CE9010 DATA SCIENCE Project

CREDIT CARD FRAUD DETECTION MODEL

TEAM MEMBERS

- ➔ Bhagwat Abhishek, U1722796C
- ➔ Yun Hong Jun,U17XXXXXX
- ➔ Chee Yeng Sung, U1820382K

Agenda

1. Context
2. EDA
3. Choice of Metrics
4. Pre-processing
5. Model Consideration
6. XGBoost
7. Random Forest
8. Logistic Linear
9. Conclusion

1 Context & Data

Context & Data

284,807 transaction data over 2 days

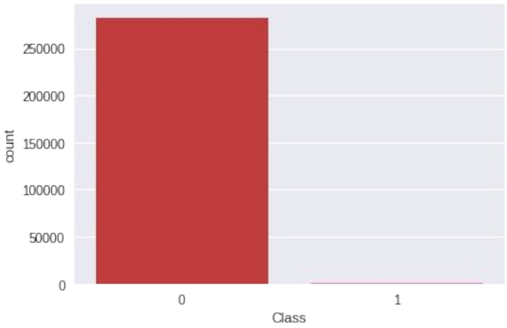
30 Features (Time, Amount 28 PCA-transformed features)

Label: Fraud / Non-fraud (our prediction)

2 Exploratory Data Analysis

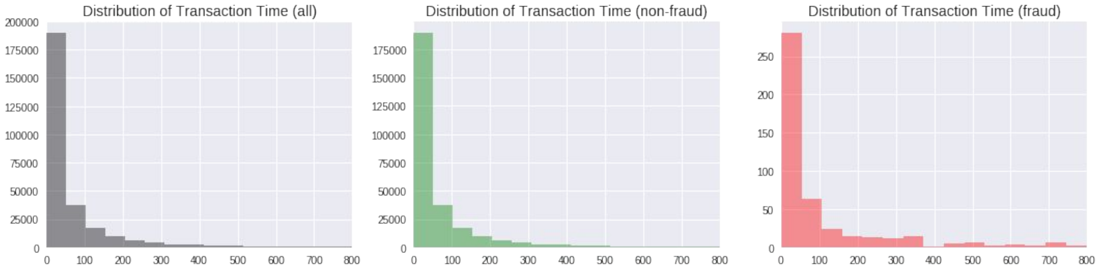
EDA

Class Distribution

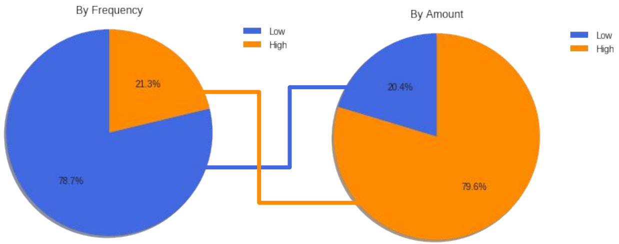


Very skewed dataset
Non-Fraud: 99.84%
Fraud: 0.16%

Amount



Pareto Principle

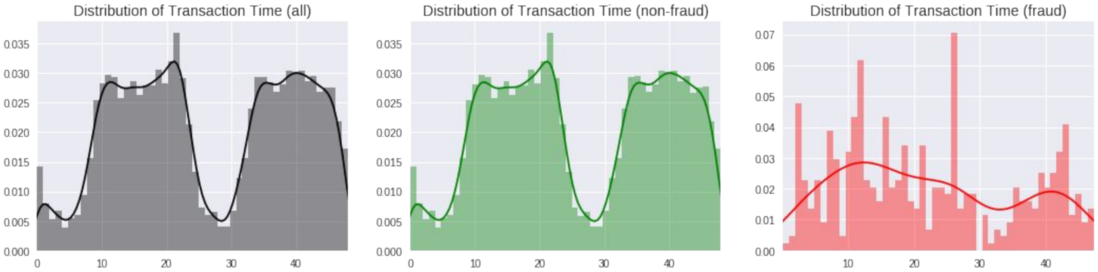


Amount

Most of the transactions are low valued.
Pareto Principle (20:80 rule) applies:

- 80% of data makes up 20% of total value.
- It’s the other 20% of data makes up 80% of total value

Time

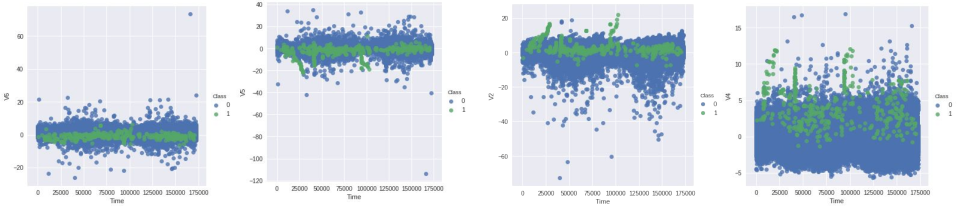


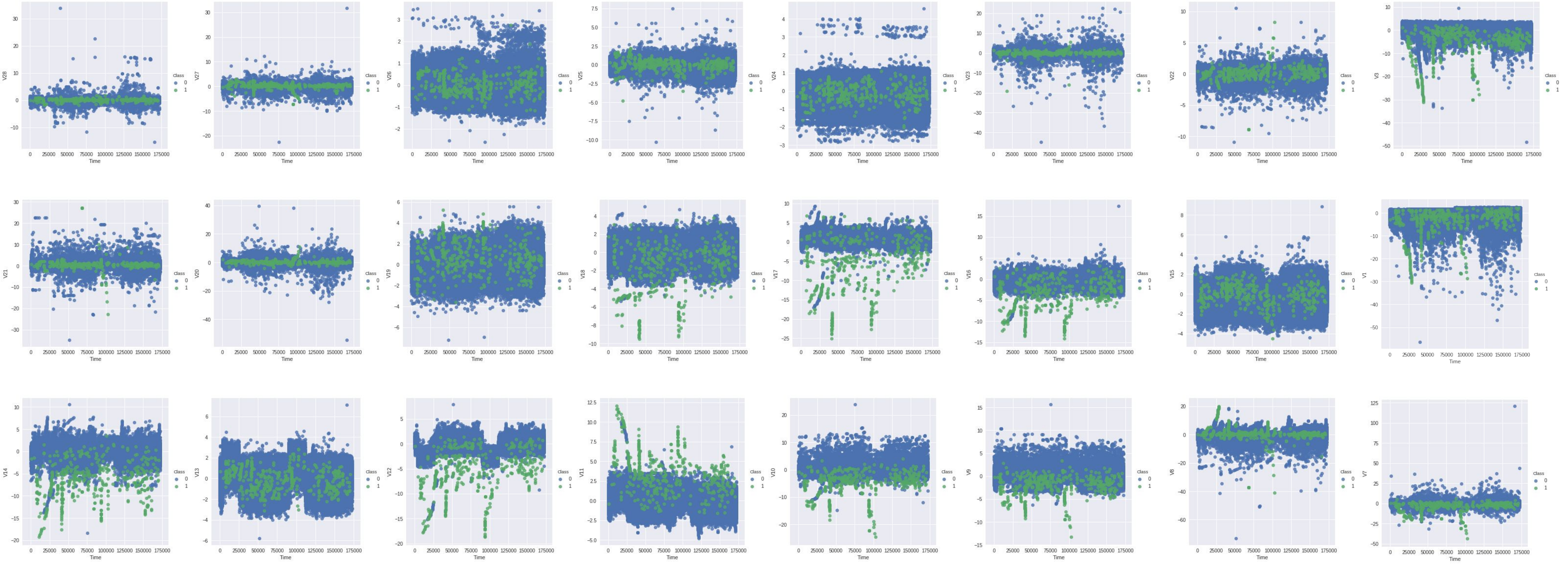
Time

High number of transaction from 8AM to 10PM. Low during bed time.

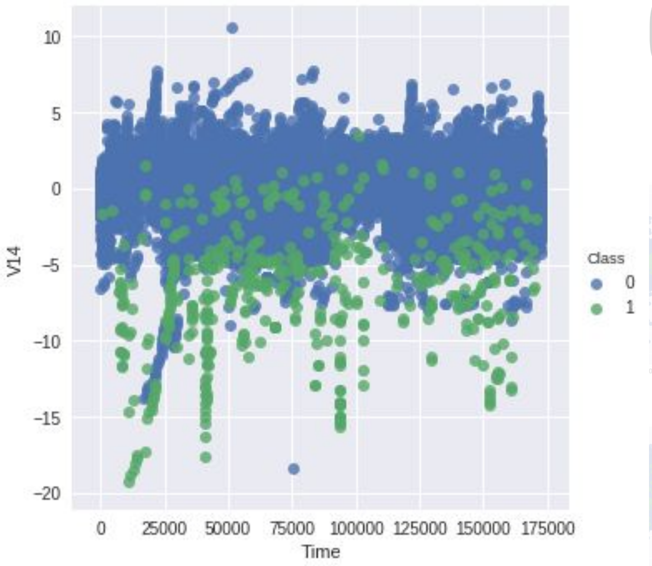
No specific pattern in fraud.
(Low importance expected)

EDA - Pca Features Plotting

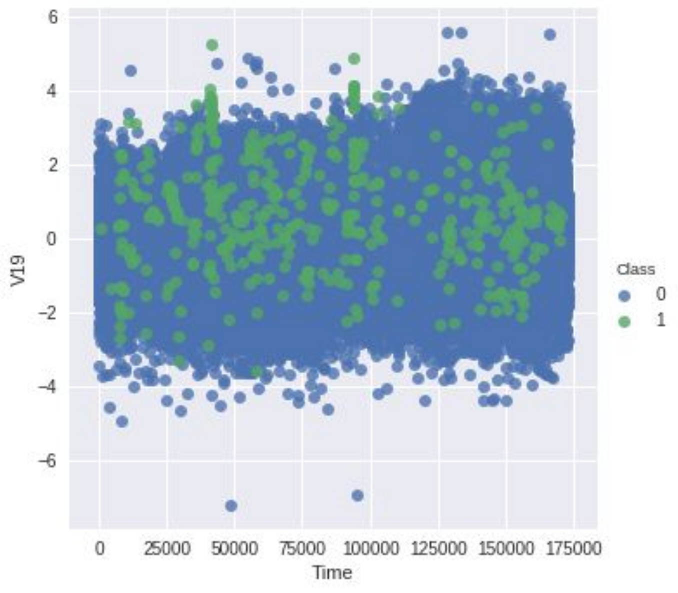




EDA - G



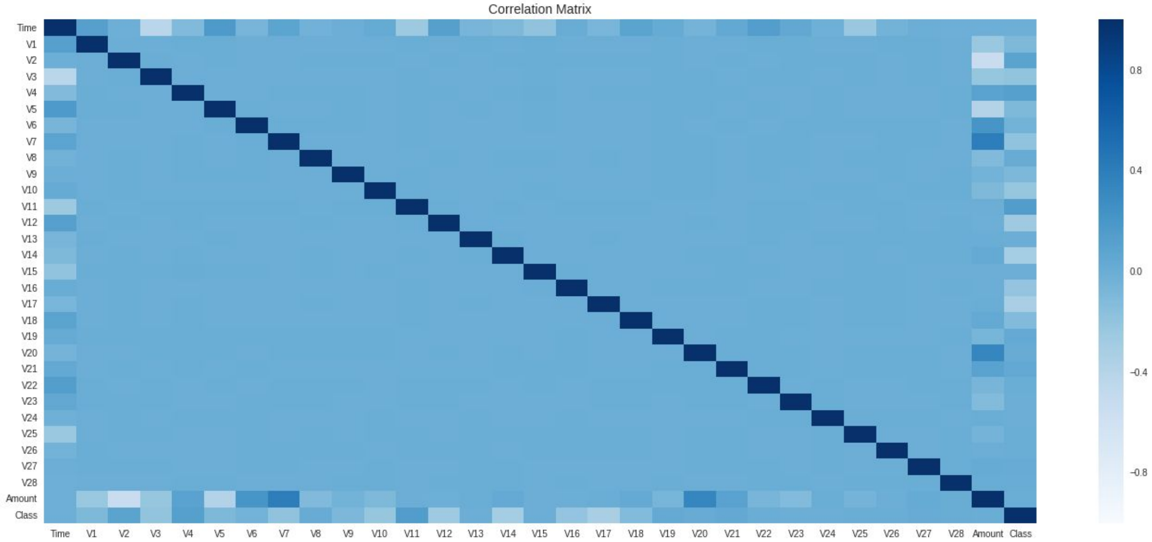
V14



V19

There is a clearer distinction between the fraud and non-fraud in V14 compared to other features such as V19.
V14 could be an important feature in determining whether a transaction is Fraud or Non-Fraud.

EDA - Correlation



No correlation found between the PCA transformed features.

PCA removed the correlation between the features.

3 Choice of Metric

Choice of Metrics

Accuracy

Measures the % of correctness of prediction. Misleading when data is very imbalanced.

F1 Score

Harmonic Average between recall and precision.

Precision

Focuses on Correctness of the Positive Predictions. Good when focus on minimising false positives.

Recall

Focuses on identifying as much frauds as possible. Good when focus is on minimising the false negatives.

4

Pre-Processing

PreProcessing

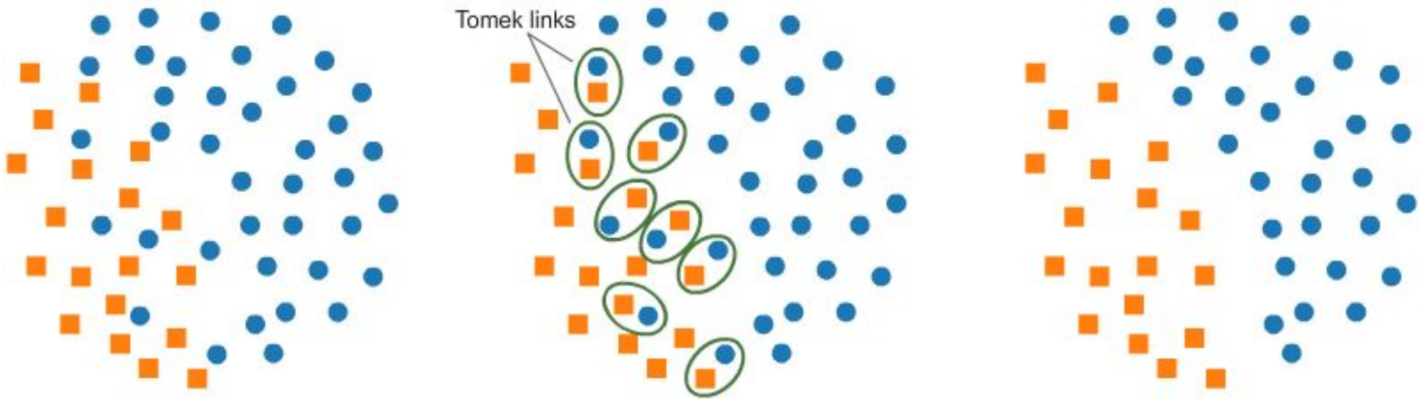
1. The first step would be to separate the features from the labels and understand the problem. In our dataset, we have chosen the ‘Class’ column to be the label, as it is that which we will be predicting in this Classifier Model. The remaining columns are taken as features.

2. We have made use of Stratified k - fold cross validation technique for splitting the data set into test data and train data. It is generally a better approach when dealing with both bias and variance. A randomly selected fold might not adequately represent the minor class, particularly in cases where there is a huge class imbalance.

Preprocessing

SMOTE + TOMEK LINKS

As the dataset we are using is a very imbalanced one, we chose to oversample the minority class. The most efficient way for our model would be to use SMOTE to oversample the class. We have used TOMEK links used as an under-sampling method or as a data cleaning method. Tomek links to the over-sampled training set as a data cleaning method. Thus, instead of removing only the majority class examples that form Tomek links, examples from both classes are removed



5 XGBoost

XGBoost

What it’s about

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient.It provides a parallel tree boosting that solve many data science problems in a fast and accurate way.

Hyperparameter tuning

Grid Search:
N_estimators: 50

Results

Confusion Matrix

True Label

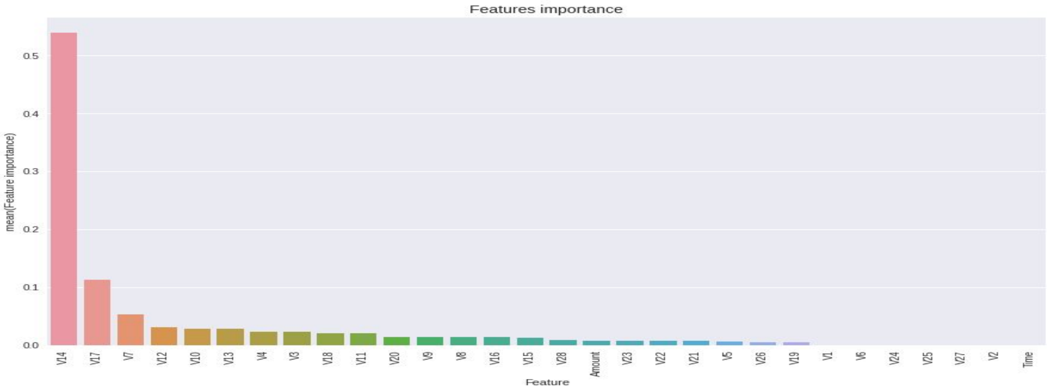
	0	1
0	56306	23
1	207	2443

Predicted Label

Scoring

Recall: 0.921
Precision: 0.990
F1 Score: 0.954

Feature importance



6 Random Forest

Random Forest

What it's about

Learning method for both classification and regression.
Construct **decision trees**.
Output the class that's mode of class of each trees

Hyperparameter tuning

Grid Search
Max_depth: 3
N_estimators: 3

Results

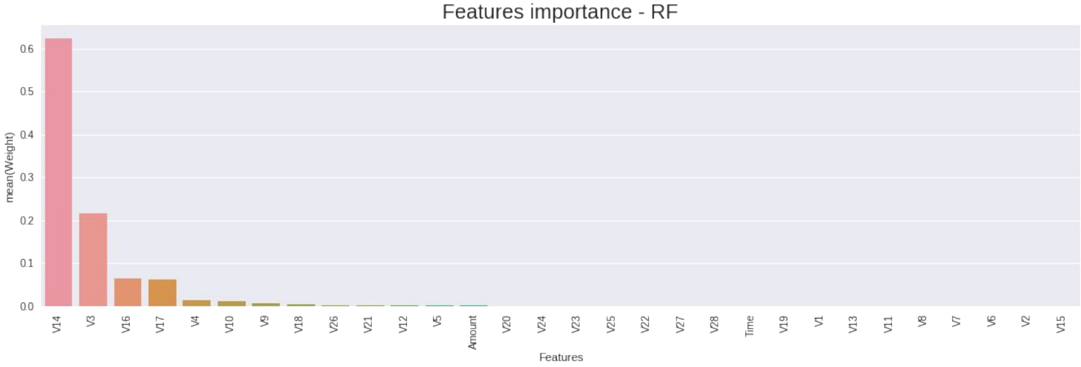
Confusion Matrix

		Predicted Label	
		0	1
True Label	0	56324	14
	1	362	2297

Scoring

Recall: 0.864
Precision: 0.994
F1 Score: 0.924

Feature importance



7 Logistic Linear

Logistic Linear Regression

What it's about

Supervised learning technique for classification in which its logistic regression (probability) function is used to determine the decision boundary, to compute the loss by and to apply gradient descent technique.

Hyperparameter tuning

Cross-validation:

$1/\lambda = 0.1$

$\lambda = 10$

Results

Confusion Matrix

True Label

0	56166	170
1	432	2225

Predicted Label

Scoring

Recall: 0.8374

Precision: 0.9290

F1 Score: 0.8808

8 Conclusion & Further Research

Conclusion

As mentioned above, we have used 3 metrics to evaluate our model -

1. Precision
2. Recall
3. F1 score

Model	Precision	Recall	F1 Score
Logistic Regression	0.929	0.837	0.880
Random Forest	0.994	0.864	0.924
XGBoost	0.990	0.921	0.954

From the table above, you can notice that the main parameters required for the evaluation, Recall and F1 Score, are the best for XGBoost.
Hence we can conclude that it is the best model which can predict credit card fraud activity most accurately.

Further Research

1. Anonymised Features

- Limited insights could be extracted as we could not identify what each features meant. More meaningful insights could be drawn if we had access to the context of the features
- For example, behavioural analytics may be used and separate models can be built for each cardholder to learn the spending pattern of each cardholder, therefore having a stronger model to predict.

2. Non-representative data

- The data is collected for only about 2 days and hence it may not be the best sample that represent the population. Also, the number of frauds was very limited. Therefore, more data would help the model to have better prediction.

3. Different types of frauds

- There are different types of credit card frauds such as friendly fraud, account takeover and unauthorised transactions. Based on our assumption, the features may differently depending on the types of the frauds. In the given dataset, the class was binary (only whether it's a fraud or non-fraud). This could be split into multi-class that have different types of frauds, hence the model can be more accurate in classifying different types of frauds.