Credit Card Fraud detection

IBM Data Science - Coursera final assignment

Introduction

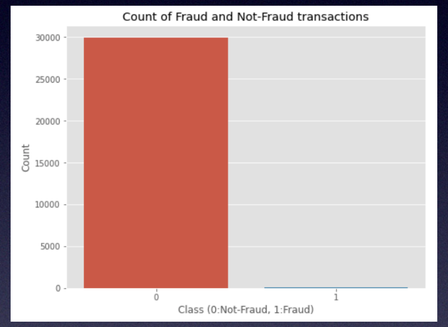
- The number of credit card owners is projected close to overpass 1 billion by 2022.
- Card Fraud detection is highly important activity requiring 24/7 data processing and alerting of potential CC miss-use.
- To ensure security of credit card transactions, it is essential to monitor fraudulent activities.

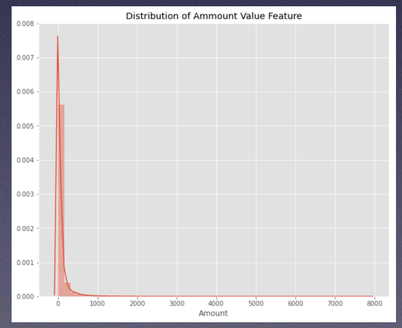
Data set sourcing

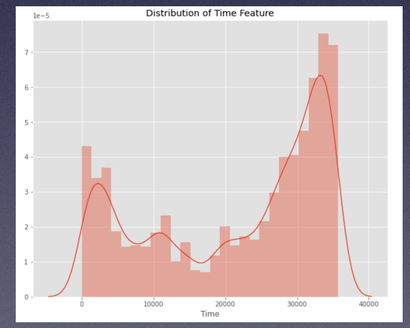
- The datasets contains transactions samples made by credit cards in September 2013 by European cardholders (file `creditcard.csv`).
- This dataset presents transactions that occurred in several hours (approximately half of the day).
- The number of fraud transactions is **94**. Number of transactions in the sample overall is **30,000**.
- The dataset is highly unbalanced, the positive class (frauds) account for **0.31%** of all transactions.

Data set overview

- Feature [Time] contains the number of seconds between each transaction and the first transaction in the dataset.
- The feature [Amount] is the transaction amount in Eur.
- Feature [Class] is the response/dependant variable and it takes value 1 in case of FRAUD and 0 in case of NOT-FRAUD.
- [Amount] feature is highly right-skewed (large number of small-amount transactions in comparison to the number or high-amount ones).
- The [Time] feature has visible heights and low parts through the period of ~10 hours. We can assume that small number of transactions happens during the night hours.

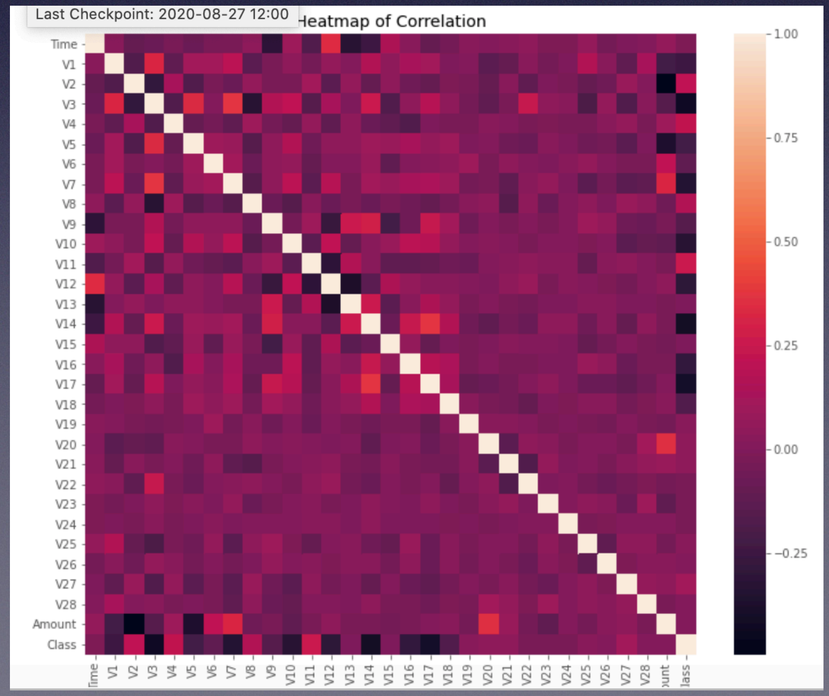






Correlations

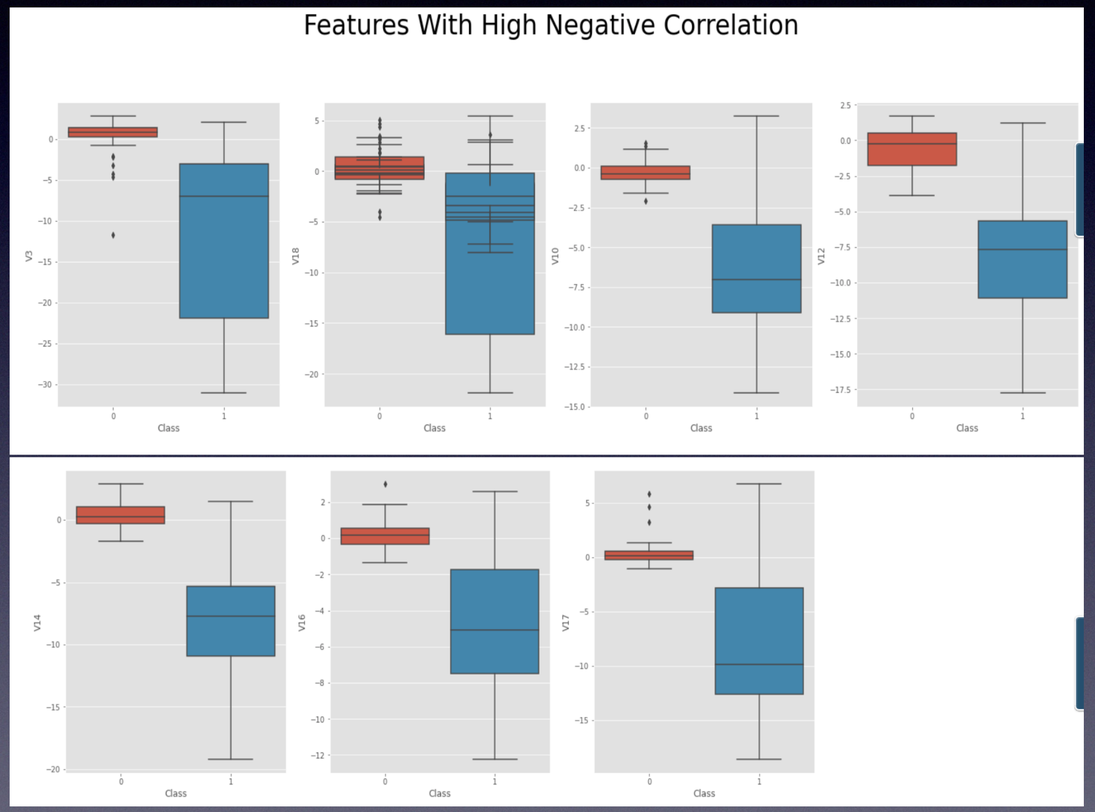
- There are relatively little significant correlations for such a big number of variables, most probably because huge [Class] imbalance distorts the importance of certain correlations with regards to our [Class] variable.

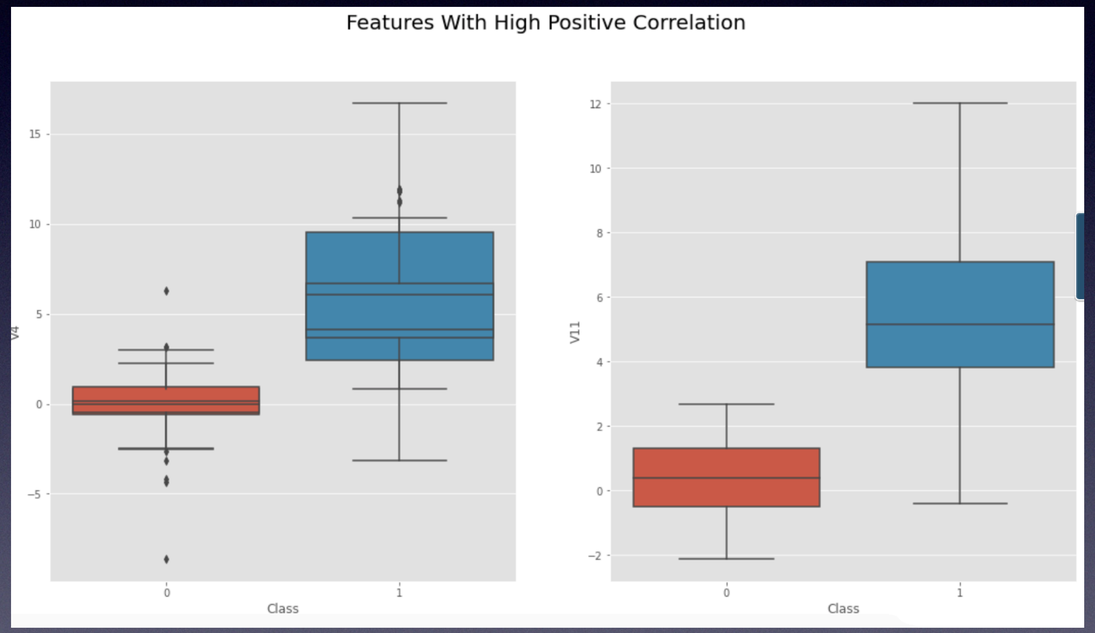


Methodology

- We will start with data preparation in a way to scale the [Time] and [Amount] features in order to avoid bias in machine learning models.
- Using the original data set would not provide required prediction strength. Since >99% of transactions are [NOT-FRAUD]
  - Trivial algorithm that always predicts that the transaction is [NOT-FRAUD] would have an accuracy higher than 99%!
- To create our balanced training data set, we will take all of the fraudulent transactions in our data set and count them. Then, we will randomly selected the ~same number of non-fraudulent transactions and concatenate the two.
- After shuffling this newly created data set, visualisation will show how new data set looks like.

How it looks like after data preparation





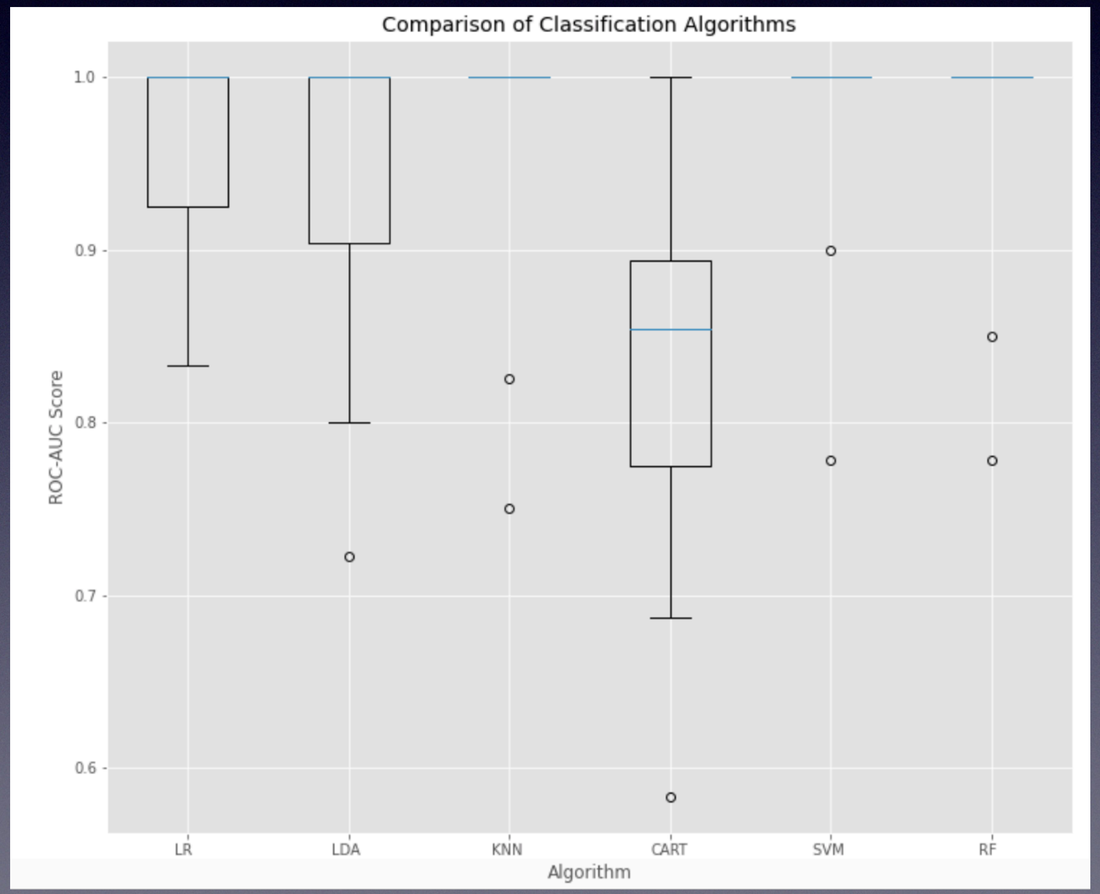
Preparation for modeling

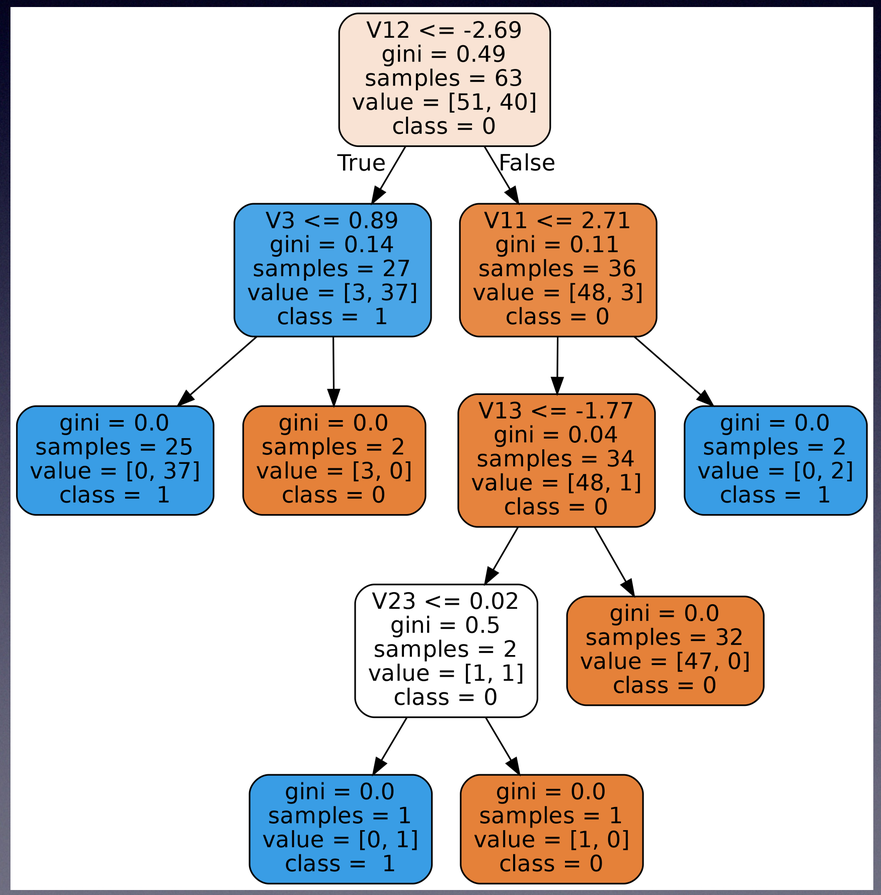
- To be able to test the performance of our algorithms, we will first perform the 80/20 train-test split.
- To prevent overfitting, we can use resampling with of K-fold cross-validation:
  - splitting training data into k parts (folds)
  - fit the model on k-1 folds before making predictions for the kth fold
  - repeat this process for every single fold and average the resulting predictions

Algorithms in use

- Logistic Regression
- Linear Discriminant Analysis
- K Nearest Neighbours (KNN)
- Classification Trees
- Support Vector Classifier
- Random Forest Classifier

Visual presentation of algorithm results





Discussion

- It seems there is a difference in algorithm performance. From the sample available, it is not fully clear if the RF should be best option to proceed with (as might be the first choice in pre-analysis [when tried with significantly bigger data-set on local machine]).
- I have added the decission tree as an addition to random forest.

Conclusion

- Credit Card Fraud detection has confirmed to be very complex issue that requires a substantial amount of planning, data preparation, cleansing and sclaing the data, bringing down the highly unbalanced dat to balanced set.

- In case of non-anonymised features available, we could make further analysis in regards the most important features in creating fraud case.

- Future work will include a comprehensive tuning of the Random Forest algorithm and, potentially, much higher data set.