Credit Card Fraud Detection

PIB3
Buu NGUYEN
Anh-Thu DOAN
Casper CORNELIS

Agenda

01 Introduction

02 Methods

03 Results & Discussion

04 Conclusion

01

Introduction

- An inclusive term for fraud committed using payment card.
- Can happen through skimming or copying the card details.

- The data was collected from over 285,000 anonymized transactions made by credit cards in September 2013 by European cardholders
- Features V1 to V28 were the principal components obtained with PCA.
- Time, Amount, Class

	V2	V3	V4	V5	V6	V7	V8
35980713	-0.07278117	2.53634674	1.37815522	-0.338320770	0.462387778	0.239598554	0.0986
19185711	0.26615071	0.16648011	0.44815408	0.060017649	-0.082360809	-0.078802983	0.0851
35835406	-1.34016307	1.77320934	0.37977959	-0.503198133	1.800499381	0.791460956	0.2476
96627171	-0.18522601	1.79299334	-0.86329128	-0.010308880	1.247203168	0.237608940	0.3774
15823309	0.87773675	1.54871785	0.40303393	-0.407193377	0.095921462	0.592940745	-0.2705
42596588	0.96052304	1.14110934	-0.16825208	0.420986881	-0.029727552	0.476200949	0.2603
22965763	0.14100351	0.04537077	1.20261274	0.191880989	0.272708123	-0.005159003	0.0812
64426944	1.41796355	1.07438038	-0.49219902	0.948934095	0.428118463	1.120631358	-3.8078
89428608	0.28615720	-0.11319221	-0.27152613	2.669598660	3.721818061	0.370145128	0.8510
33826175	1.11959338	1.04436655	-0.22218728	0.499360806	-0.246761101	0.651583206	0.0695
44904378	-1.17633883	0.91385983	-1.37566665	-1.971383165	-0.629152139	-1.423235601	0.0484
38497822	0.61610946	-0.87429970	-0.09401863	2.924584378	3.317027168	0.470454672	0.5382
24999874	-1.22163681	0.38393015	-1.23489869	-1.485419474	-0.753230165	-0.689404975	-0.2274
06937359	0.28772213	0.82861273	2.71252043	-0.178398016	0.337543730	-0.096716862	0.1159
79185477	-0.32777076	1.64175016	1.76747274	-0.136588446	0.807596468	-0.422911390	-1.9071
75241704	0.34548542	2.05732291	-1.46864330	-1.158393680	-0.077849829	-0.608581418	0.0036
10321544	-0.04029621	1.26733209	1.28909147	-0.735997164	0.288069163	-0.586056786	0.1893
43690507	0.91896621	0.92459077	-0.72721905	0.915678718	-0.127867352	0.707641607	0.0879
40125766	-5.45014783	1.18630463	1.73623880	3.049105878	-1.763405574	-1.559737699	0.1608
49293598	-1.02934573	0.45479473	-1.43802588	-1.555434101	-0.720961147	-1.080664130	-0.0531
69488478	-1.36181910	1.02922104	0.83415930	-1.191208794	1.309108819	-0.878585911	0.4452
96249607	0.32846103	-0.17147905	2.10920407	1.129565571	1.696037686	0.107711607	0.5215
16661638	0.50212009	-0.06730031	2.26156924	0.428804195	0.089473517	0.241146580	0.1380

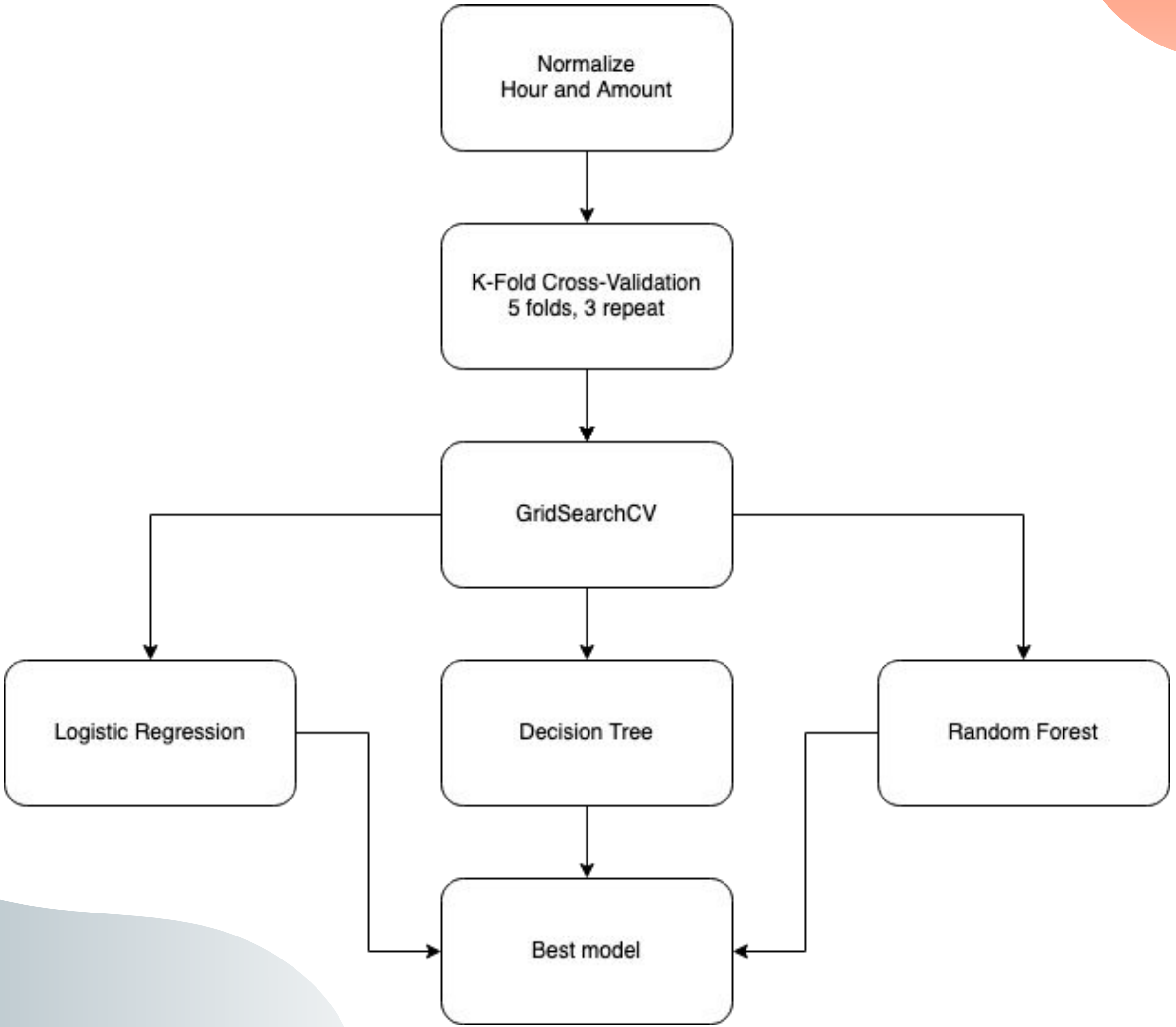
02 Method:

- R and its libraries (tidyverse, ggplot2, etc.) were used to explore the interactions and correlations between features then visualize them.

- Python were used to create a machine learning pipeline in this study.

F2 score was the primary metric of our models.

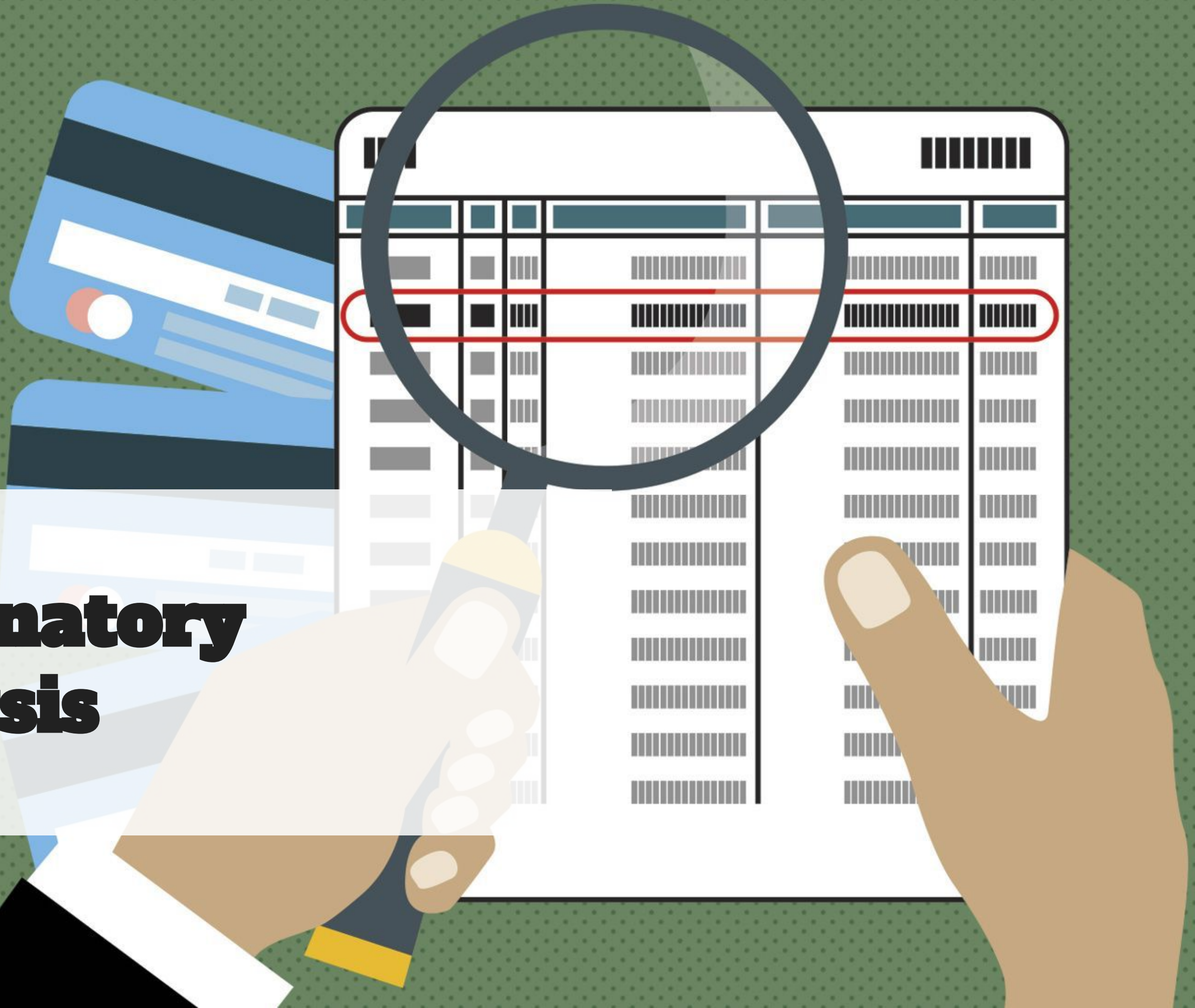
$$F_2 = \frac{TP}{TP + 0.2FP + 0.8FN}$$


Normalize
Hour and Amount
K-Fold Cross-Validation
5 folds, 3 repeat
GridSearchCV
Logistic Regression
Decision Tree
Random Forest
Best model

03 Results & Discussion

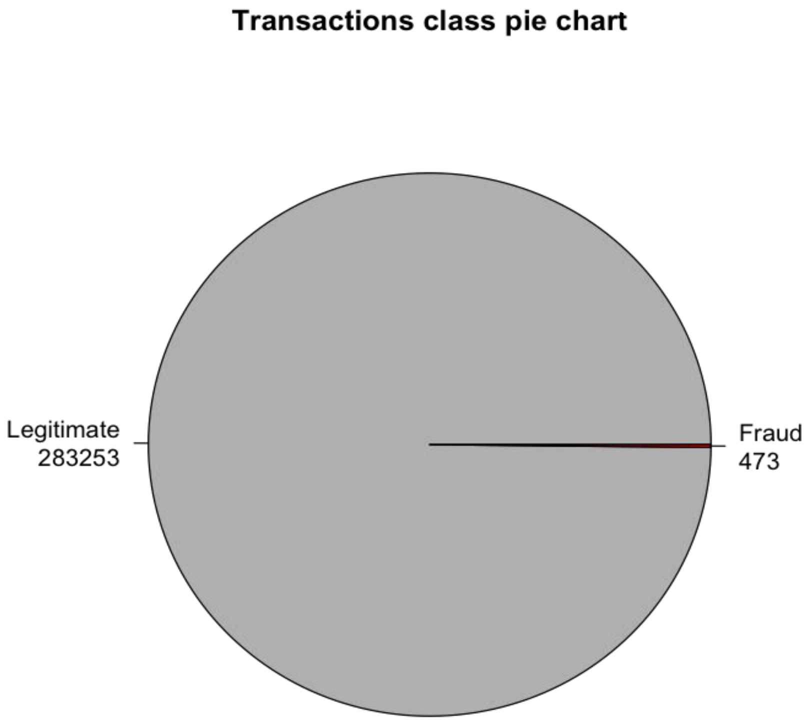
- Explanatory Analysis
- Classification Models

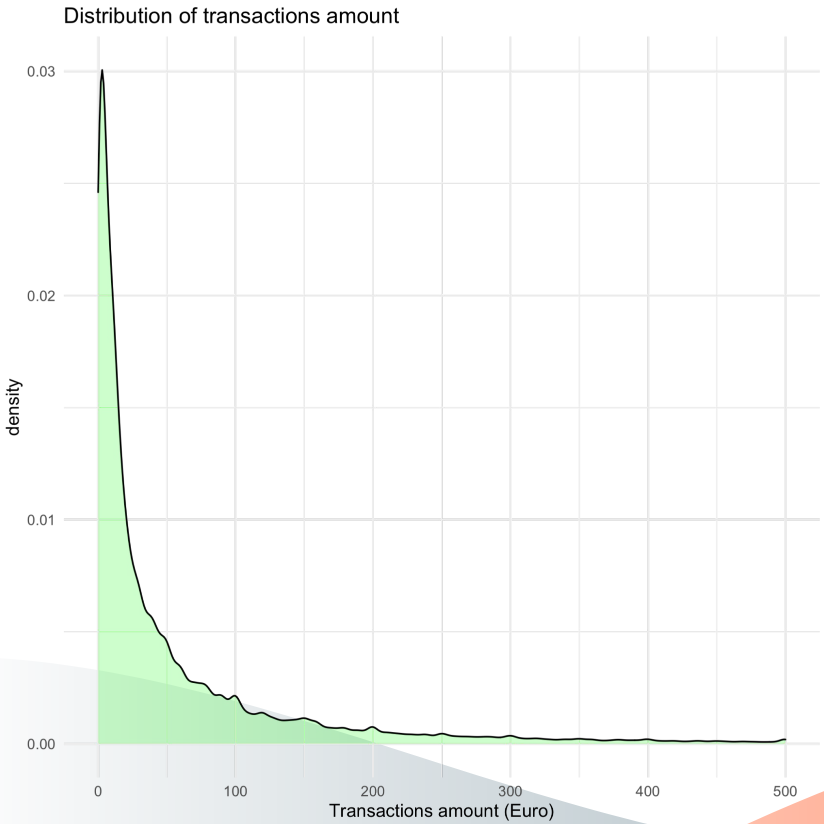
Explanatory Analysis



- The majority of the samples are legitimate transactions.
- The illegal transactions only held 0.17%

➢ This dataset is very unbalanced

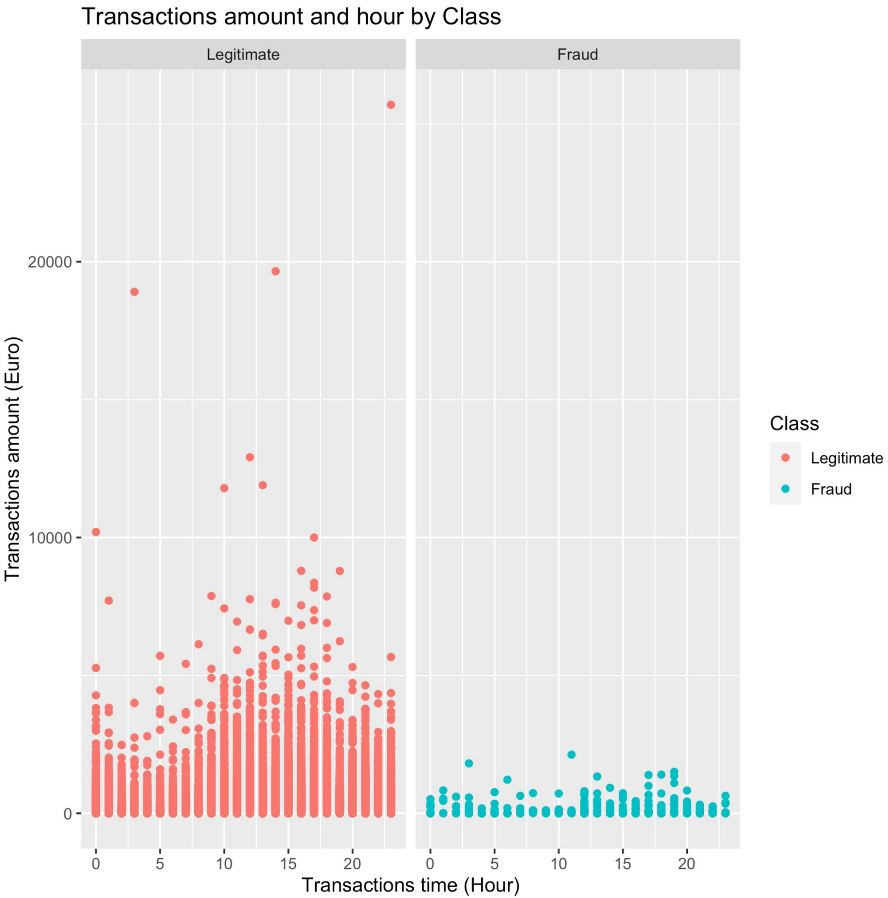




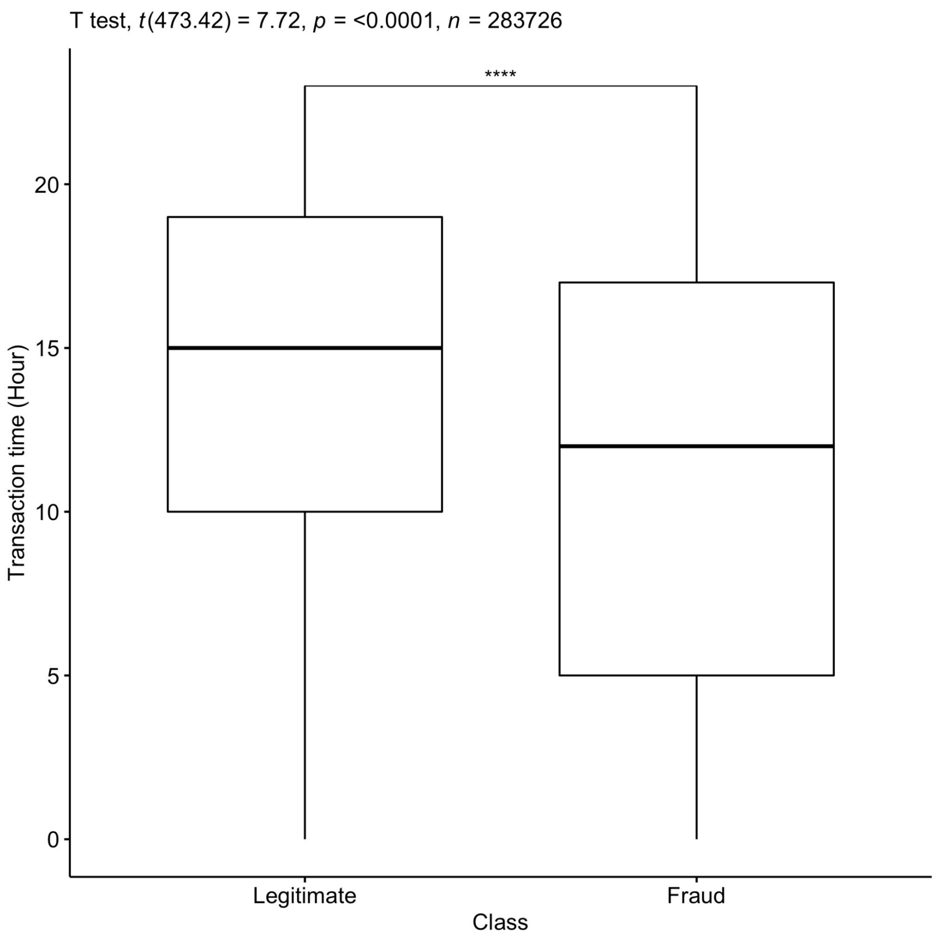
	Fraud	mean(Amount)	median(Amount)	sd(Amount)
1	Fraudulent	122.21	9.25	256.68
2	Legitimate	88.29	22.00	250.11

Table 1: Descriptive statistics table of transactions class

- **80% of the amounts are between 0 and 100 euros => daily expenses**
- **Several small fraudulent transactions => unnoticed**



- **The legitimate transactions decreased throughout the night, and it increased at the beginning of the day and kept remaining the whole day.**
- **The illegal transactions are more likely to evenly spread the whole time.**

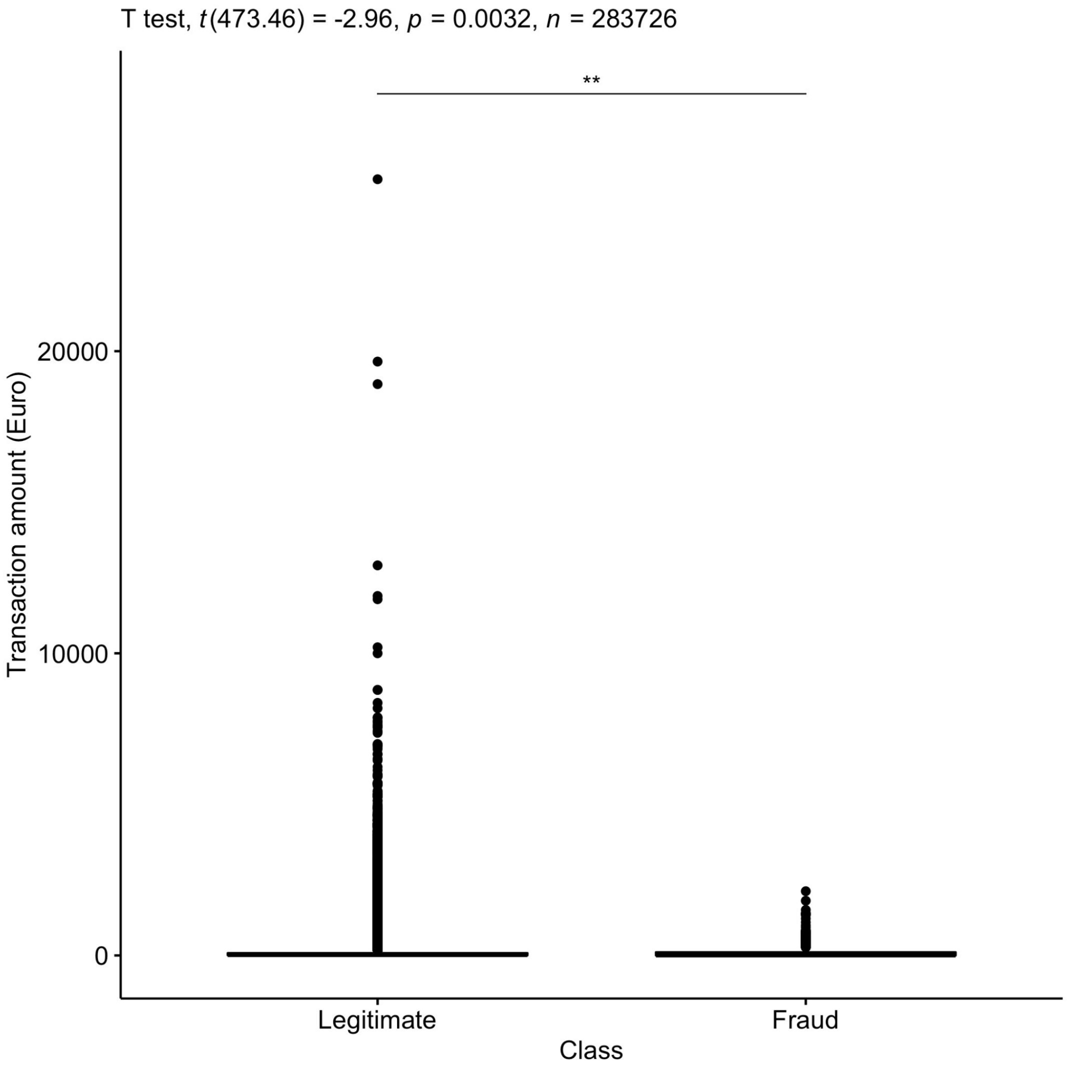


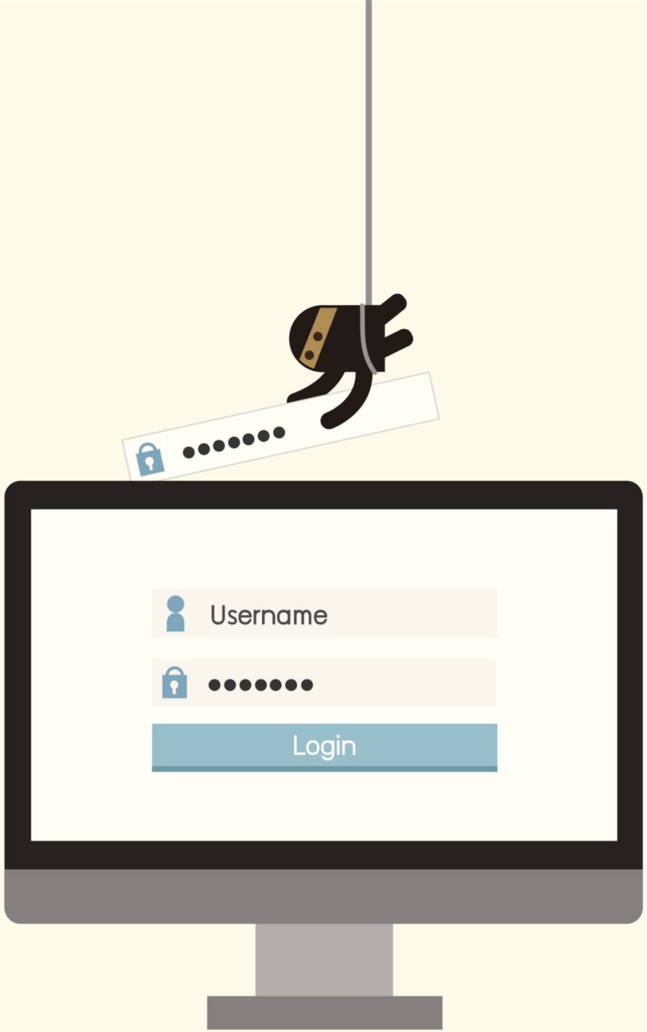
- **The mean score: Legitimate > Fraud.**
- **The magnitude of the differences in the means was significant.**

➢ **There is a significant difference in transaction time between Legitimate and Fraud transactions.**

- **The mean score of the Legitimate is lower than Fraud transaction**
- **The magnitude of the differences in the means was significant.**

➢ **There is a significant difference in transaction amount between Legitimate and Fraud transactions.**

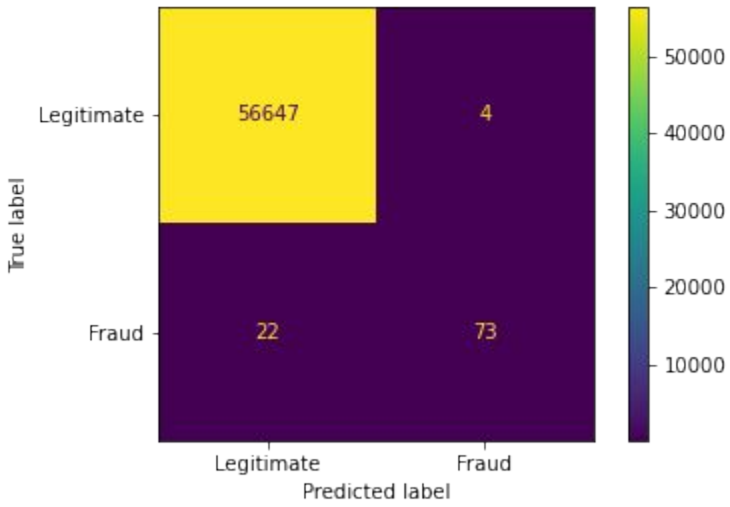




Classification Models

	Mean F_2 Score	Hyperparameters
Logistic Regression	0.6475	C = 100
Decision Tree	0.7877	max_depth = 5
Random Forest	0.8041	max_features = 5 n_estimators = 25

Result model



Our final model is a Random Forest Classifier with 25 trees in the forest and 5 features to consider when looking for the best split.

F2 score on test set: 0.7987

Conclusion

Conclusion

- Nowadays, a solution that minimizes the threat of credit card fraud and is also not too aggressive to block legitimate transactions is needed.
- Therefore, we emphasize developing an efficient and secure system for detecting fraudsters in our study.
- This study concludes that the model that returns the best results in this dataset is the Random Forest Classifier with 25 trees in the forest and five features to consider when looking for the best split.

Limitation

- Since 28 over 31 variables have already been transformed to principal components, we only got three variables to play with.
- Therefore, we did not gain much insight from our explanatory analysis.
- Our method is a naive approach to this problem, so our model may not be good enough to apply in production.
- Moreover, we have not dealt with the most critical part of this dataset: **the imbalance**.

Thank you for your attention!