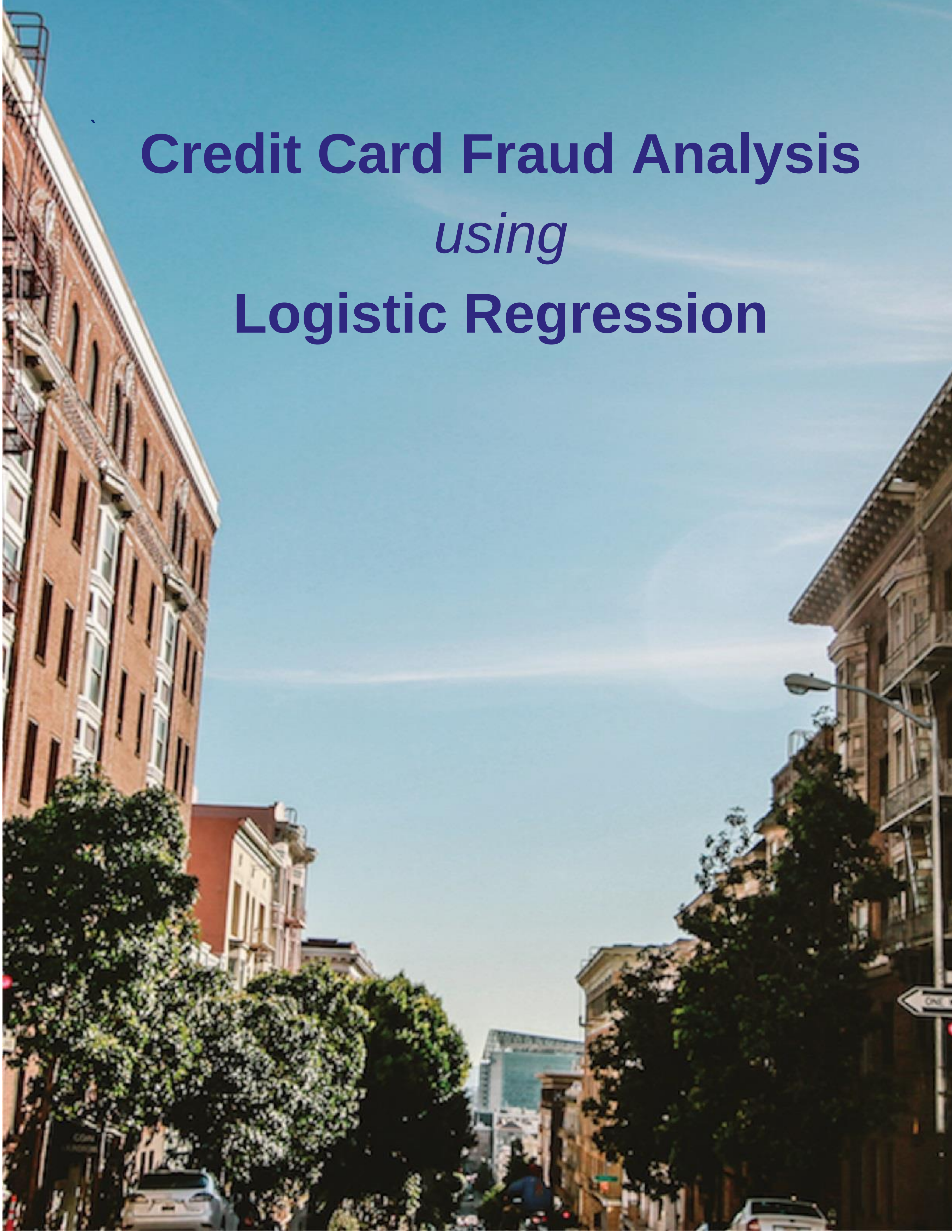
Credit Card Fraud Analysis

using

Logistic Regression



Contents

Credit Card Fraud Analysis

This data is taken from KAGGLE.
Link: https://www.kaggle.com/samkirkiles/credit-card-fraud/data

Data

The datasets contain credit card transactions made by credit cardholders. In the dataset we have 284,807 transactions of which of which 492 are fraud transactions (0.172% of all transactions).

The dataset contains total of 31 variables. Explanation given below:
Time: time elapsed in seconds between the transaction and the first transaction
V1, V2…V28: No explanation given
Amount: transaction amount
Class: response variable – 1 for fraud and 0 for not-fraud

Task

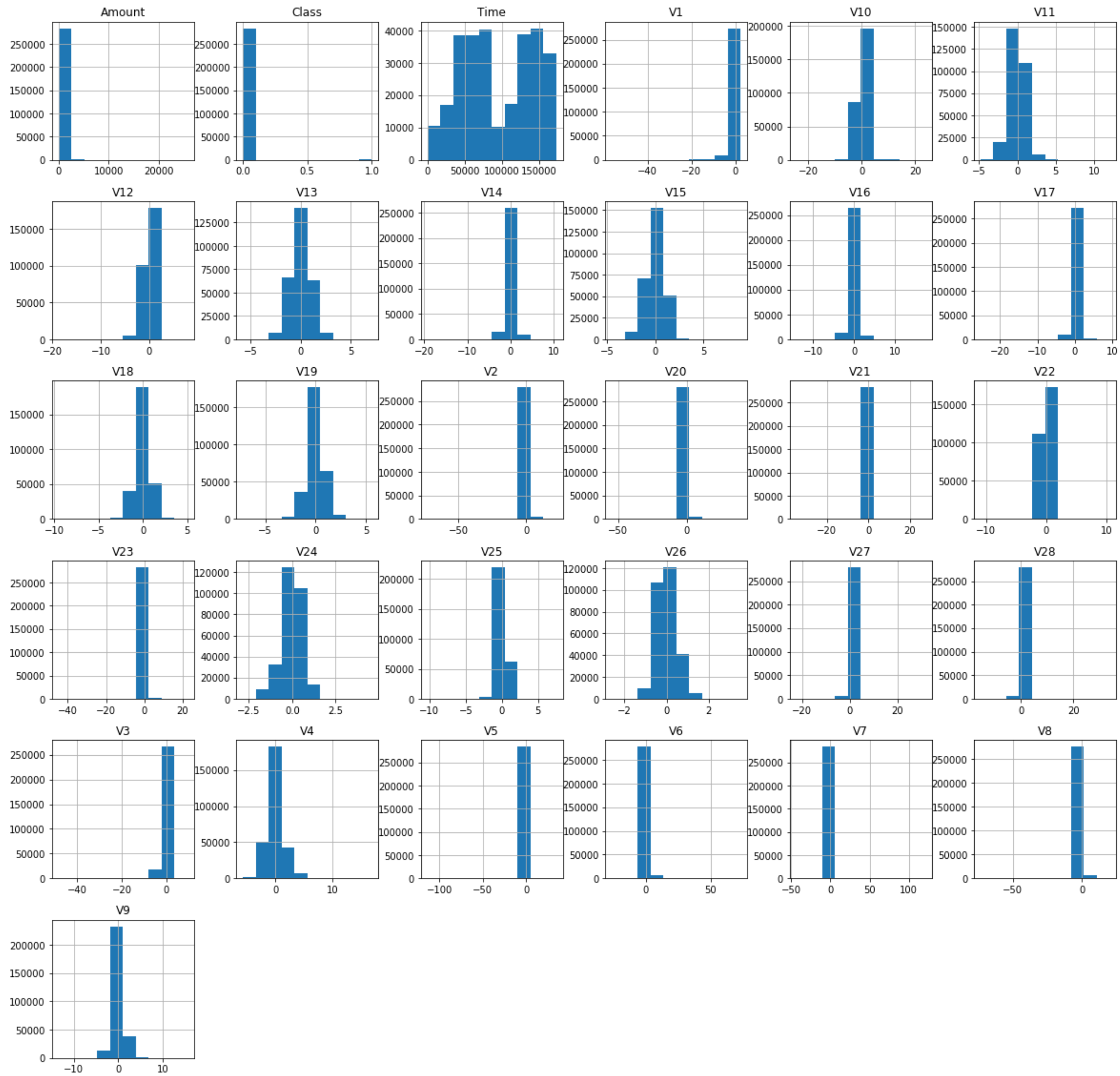
To determine whether a credit card transaction is fraud or genuine

Motivation

Motivation for choosing dataset: To learn logistic regression. In this focus, we have tried to incorporate everything pertaining to data analysis for example reviewing data, checking missing values, and its treatment, doing descriptive analysis and applying logistic regression and finally evaluating model performance.
Motivation for choosing python: As in class we tasted the flavor of R, we wanted have hands on using Python as well. That is why, we did the analysis using Python and associated libraries.

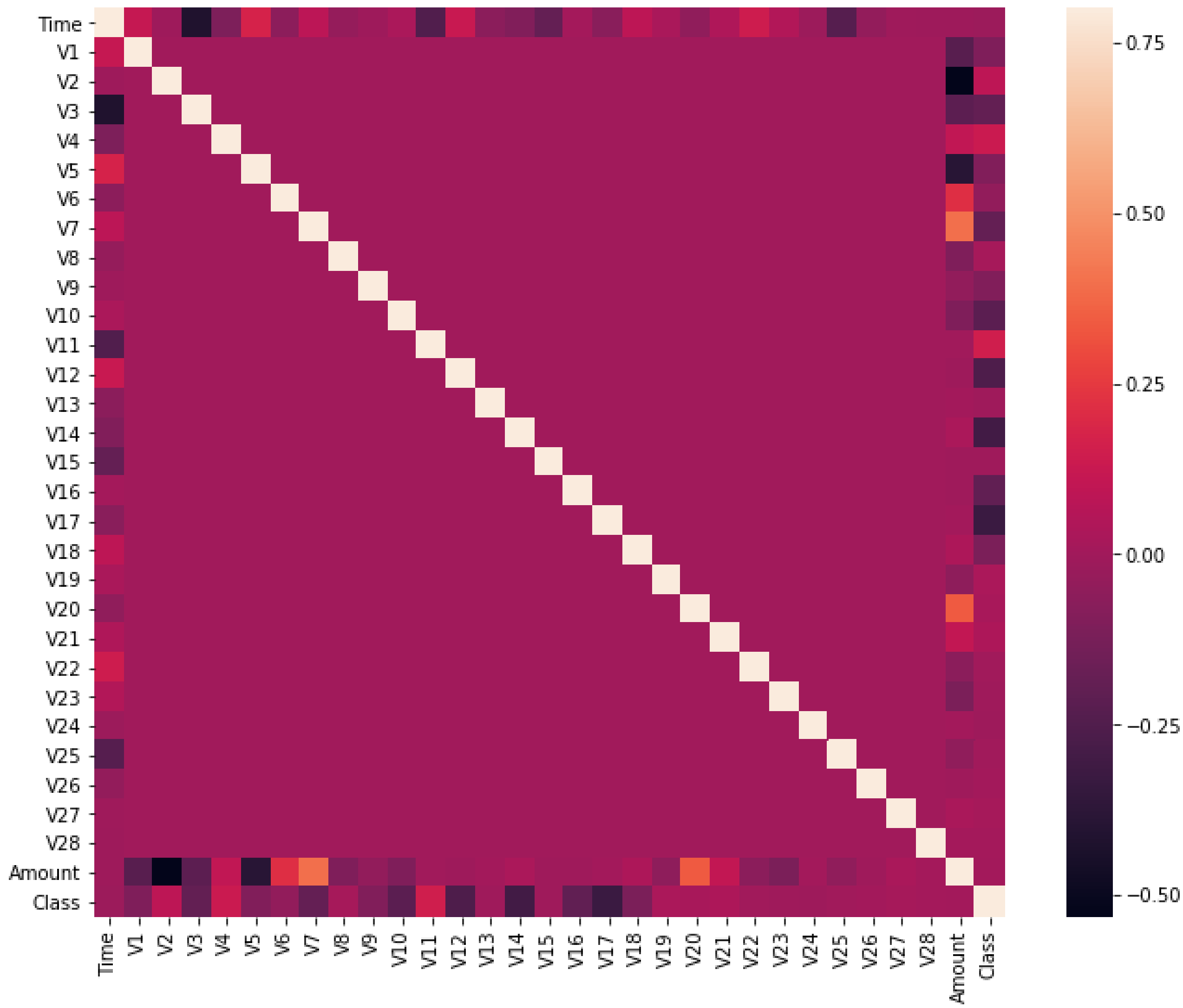
Descriptive Analysis

Fig1: Overall Data Comparison



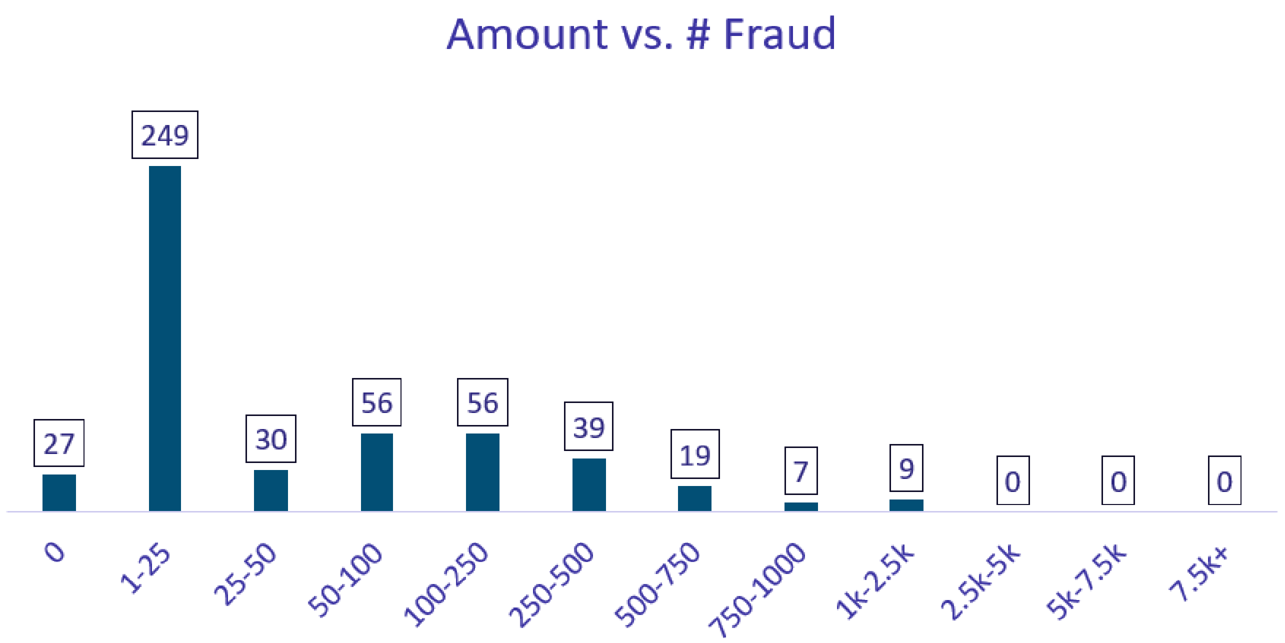
When comparing the histograms of the data above, it is clearly observed that most of the histograms are densely populated around zero. Total #Fraud cases is 492 (0.17%) of 284,315 transaction

Fig 2: Correlation Heat Map



The heatmap clearly shows different correlation between various V# parameters and Time, Amount and Class. Observing from the heatmap, it can be said that most data in a one to one relation are fairly similar in the point that the mean is at ‘0’. There are discripency when it comes to a comparison between the Class and other V columns.

Fig 3: Amount vs #Frauds



Here, we see an interesting trend. We find that there are 27 frauds with $0 ticket size. More than 50% of fraud transactions done for $1 to $25, small ticket transactions. And rest, above $25. Interestingly, there is no fraud observed for bigger transactions. Highest amount of fraud is recorded as $2,526.

Predictive Analysis

As the target variable is bi-variate, we decided to approach this problem with logistic regression.

Step 1: Resampling: We used underscore method for resampling as we had only 0.17% event rate. We tried different sampling ratios for fraud/ non-fraud for best model selection and tried to keep event rate between 2%-5% to get more events to build good model

Step 2: Train Test Split: We split the data into 70:30 ratio for training and testing

Step 3: Normalize data: We normalize the data so that all columns can have similar influence (Example: V1 vs Amount, V1 is between -10 to 10 whereas Amount is between $0 to ~ $25,000)

Step 4: Model Creation: Created logistic models using different C-parameter and classifying probabilities

Step 5: Model Validation: Based on confusion matrix, classification reports and ROC curve, choose optimal model for fraud detection

Snapshot of Model Creation:

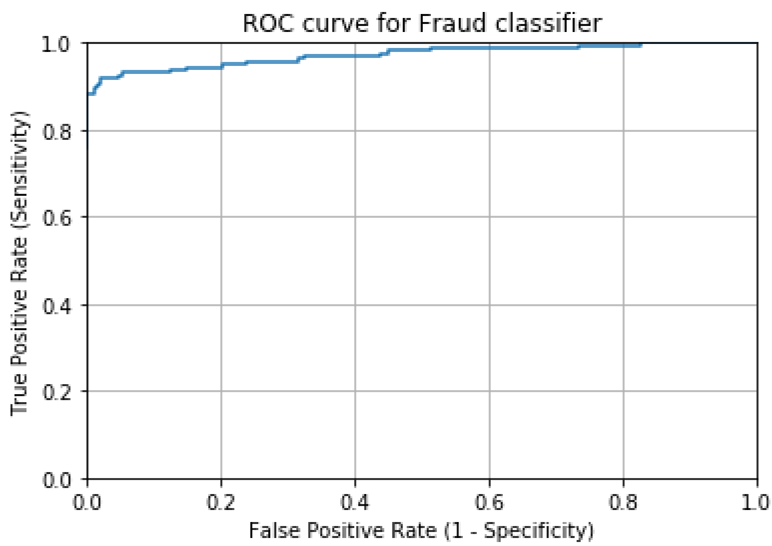
Step1: Resampling: Divided dataset into fraud vs non-fraud in 1:19 ratio

Step 2: Train Test Split: Spited data in train vs test in 70:30 ratio

Step 3: Normalized the data

Step 4: Model Creation using following C-parameter and Classifying probabilities

- C parameter (.01, 0.1,1,10,100,1000)
- Probabilities (0.025,0.05,0.075,0.1,0.125,0.150,0.175,0.2,0.5).
  - 0.5 is default
  - Selected these probabilities based on roc curve



Step 5: Model Validation

From observation, best fit for model with 5% event rate is at

- C-parameter =1000
- Probability at 20%. Means probability <=20.00% classified as 0 (non-fraud) vs probability>20.00% classified as 1 (fraud)

- Confusion Matrix:

	Classified as Not Fraud	Classified as Fraud
True Not Fraud	2,780	17
True Fraud	14	141

- Classification Report:

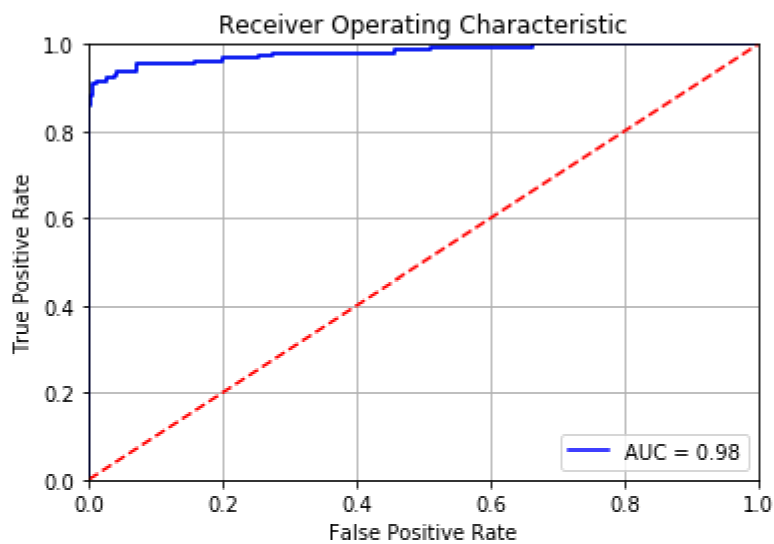
	precision	recall	f1-score	support
Not Fraud	0.99	0.99	0.99	2,797
Fraud	0.88	0.91	0.89	155
avg/ total	**0.99**	**0.99**	**0.99**	**2,952**

Fraud Recall = 91%

Fraud Precision = 88%

Overall Model Accuracy: 99%

- AUC = 98.13%



Summary

- **With the help of logistic regression, we can now predict 92% of the fraud compared to no fraud detection earlier, mapping customer behavior and deterring fraud eventually saving $51K (~85% of 60K) loss due to fraud.**

Future Scope

1. Perform the same task with outlier treatment
2. Explore other classification-based techniques such as SVM, Random Forest or Local Outlier Factor, etc. and make model more robust