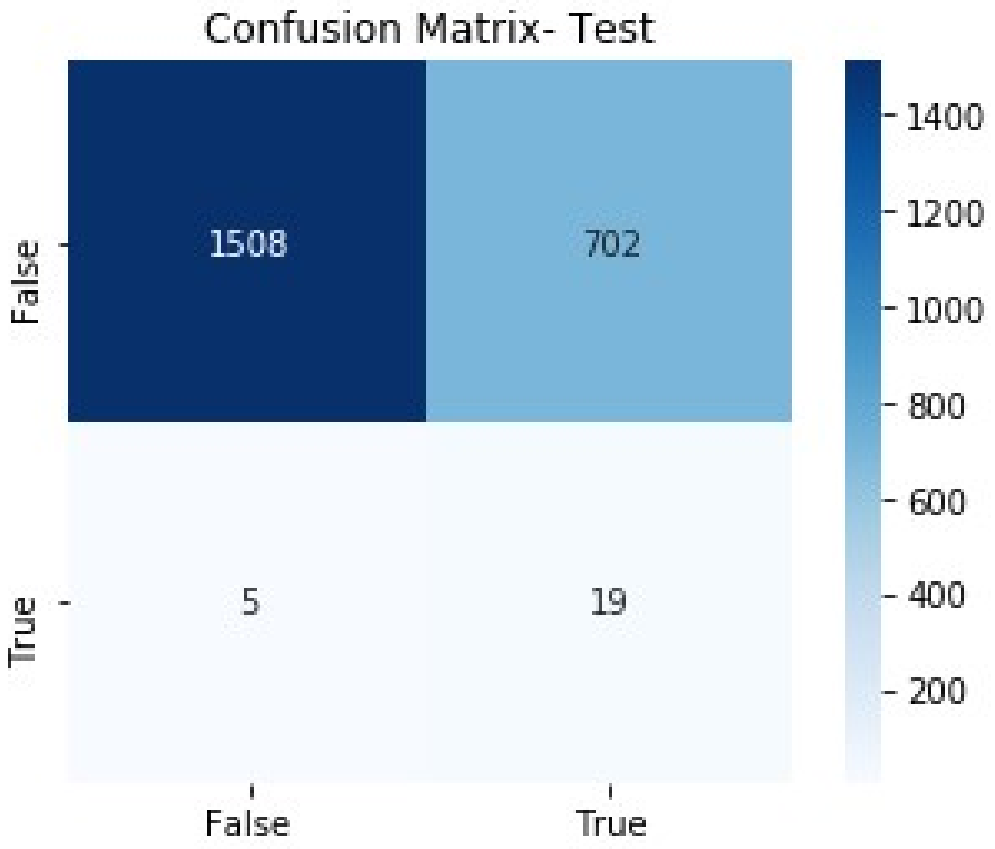
6.3 - Modeling – Random Forest

Random forest training results.



6.3 - Modeling – Random Forest

Random forest training results.



6.4 - Modeling – Review

Results comparation

Validation result

	Model	Best Threshhold	F1 Score	Accuracy	Recall	Precision
0	logistic	0.367347	0.884615	0.875000	0.958333	0.821429
0	NN	0.473684	0.721311	0.645833	0.916667	0.594595
0	RF	0.368421	0.872727	0.854167	1.000000	0.774194

Test result

	Model	F1 Score	Accuracy	Recall	Precision
0	logistic regression	0.050919	0.699642	0.750000	0.026354
1	MM	0.036597	0.528648	0.833333	0.018709
2	RF	0.051007	0.683527	0.791667	0.026352

6.4 - Modeling – Review

The analyze of the training results show a high precision and sensitivity rates, highlighting the techniques of Reverse Logistics and Random Forest, with an F1 score of 0.88 and 0.87.

However, when using the fit models on an unbalanced test base, there are a large drop in F1 values and a significative drop in precision score.

The reason may be the non-removal of outliers and the use of undersample technique for balancing the dataset.

Depending on the business problem the drop in the scores, mainly the precision score could invalidate using the model in production. It is suggest in the next step retraining the models using other balancing techniques, like SMOTE and removing the outliers.

7 – Reference

- https://towardsdatascience.com/under-the-hood-logistic-regression-407c0276c0b4
- https://towardsdatascience.com/detecting-financial-fraud-using-machine-learning-three-ways-of-winning-the-war-against-imbalanced-a03f8815cce9
- https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8
- https://towardsdatascience.com/credit-card-fraud-detection-9bc8db79b956
- http://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/27166/Dissertacao_Joao_Carlos_Pacheco_VFinal_2.pdf?sequence=3&isAllowed=y
- https://medium.com/omixdata/estat%C3%ADstica-an%C3%A1lise-de-regress%C3%A3o-linear-e-an%C3%A1lise-de-regress%C3%A3o-log%C3%ADstica-com-r-a4be254df106
- https://medium.com/towards-artificial-intelligence/credit-card-fraud-prediction-using-machine-learning-f47e52a0dbc2
- https://www.kaggle.com/marcelotc/creditcard-fraud-logistic-regression-example
- https://towardsdatascience.com/eveything-you-need-to-know-about-interpreting-correlations-2c485841c0b8
- https://towardsdatascience.com/credit-card-fraud-detection-a1c7e1b75f59

7 – Reference

•https://towardsdatascience.com/linear-regression-under-the-hood-583003d0bf38
•https://towardsdatascience.com/real-time-fraud-detection-with-machine-learning-485fa502087e
•sciencedirect.com/science/article/abs/pii/S0167923617300027
•https://link.springer.com/article/10.1007/s10479-008-0371-9
•researchgate.net/profile/Saravanan_Sagadevan2/publication/326986162_Credit_Card_Fraud_Detection_Using_Machine_Learning_As_Data_Mining_Technique/links/5b70a251a6fdcc87df733637/Credit-Card-Fraud-Detection-Using-Machine-Learning-As-Data-Mining-Technique.pdf
•https://www.researchgate.net/profile/Yo-Ping_Huang/publication/4073793_Survey_of_fraud_detection_techniques/links/541771590cf203f155ad5825/Survey-of-fraud-detection-techniques.pdf
•https://www.ripublication.com/ijaer19/ijaerv14n2_08.pdf