deti universidade de aveiro
departamento de electrónica, telecomunicações e informática

Credit Card Fraud Detection

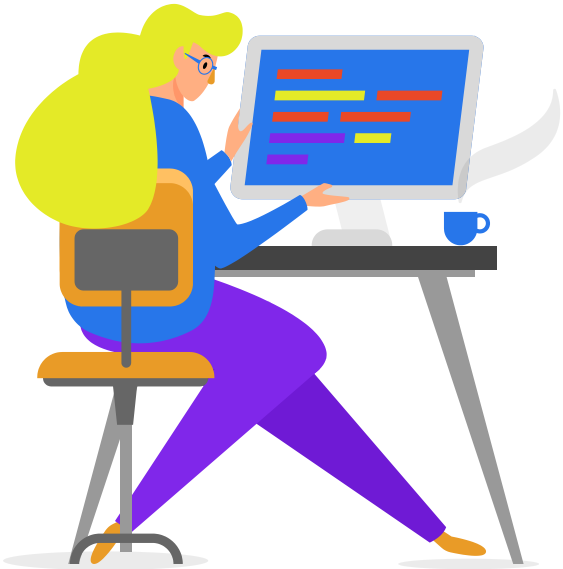
Aprendizagem Automática

Pétia Georgieva Georgieva

2022/2023

Paulo Pereira, 98430

Table of Contents



Context

01

Why this theme?

An interesting problem with a huge impact on the finance sector

02

Dataset

Unbalanced dataset available on Kaggle

Credit Card Fraud Detection Dataset | Kaggle

kaggle

03

ML models used

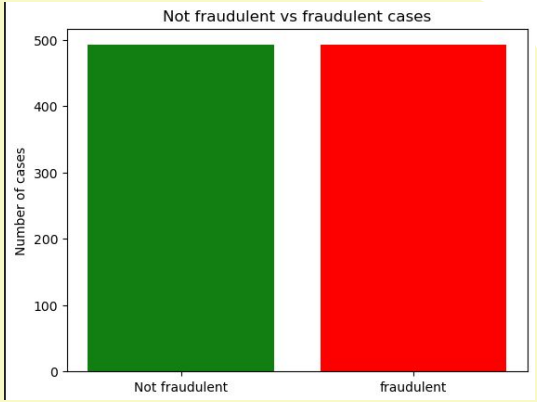
Data Analysis and Pre-Processing

Dataset analysis

- **31 columns-** including Amount, Time and Class
- **Unbalanced Dataset**
- **284807 entries-** 492 labeled as frauds
- **28 features:**
    - Used Vnumber for security reasons
- **1 output:**
    - 0 = doesn't predict heart attack
    - 1 = predicts heart attack

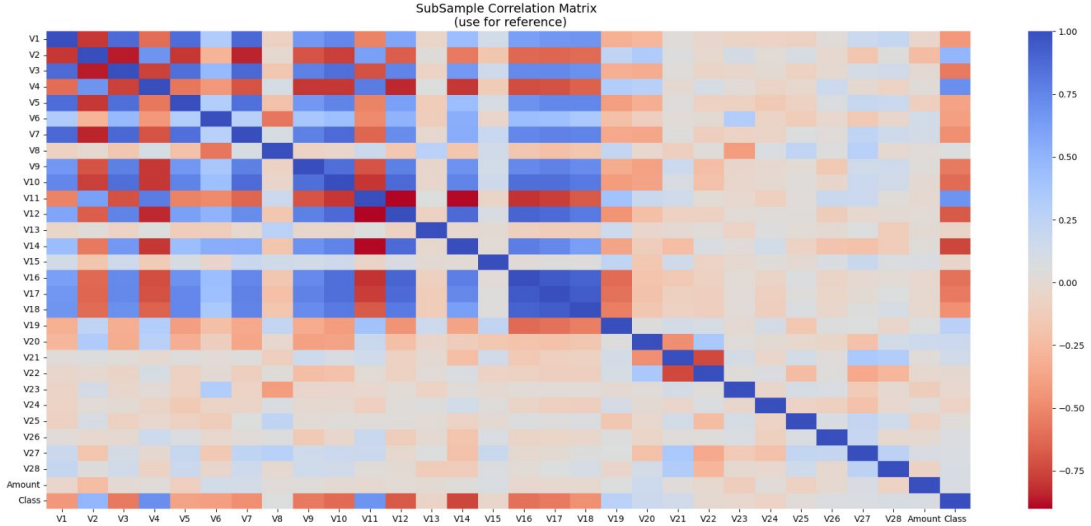
Data Analysis and Pre-Processing

Dataset pre-processing



- **Sub-Sampled-** Until equality was achieved
- **Find null values because row couldn't be used if present**
- **Naived Predictor**
  - Sub-Sampled dataset well balanced
- **Normalized data using StandardScaler-** Removed mean and scaling to unit variance

Data Analysis and Pre-Processing



Correlation Matrix

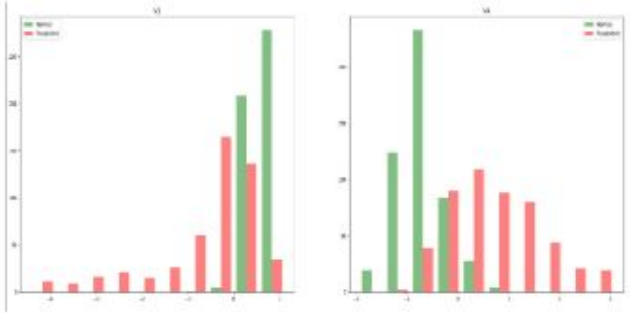
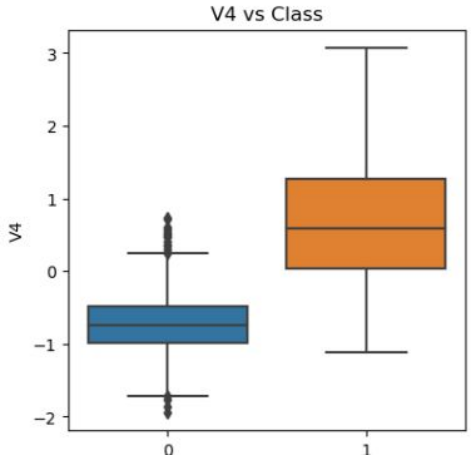
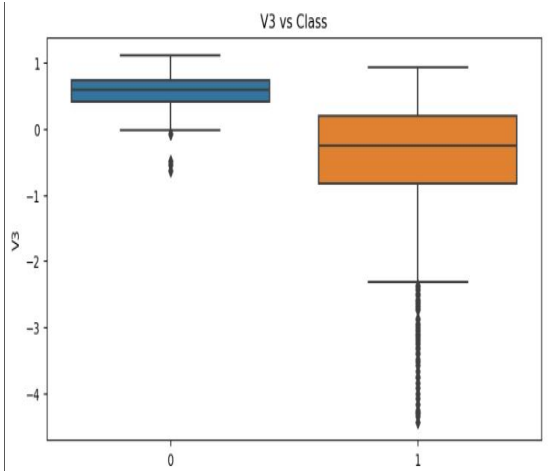


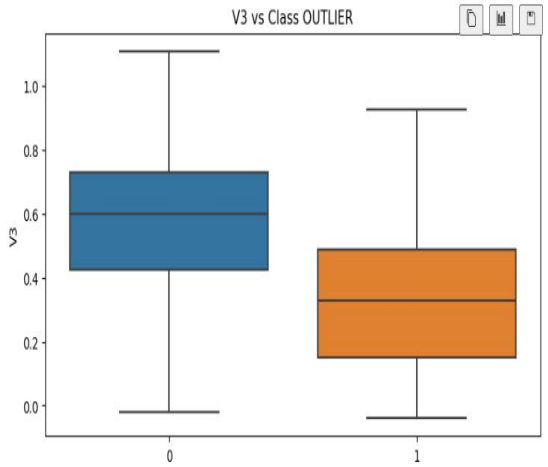
Figure 4 - Example where there is a big difference between both cases (V3/V4)

Data Analysis and Pre-Processing

Boxplot analysis







Models

Three Models were **explored** and **evaluated** with the **same metrics.**

Neural Networks

Improves performance over time

03

Support Vector Classification

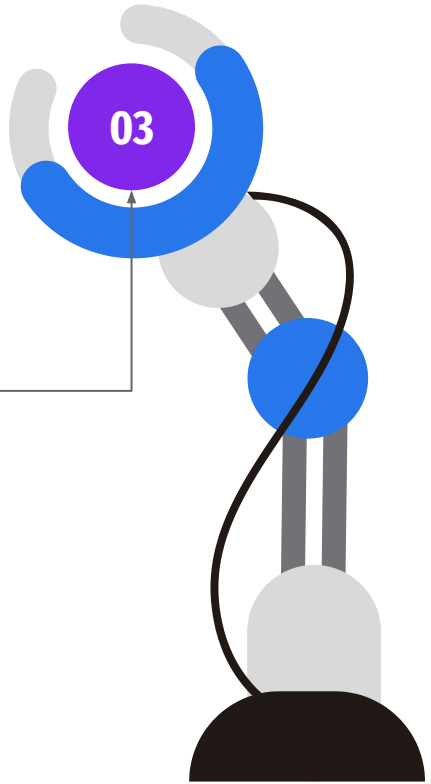
Using the rbf kernel parameter

02

Logistic Regression

Applies the logistic sigmoid function

01



Equations

Precision Score:
System's ability to not label as positive a sample that is negative

$$precision = tp/(tp + fp)$$

Recall Score:
System's ability to find all positive samples.

$$recall = tp/(tp + fn)$$

F1 Score:
Harmonic mean of the precision and recall

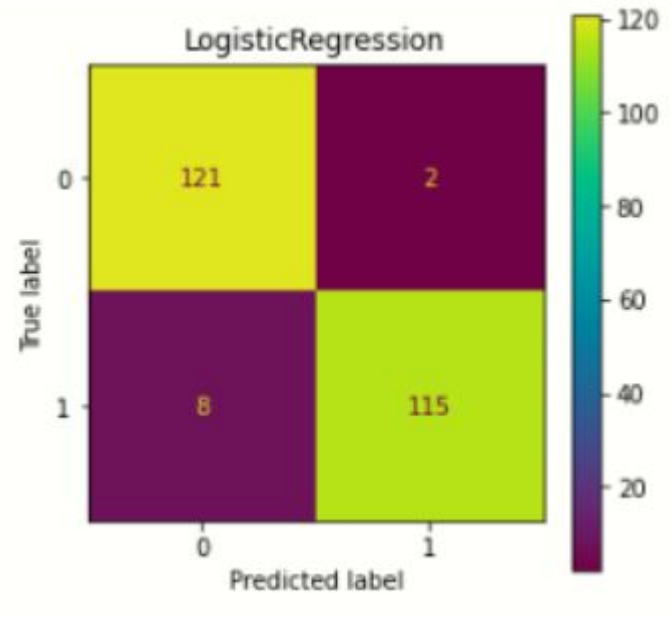
$$F1 = 2 * (precision * recall)/(precision + recall)$$

Accuracy Score:
Represents the overall performance of the mode
Given by how many Classes were correctly classified I

Logistic Regression (no penalty)

1

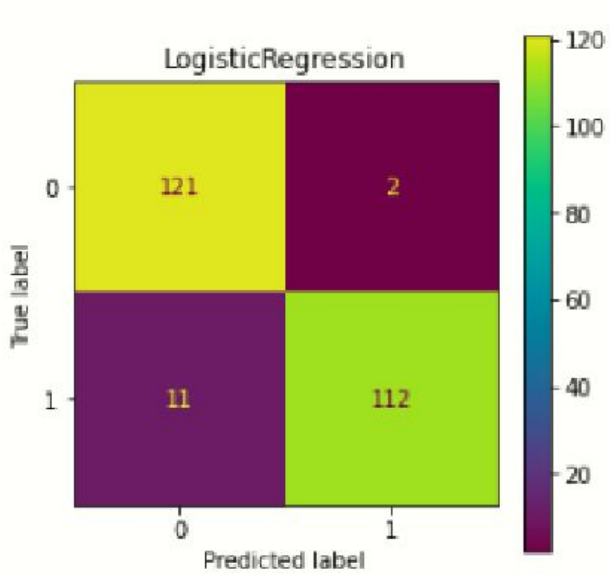
Scores	All Fea-tures	Best Fea-tures
F1	0.9184	0.9593
Accuracy	0.9187	0.9593
Precision	0.9725	0.9829
Recall	0.8618	0.9350



Logistic Regression (L2 penalty)

2

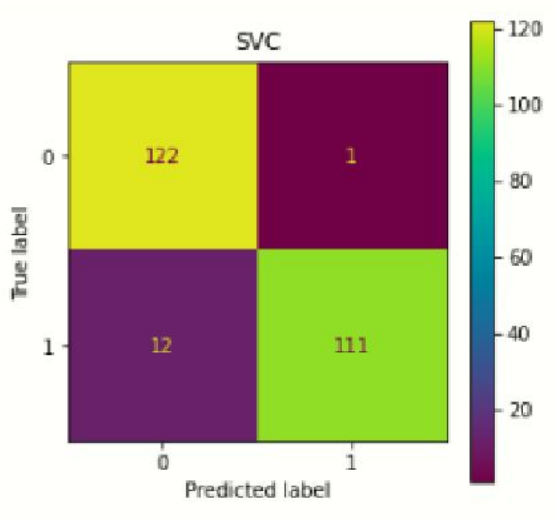
Scores	All Fea-tures	Best Fea-tures
F1	0.9184	0.9471
Accuracy	0.9187	0.9472
Precision	0.9813	0.9825
Recall	0.8537	0.9106



SVC

3

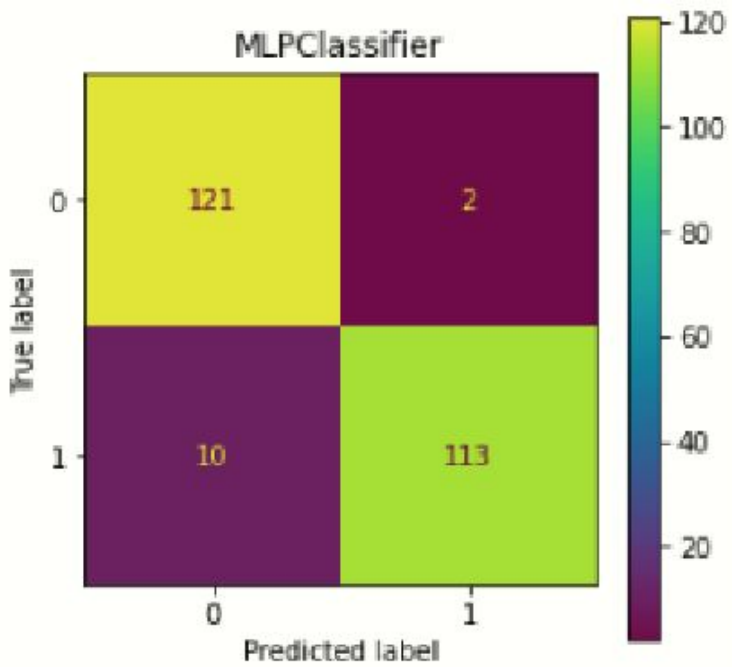
Scores	All Features	Best Features
F1	0.9143	0.9470
Accuracy	0.9146	0.9472
Precision	0.9722	0.9911
Recall	0.8537	0.9024



Neural Networks

4

Scores	All Fea-tures	Best Fea-tures
F1	0.8861	0.9512
Accuracy	0.8862	0.9512
Precision	0.9060	0.9826
Recall	0.8618	0.9187



Results

5	Accuracy	F1-Score
Logistics Regression	95.93%	95.93%
Logistics Regression “L2”	94.71%	94.72%
SVC	94.70%	94.72%
NN	95.12%	95.12%

6	FP+FN	Fit Time Rank
Logistics Regression	10	3rd
Logistics Regression “L2”	13	2nd
SVC	13	1st
NN	12	4th

Tuning Hyper-Parameter (Logistic Regression/SVC)

7

Solver	Max Iterations	class_weight	penalty	C
lbgfs	5000	balanced	L2	1

8

Kernel	C	gamma
rbf	3	1

Tuning Hyper-Parameter (NN)

9

Solver	Max Iterations	Hidden layers size	activations	alpha	Learning rate	Learning rate init
dam	5000	(12,12)	tanh	1e-3	constant	0,001

K fold

3

Model	Average Accuracy(%)
Logistic Regression	92.81
NN	93.76
SVC	92.81

Final Results

10

Model	F1 Score	Accuracy	Precision	Recall
Logistic Regression	0.9868	0.9769	0.0637	0.9024
SVC	0.9999	0.9999	0.9934	0.9228
NN	0.9995	0.9995	0.8876	0.8028

Novelty and Contributions

- Janio Martinez Bachmann’s notebook
- Joparga3’s notebook