**Credit Card Fraud Detection**

**Description From Kaggle:**

**Context**

It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

**Content**

The datasets contains transactions made by credit cards in September 2013 by european cardholders.

This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-senstive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

**Inspiration**

Identify fraudulent credit card transactions.

Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification.

**Acknowledgements**

The dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.

More details on current and past projects on related topics are available on https://www.researchgate.net/project/Fraud-detection-5 and the page of the DefeatFraud project

Please cite the following works:

Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015

Dal Pozzolo, Andrea; Caelen, Olivier; Le Borgne, Yann-Ael; Waterschoot, Serge; Bontempi, Gianluca. Learned lessons in credit card fraud detection from a practitioner perspective, Expert systems with applications,41,10,4915-4928,2014, Pergamon

Dal Pozzolo, Andrea; Boracchi, Giacomo; Caelen, Olivier; Alippi, Cesare; Bontempi, Gianluca. Credit card fraud detection: a realistic modeling and a novel learning strategy, IEEE transactions on neural networks and learning systems,29,8,3784-3797,2018,IEEE

Dal Pozzolo, Andrea Adaptive Machine learning for credit card fraud detection ULB MLG PhD thesis (supervised by G. Bontempi)

Carcillo, Fabrizio; Dal Pozzolo, Andrea; Le Borgne, Yann-Aël; Caelen, Olivier; Mazzer, Yannis; Bontempi, Gianluca. Scarff: a scalable framework for streaming credit card fraud detection with Spark, Information fusion,41, 182-194,2018,Elsevier

Carcillo, Fabrizio; Le Borgne, Yann-Aël; Caelen, Olivier; Bontempi, Gianluca. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization, International Journal of Data Science and Analytics, 5,4,285-300,2018,Springer International Publishing

Bertrand Lebichot, Yann-Aël Le Borgne, Liyun He, Frederic Oblé, Gianluca Bontempi Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection, INNSBDDL 2019: Recent Advances in Big Data and Deep Learning, pp 78-88, 2019

Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Frederic Oblé, Gianluca Bontempi Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection Information Sciences, 2019

**Description of the problem**

Credit card fraud is an inclusive term for fraud committed using a payment card, such as a credit card or debit card. The purpose may be to obtain goods or services, or to make payment to another account which is controlled by a criminal. The Payment Card Industry Data Security Standard (PCI DSS) is the data security standard created to help businesses process card payments securely and reduce card fraud.

Credit card fraud can be authorised, where the genuine customer themselves processes a payment to another account which is controlled by a criminal, or unauthorised, where the account holder does not provide authorisation for the payment to proceed and the transaction is carried out by a third party. In 2018, unauthorised financial fraud losses across payment cards and remote banking totalled £844.8 million in the United Kingdom. Whereas banks and card companies prevented £1.66 billion in unauthorised fraud in 2018. That is the equivalent to £2 in every £3 of attempted fraud being stopped.

Credit cards are more secure than ever, with regulators, card providers and banks taking considerable time and effort to collaborate with investigators worldwide to ensure fraudsters aren't successful. Cardholders' money is usually protected from scammers with regulations that make the card provider and bank accountable. The technology and security measures behind credit cards are becoming increasingly sophisticated making it harder for fraudsters to steal money.

From Wikipedia.

This analysis can help for the banks to detect the fraud credit card.

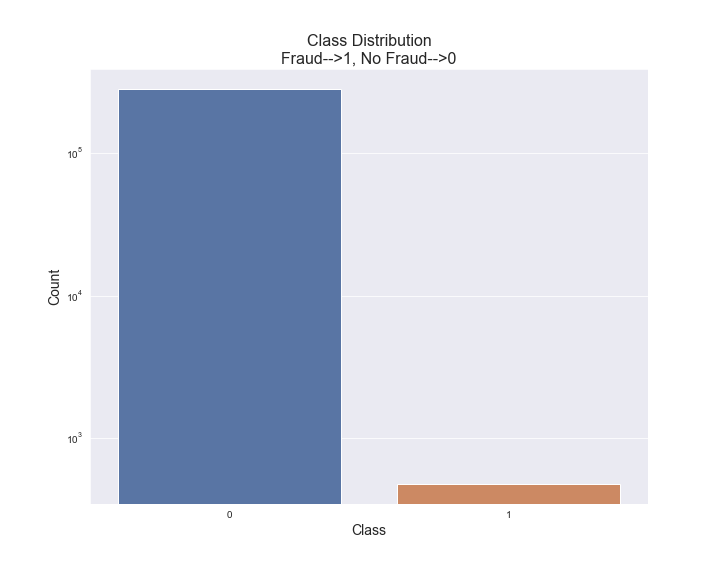
**Predicting the fraud credit cards**

* Use different machine learning models to predict the fraud cards. The models show around 90% sure to find the fraud cards.
* The percentage of fraud is 0,17!
* The fraud distribution is not so high by the amount of fraud, if we compare the global picture. The maximum amount is 2126 dollar.
* The fraud transactions follow the normal transactions, so we cannot find a specific time to fraud.

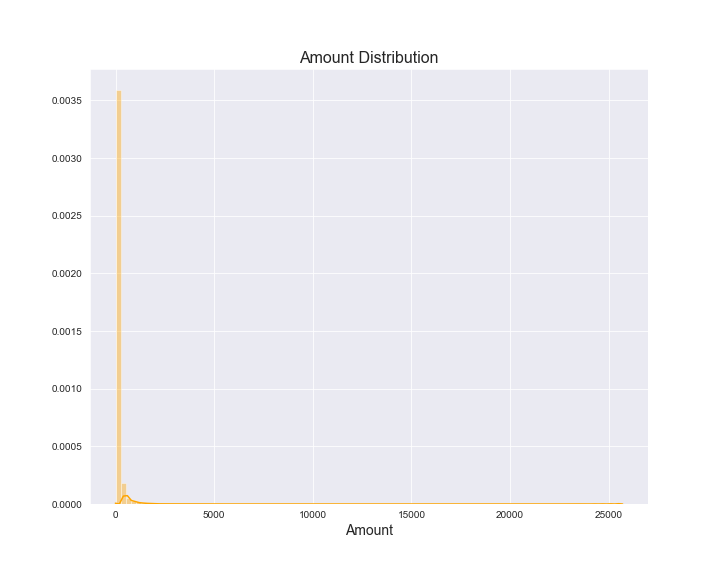
**Data description and cleaning**

* Unfortunately, I cannot describe the data, because they are secret! However with these data we are able to find a solution for this kind of problem.
* The first step is to understand what contains the dataset. You can go on with the delete duplicate, set the types, drop the NA values and so on.

**Class Distribution:**

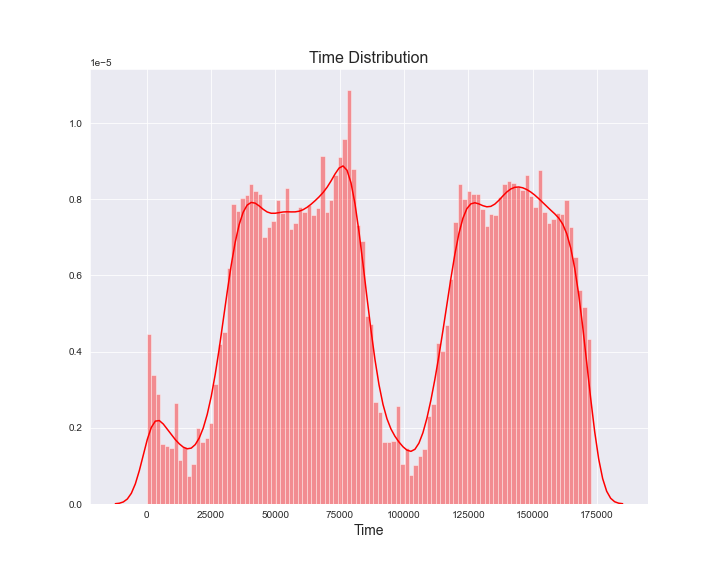
As we can see the distribution is not balanced! For the models, it is better to balance it.

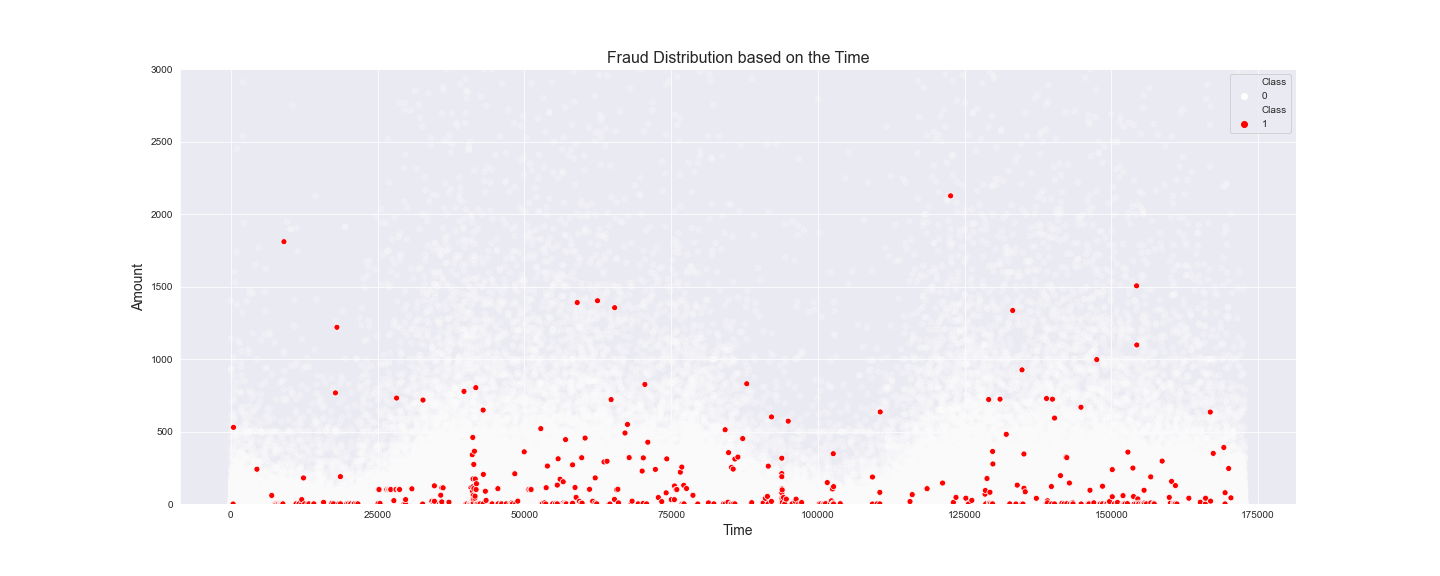
**Amount Distribution:**

The amounts are grouped together at the low level.

**Time Distribution:**

Before 25000 and between 100000 and 125000 are Transactions.

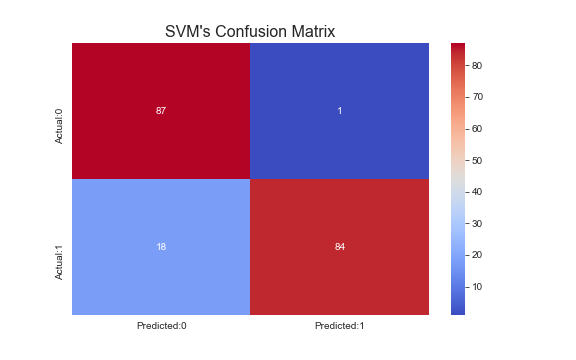
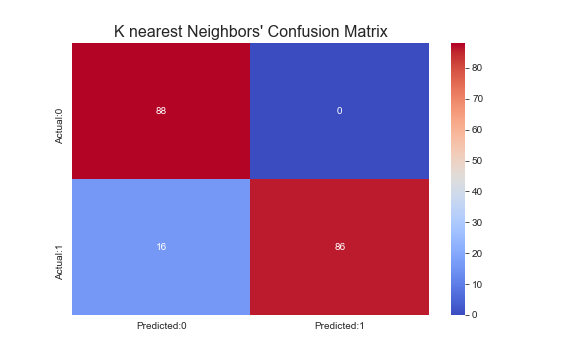
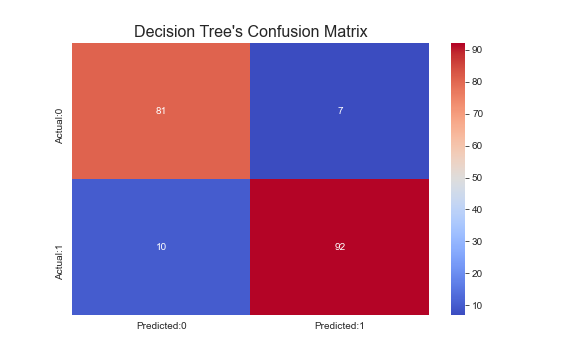


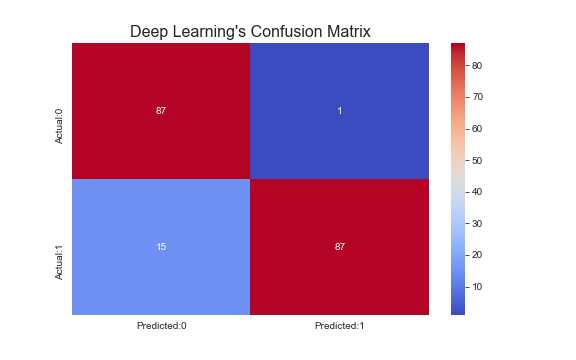
**Fraud Distribution:**

Before 25000 and between 100000 and 125000 are less Fraud transaction. it is related to the "Time distribution".

**Machine Learning models**

For this analyses I have used some diferrent models such as Ligistic Regression, K Nearest Neighbors, Decision Tree, Support Vector Machine and the final one: a Deep Learning model. On the next page you can see the results.

**Machine Learning models – Confusion Matrix**



**Machine Learning models – Results**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Models: | Decision Tree | Support Vector Machine | Logistic Regression | K Nearest Neighbors | Deep Learning model | Avarage |
| Accuracy score | 91,05% | 90,00% | 91,58% | 91,58% | 92,00% | 91,24% |
| True Positive | 81 | 87 | 86 | 88 | 87 | 85,8 |
| False Positive | 7 | 1 | 2 | 0 | 1 | 2,2 |
| False Negative | 10 | 18 | 14 | 16 | 15 | 14,6 |
| True Negative | 92 | 84 | 88 | 86 | 87 | 87,4 |
| Sum | 190 | 190 | 190 | 190 | 190 |  |

**Conclusion and future directions**

* Collect more data to test it and get more accurated models.
* Add new independent variables to get new insights.