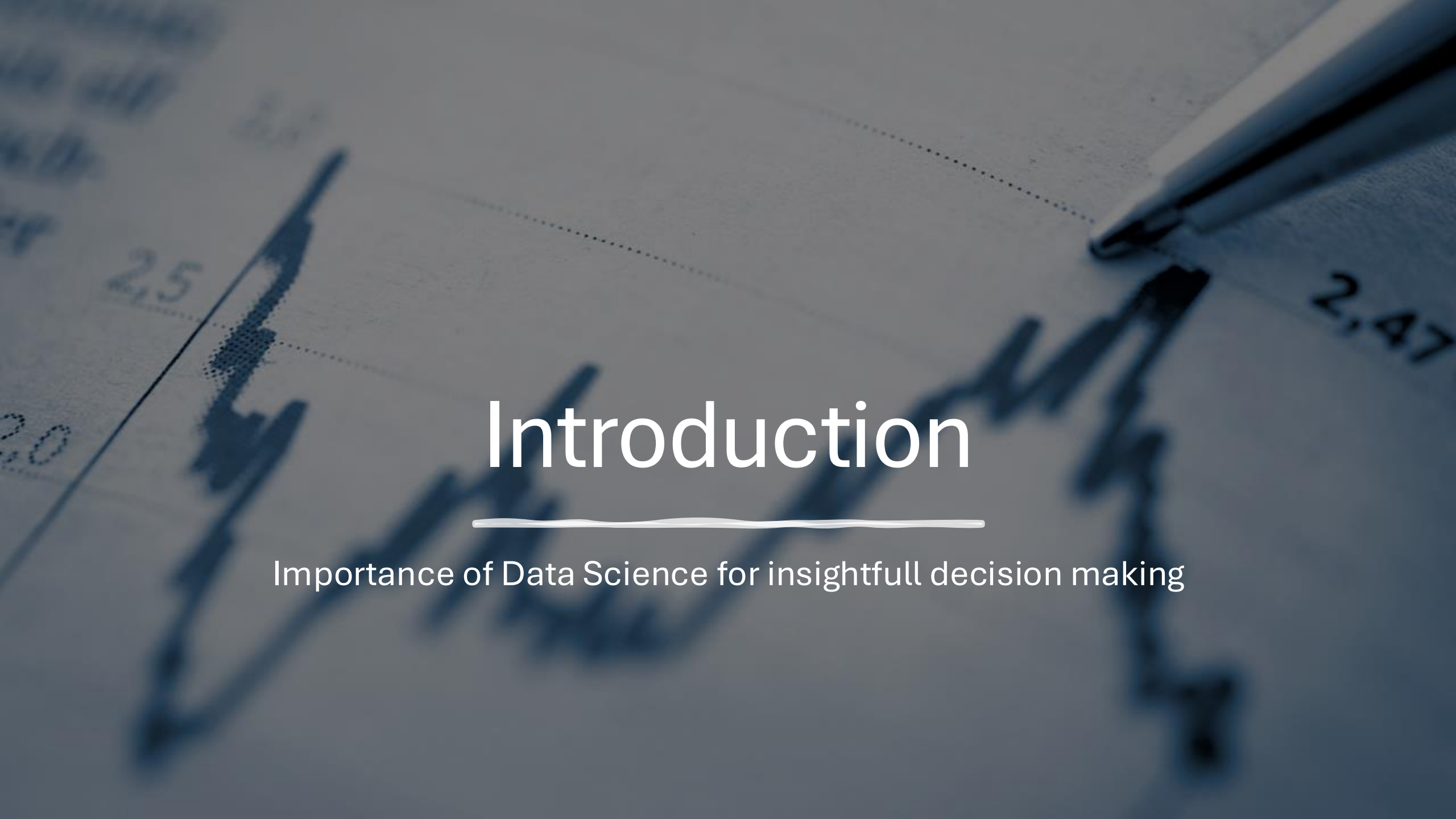
Credit Card Fraud
Transaction detection System

My Paypal



Introduction

Importance of Data Science for insightfull decision making

Data collection and exploration

The data we used in this project is an open source dataset containing 28 variable resulting of the PCA transformation in order to maintain customer privacy.

Here is a brief view of the variables in the dataset:

2- Data Exploration

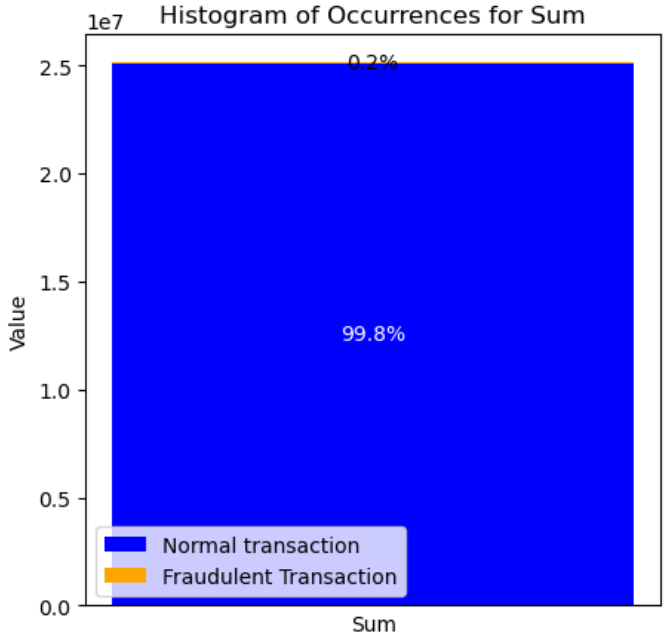
```
df.head()
```

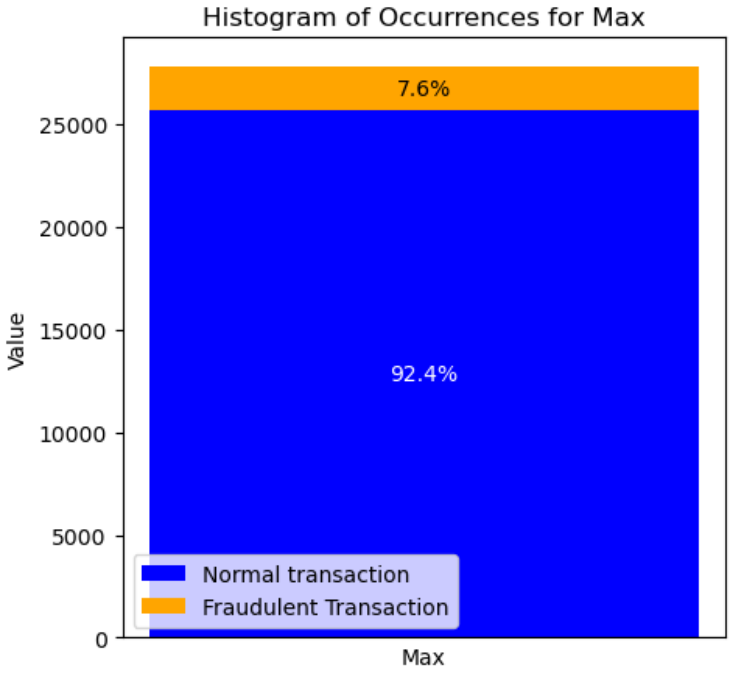
	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	149.62	0
1	0.0	1.191857	0.266151	0.166480	0.448154	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

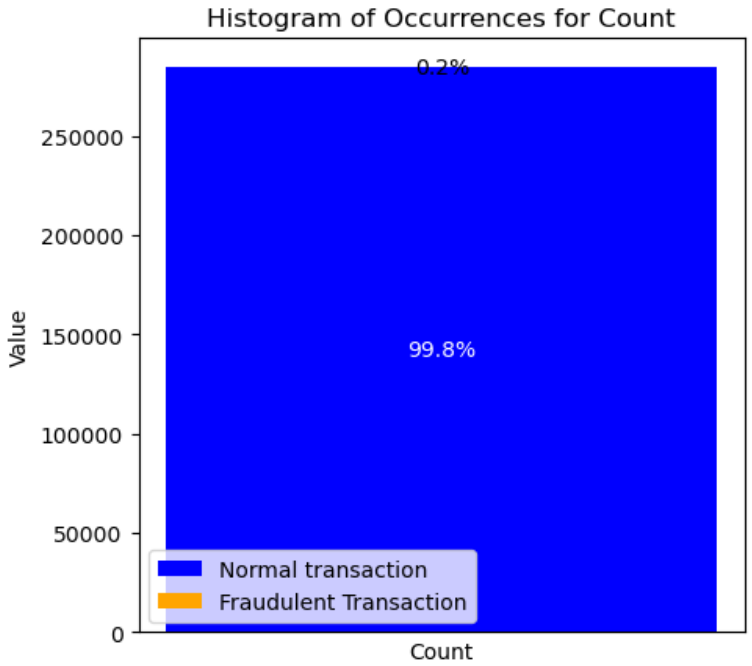
5 rows × 31 columns

Data Visualization

Target variable distribution and Correlation Heatmap







1- By analysing the above table, we noticed the number of fraudulent transaction are drasticaly inferior comparing to the the normal transaction which suggest an unbalanced repartition on the class. This situation can affect our model during its learning process.

2-The maximun amount of frandulant transaction in lower than the maximun amount of normal transaction. We conclue that a frandulant transaction is not necessarly a very High amount transaction.

Machine Learning Model

Best ML Model

Since our dataset is highly imbalanced, we perform couple of machine learning techniques in order to have a powerful model able to detect fraud easy and accurately. We used the following ensemble models from imblearn.ensemble : **BalancedRandomForestClassifier , EasyEnsembleClassifier**

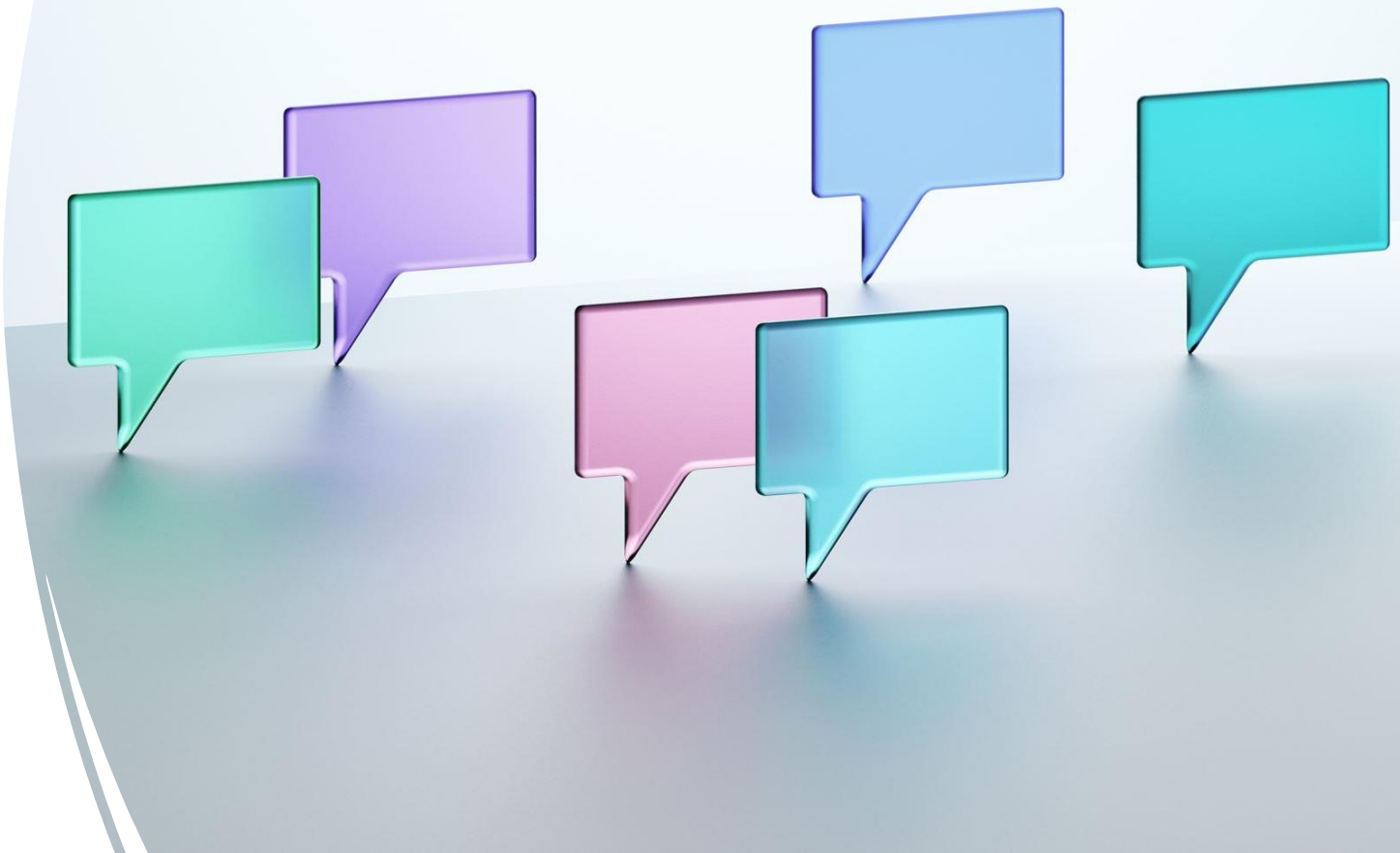
Interpreation

•Since our dataset is highly imbalanced the more informatives metrics are Precisions-Recall AUC.

•`Precision-Recall` AUC: Both models (0.83-> EassyEnsembleClassifier , 0.82 -> BalancedRandomForestClassifier) perform well in distinguishing between the classes,particularly in the minority class(positive class)

```
Precision-Recall AUC: 0.83
              precision    recall  f1-score   support

           0       1.00      0.97      0.98     56860
           1       0.05      0.87      0.09       102

    accuracy                           0.97     56962
   macro avg       0.52      0.92      0.54     56962
weighted avg       1.00      0.97      0.98     56962

Model: BalancedRandomForestClassifier
Precision-Recall AUC: 0.82
              precision    recall  f1-score   support

           0       1.00      0.98      0.99     56860
           1       0.08      0.88      0.14       102

    accuracy                           0.98     56962
   macro avg       0.54      0.93      0.57     56962
```

COMMUNICATION



Conclusion

- Overall we use two Machine learnig models that perform well with classification on imbalanced dataset. And instead of using accuracy to evaluate the performance of our model we used Area Under the Curve (AUC) which is a more informative metrics for imbalanced dataset.
- 83% AUC show that our model is good at predicting the positive class. In this case at identifying if a transaction is a fraudulent one.