Credit Card Fraud Detection

Using Machine Learning

Presented To:
Dr. Upasna Talukdar

Presented By: Radha Agrawal
Priya Kumari

Motive

- Identify fraudulent credit card transactions.
- to find new methods for fraud detection and to increase the accuracy of results.
- to detect the credit card fraud in the dataset obtained from ULB by applying Logistic regression, Decision tree to evaluate their Accuracy, sensitivity, specificity, precision using different models and compare and collate them to state the best possible model to solve the credit card fraud detection problem.

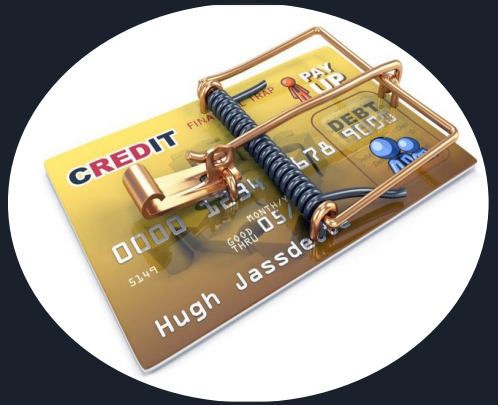
Fraud

- Theft and fraud using a credit or similar payment mechanism as a fraudulent source or funds in a transaction .

- Purpose being to obtain goods without paying or obtain unauthorized funds.

Types of Credit Card Fraud

- Stolen Cards
- Compromised accounts
- Mail/Telephone/Internet order
- Account takeover
- Skimming
- Phishing



Challenges Involved

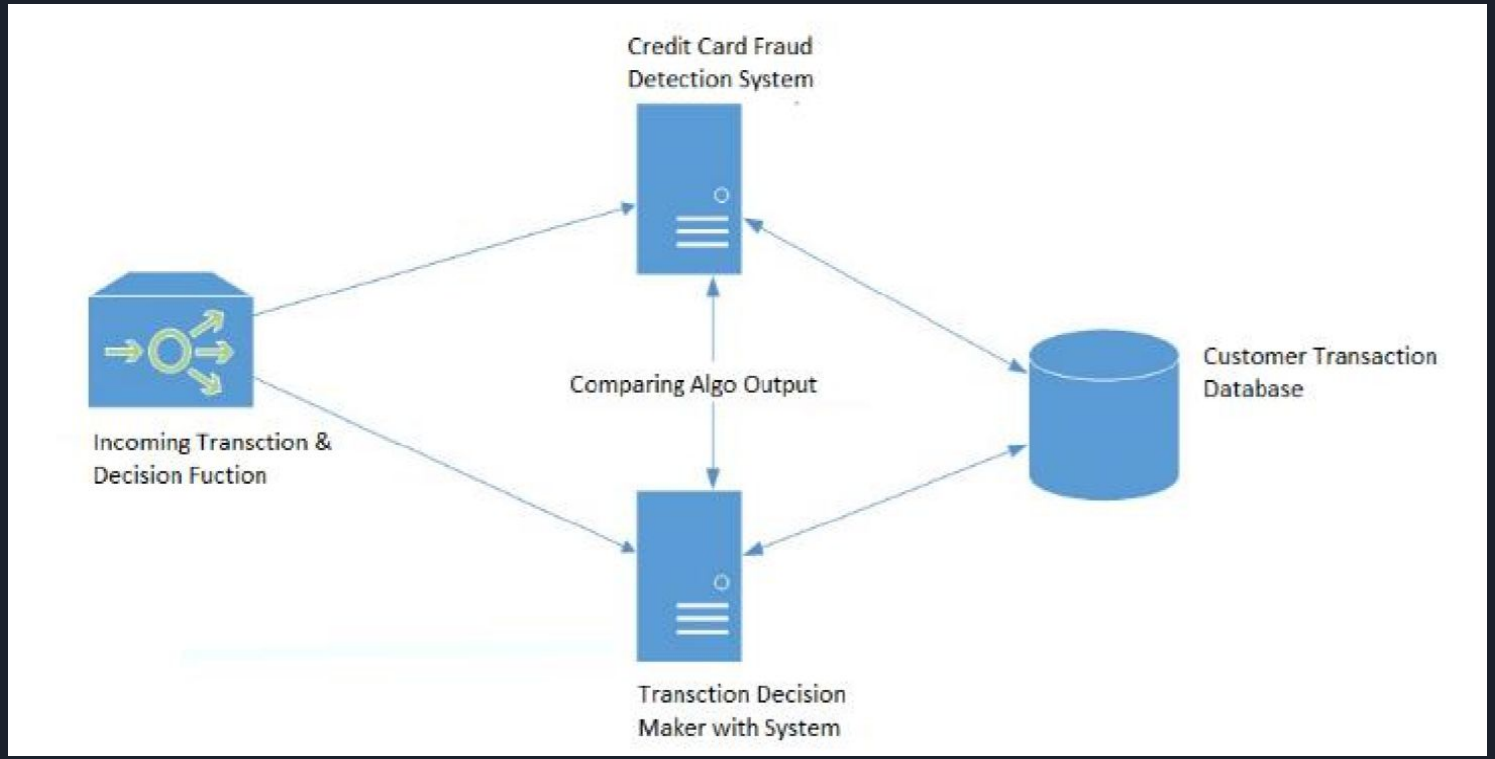
Credit card fraud detection is:

- One of the most explored domains of fraud detection
- Relies on the automatic analysis of recorded transactions

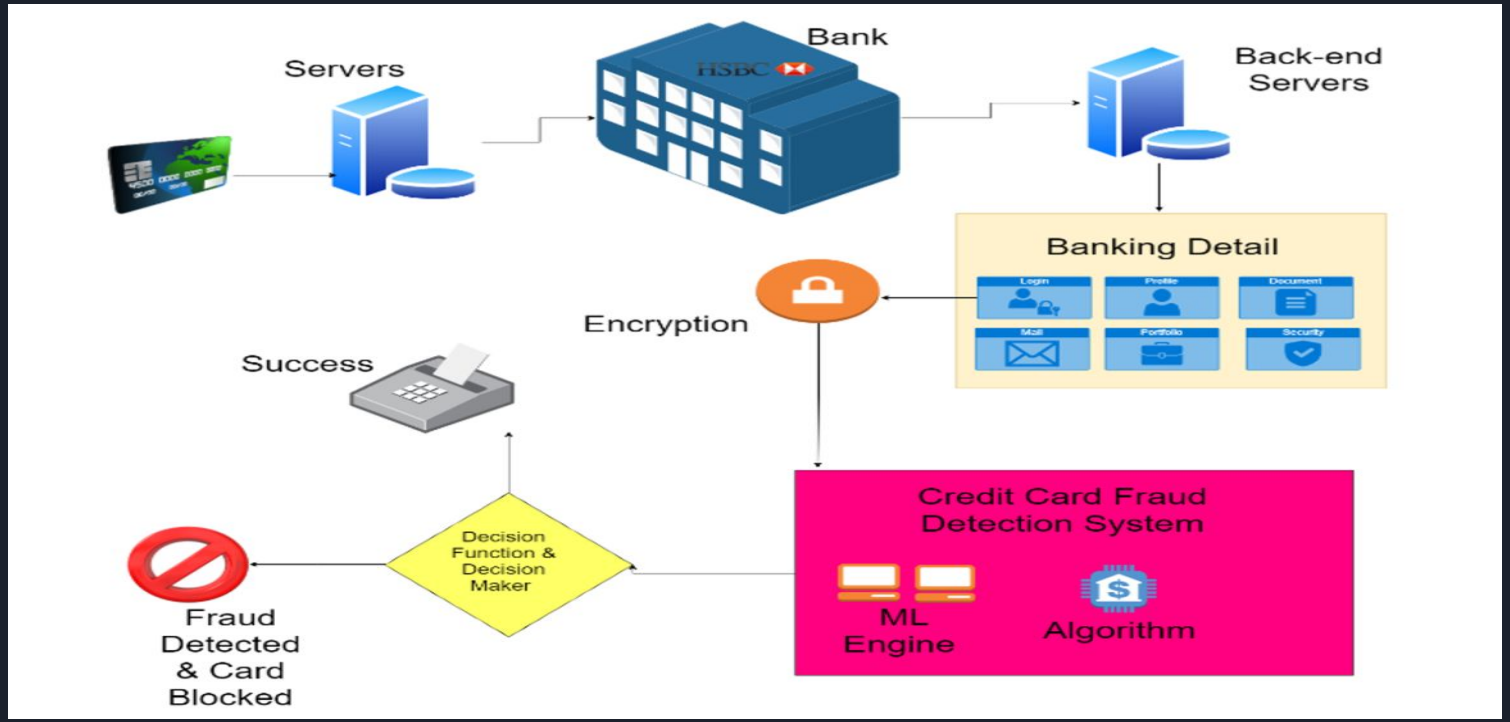
The main challenges in credit card fraud detection are:

1. Huge size of data
2. Imbalanced dataset
3. Availability of data

The basic rough architecture:



The full architecture diagram:



Dataset*

- 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.
- It contains only numeric input variables which are the result of a PCA transformation.

*downloaded from kaggle ,sourced from ULB Machine Learning Group

Content

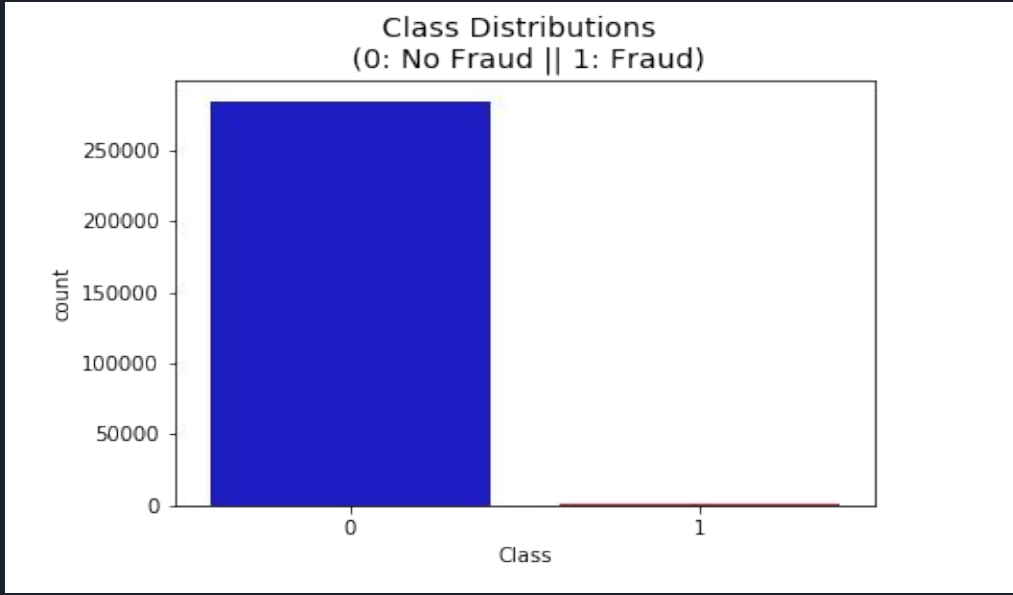
There are ~285,000 rows and 30 features and one class where time and amount are actual variables.

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 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Class Distributions

The dataset is highly unbalanced and skewed towards the positive class and positive class that is fraud cases make up 0.173% of the transactions data.



METRICS FOR PERFORMANCE EVALUATION

Due to class imbalance problem, the accuracy metric will not only help us in identifying the best classifier model

Consider alternative measures:

- Precision
- Recall
- F-measure

Performance Metrics

Precision – Fraction of records that are actually positive in the group that the classifier predicted as positive.

Precision = True Positives / True Positives + False Positives

Recall – Fraction of positive examples correctly predicted by the classifier.

Recall = True Positives / True Positives + False Negatives

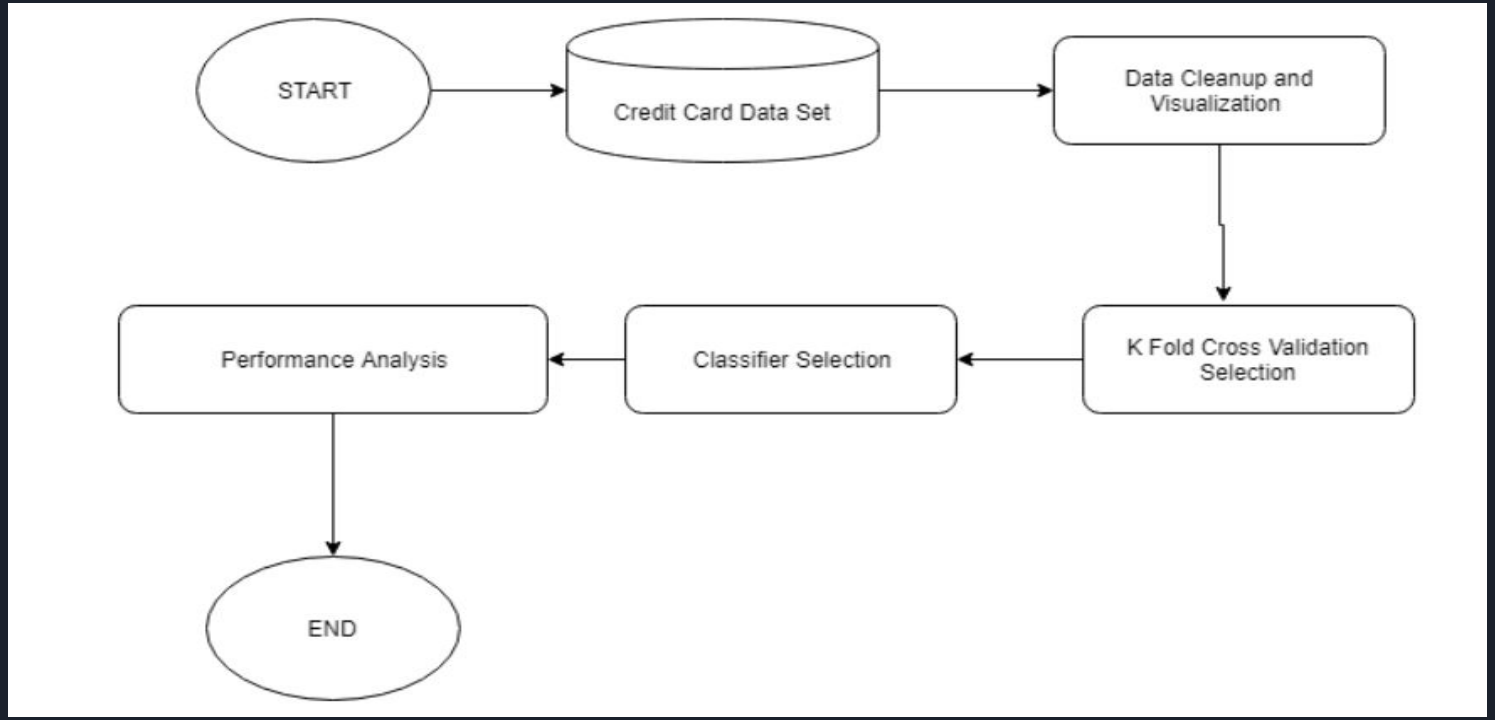
F-measure – A Harmonic mean between recall and precision.

F-measure = 2*Recall*Precision / Recall + Precision

Performance of classifiers

- Classifier models based on and logistic regression,naive bayes and decision tree,Random Forest are developed.
- To evaluate these models, 70% of the dataset is used for training while 30% is set aside for validating and testing.
- Accuracy, sensitivity, specificity, precision are used to evaluate the performance of the classifiers.

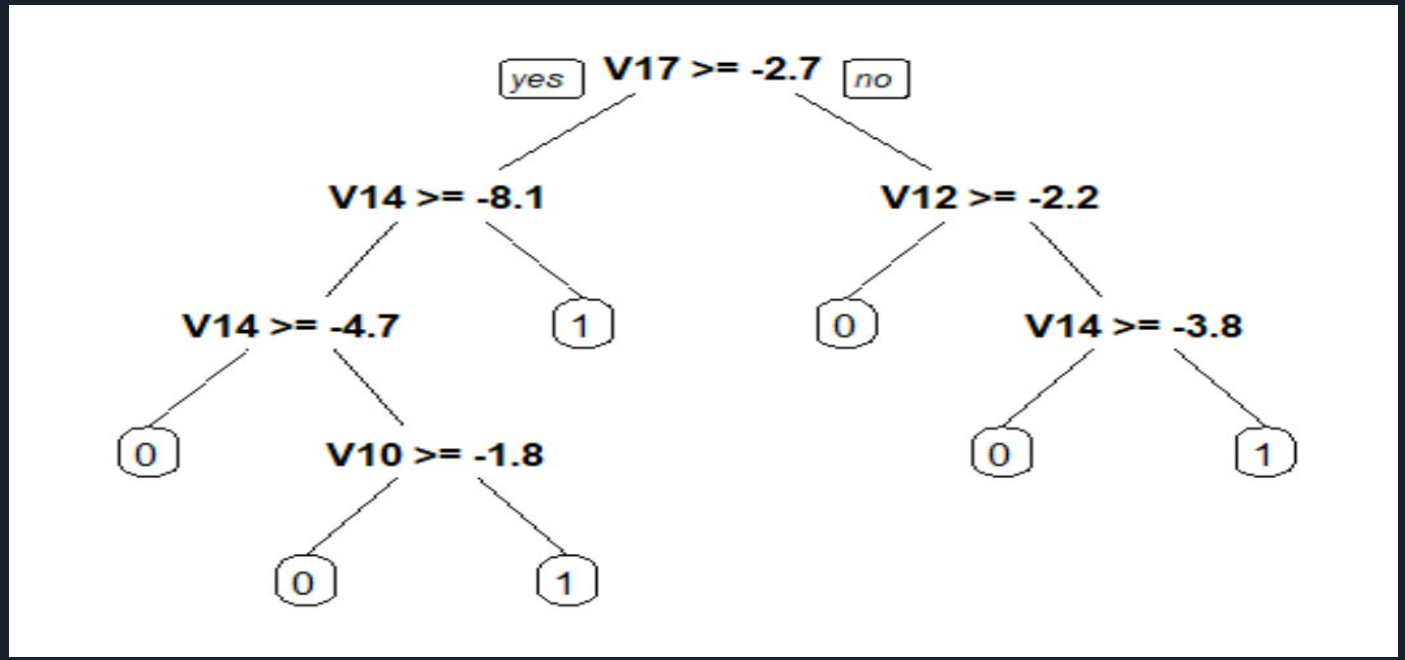
contd.


START
Credit Card Data Set
Data Cleanup and Visualization
K Fold Cross Validation Selection
Classifier Selection
Performance Analysis
END

Results

Metrices	Classifier					
	Logistic Regression	Decision Tree	Naive Bayes	KNN	K Means	RandomForest
Accuracy	.93	.99	.86	0.99	0.79	0.99
Precision	.97	.99	.98	0.97	0.05	0.96
Recall	.90	1.0	.74	0.75	0.61	0.73
F-measure	.93	0.99	.84	0.85	0.01	0.83

contd.



Decision Tree Reference

Conclusion

- Credit card fraud is without a doubt an act of criminal dishonesty.
- We have listed out the most common methods of fraud along with their detection methods and explained in detail, how machine learning can be applied to get better results in fraud detection
- While the algorithm does reach over 99.6% accuracy.This high percentage of accuracy is to be expected due to the huge imbalance between the number of valid and number of genuine transactions.
- While we couldn’t reach out goal of 100% accuracy in fraud detection, we did end up creating models that can, with enough time and data, get very close to that goal

Future Enhancements

- More room for improvement can be found in the dataset.
- As demonstrated before, the precision of the algorithms increases when the size of dataset is increased. Hence, more data will surely make the model more accurate in detecting frauds and reduce the number of false positives.
- However, this requires official support from the banks themselves.


THANK
YOU