# Credit Card Fraud Detection


Rajasekaran, Sudharsana

# Contents

Introduction ..... 2
What is Credit card Fraud? ..... 2
Why is credit card Fraud Detection Important? ..... 2
Exploratory Data Analysis: ..... 3
Classification Models and their performances: ..... 4
Classification using Logistic Regression ..... 4
Choosing a best logit Model using AIC: ..... 4
K- Fold Cross Validated Logit Model: ..... 5
Support Vector Machines: ..... 7
Neural Networks as a classifier: ..... 8
Neural Network Architecture: ..... 8
Other Classification Models: ..... 9

## Introduction

Credit Card Transactions are increasing globally every year, so it is important for the credit card companies to identify and flag those credit card transactions so that customers are not charged for items that they did not purchase. There are different algorithms that can help us flag and track these fraudulent transactions, so we prevent any unauthorized use of the cards.

## What is Credit card Fraud?

Credit card fraud is a wide-ranging term for theft and fraud committed using or involving a payment card, such as a credit card or debit card, as a fraudulent source of funds in a transaction. Credit card fraud is also the unauthorized use of another person's credit card information—to make purchases or access funds through cash advances using the victim's account.

## Why is credit card Fraud Detection Important?

Although incidences of credit card fraud are limited to about 0.1% of all card transactions, they have resulted in huge financial losses. Nilson Report estimates that in 2016, losses topped $24.71 billion. That represents a 12% increase over the previous year.

As online shopping and bill paying has skyrocketed in popularity, there is no longer a need to possess a physical credit card or debit card to make purchases, and it is possible even to open a financial account and obtain credit cards solely through online transactions. Because of this, criminals able to obtain enough personal information about other individuals may use that information to commit credit card fraud by opening new accounts, or having new cards sent to them on existing accounts. Therefore, there is an immense need for the credit card companies to identify the fraudulent transactions so they will be able to keep their customer satisfaction high and losses minimal.

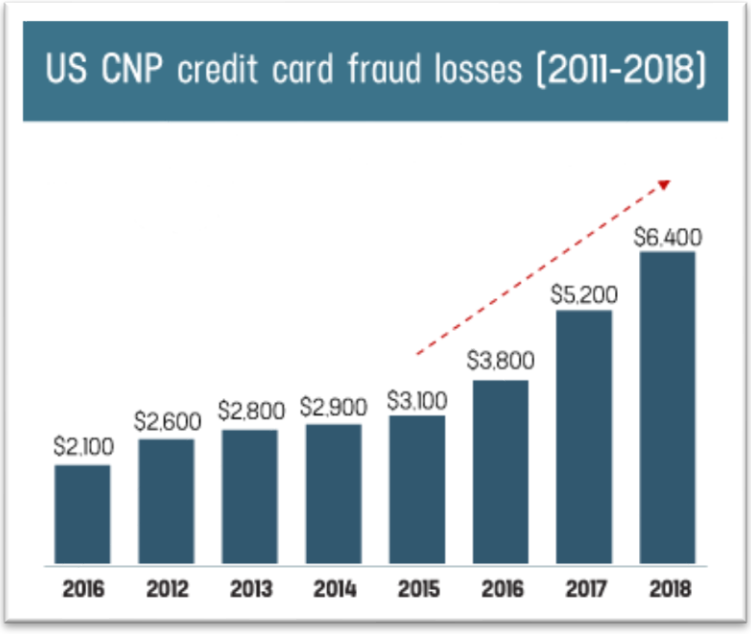


*Figure 1. Year on Year Credit Card Fraud Losses*

## Exploratory Data Analysis:

The dataset contains credit card transactions made by European cardholders. Initial dataset analysis yielded the following results,

1. **Number of Transactions:** 284807 credit card transactions are present in the dataset. There are no missing values in the dataset.
2. **Distribution of the Dataset:** The Dataset is highly unbalanced with only 492 Fraudulent transactions accounting for only 0.172% of all the transactions in the dataset.

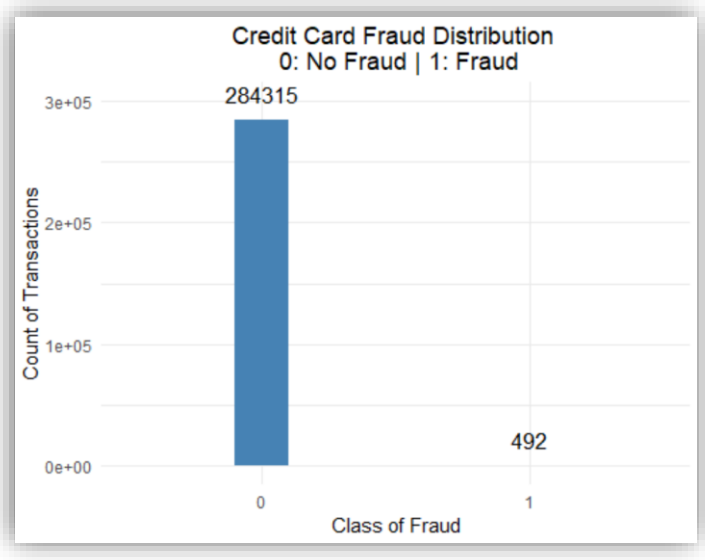


*Figure 2: Credit Card Fraud Distribution*

3. **Correlation between variables:**
   The data set consists of numerical values of 28 original features that are masked using PCA transformation, V1 - V28. Hence, there are no observed from the correlation between the variables.

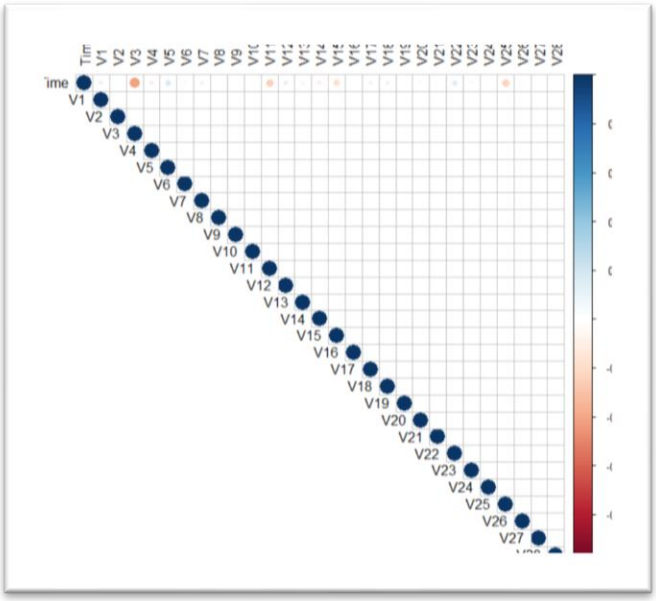


*Figure 3: Correlation Plot*

# Classification Models and their performances:

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y). In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation.

## Classification using Logistic Regression

Logistic Regression is a linear classifier and is one of the simplest methods for the go to method for binary classification problem. Logistic regression is intended for two-class classification problems. It will predict the probability of an instance belonging to the default class, which can be snapped into a 0 or 1 classification.

### Choosing a best logit Model using AIC:

Choosing a best model is very important for detecting the fraud with high level of accuracy. We chose the best logit model for this classification using the one with lowest AIC score. The model parameters can be found below

*Model Parameters:*

```
Call:
glm(formula = Class ~ V1 + V4 + V5 + V6 + V8 + V9 + V10 + V13 +
    V14 + V16 + V20 + V21 + V22 + V23 + V27 + V28 + Amount, family
= binomial,
    data = cc_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0292  -0.0298  -0.0198  -0.0128   4.5686

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.6508124  0.1484426 -58.277  < 2e-16 ***
V1           0.0642712  0.0318570   2.017 0.043644 *
V4           0.7064997  0.0686232  10.295  < 2e-16 ***
V5           0.0968806  0.0455432   2.127 0.033402 *
V6          -0.1152701  0.0753858  -1.529 0.126247
V8          -0.1971174  0.0299773  -6.576 4.85e-11 ***
V9          -0.1793705  0.0917709  -1.955 0.050637 .
V10         -0.8196306  0.0990061  -8.279  < 2e-16 ***
V13         -0.3144415  0.0869483  -3.616 0.000299 ***
V14         -0.4758231  0.0576275  -8.257  < 2e-16 ***
V16         -0.3142867  0.0662015  -4.747 2.06e-06 ***
V20         -0.3541606  0.0639991  -5.534 3.13e-08 ***
V21          0.3540675  0.0624777   5.667 1.45e-08 ***
V22          0.5666220  0.1399249   4.049 5.13e-05 ***
V23         -0.1454916  0.0587297  -2.477 0.013238 *
V27         -0.6821264  0.1370091  -4.979 6.40e-07 ***
V28         -0.2188387  0.0952007  -2.299 0.021521 *
Amount       0.0003759  0.0001594   2.358 0.018390 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5799.1  on 227845  degrees of freedom
Residual deviance: 1787.4  on 227828  degrees of freedom
AIC: 1823.4

Number of Fisher Scoring iterations: 11
```

*Confusion Matrix*

Owing to such imbalance in data, an algorithm predicts most of the transactions as non-frauds and so will also achieve an accuracy of 99.828%. Hence, accuracy is not a correct measure for validating the performance of the model in this case. The correct performance measure would be recall. Though the model does exceptionally well in identifying 64 fraudulent cases, there are 7 cases where it was misidentified. These seven cases pose severe threat as they were fraudulent and marked as legitimate transactions.

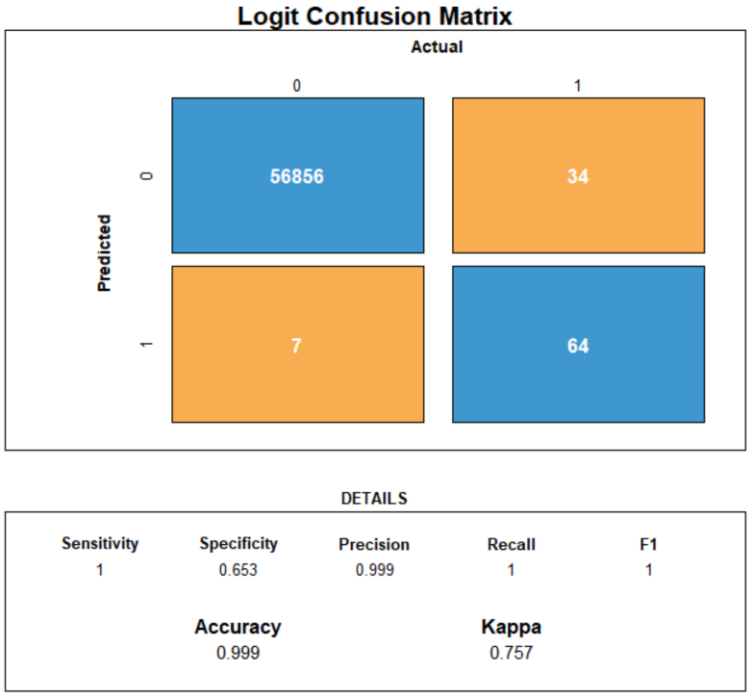


*Figure 4: Confusion Matrix of Logistic Regression*

## K- Fold Cross Validated Logit Model:

In order of get more efficient β coefficients, the chosen logit model is run through a 10-Fold cross validated to improved robustness of the model. For logistic regression models unbalanced training data affects only the estimate of the model. Therefore, the model was not able to improve specificity even after accounting for cross validation. The parameters of the model can be found below,

*Model Parameters:*

```
Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0292  -0.0298  -0.0198  -0.0128   4.5686

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.6508124  0.1484426 -58.277  < 2e-16 ***
V1           0.0642712  0.0318570   2.017 0.043644 *
V4           0.7064997  0.0686232  10.295  < 2e-16 ***
V5           0.0968806  0.0455432   2.127 0.033402 *
V6          -0.1152701  0.0753858  -1.529 0.126247
```

```
V8          -0.1971174  0.0299773  -6.576 4.85e-11 ***
V9          -0.1793705  0.0917709  -1.955 0.050637 .
V10         -0.8196306  0.0990061  -8.279  < 2e-16 ***
V13         -0.3144415  0.0869483  -3.616 0.000299 ***
V14         -0.4758231  0.0576275  -8.257  < 2e-16 ***
V16         -0.3142867  0.0662015  -4.747 2.06e-06 ***
V20         -0.3541606  0.0639991  -5.534 3.13e-08 ***
V21          0.3540675  0.0624777   5.667 1.45e-08 ***
V22          0.5666220  0.1399249   4.049 5.13e-05 ***
V23         -0.1454916  0.0587297  -2.477 0.013238 *
V27         -0.6821264  0.1370091  -4.979 6.40e-07 ***
V28         -0.2188387  0.0952007  -2.299 0.021521 *
Amount       0.0003759  0.0001594   2.358 0.018390 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5799.1  on 227845  degrees of freedom
Residual deviance: 1787.4  on 227828  degrees of freedom
AIC: 1823.4

Number of Fisher Scoring iterations: 11
```

*Confusion Matrix*

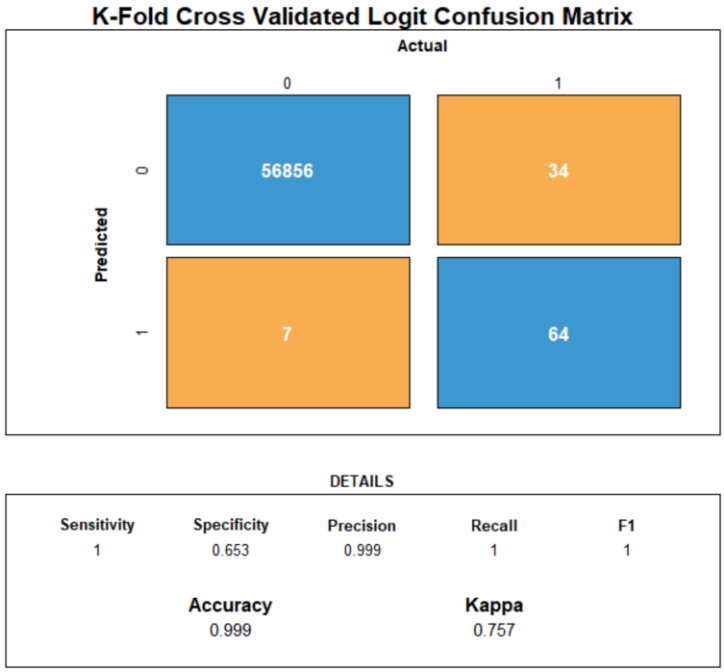


*Figure 5: Confusion Matrix of K-Fold cross validated Logistic Regression*

## Support Vector Machines:

SVM is a very good classification model for classifying highly unbalanced datasets. n SVM classifier trained on an imbalanced dataset can produce suboptimal models which are biased towards the majority class and have low performance on the minority class.

```
Call:
svm(formula = Class ~ ., data = cc_train, kernel = "radial", cost = 10,
    gamma = 0.1)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  10

Number of Support Vectors:  9346

 ( 8975 371 )


Number of Classes:  2

Levels:
 0 1
```

### *Confusion Matrix:*

Any misidentification of Fraudulent cases is highly important as this may in turn cause huge financial losses. SVM model categorizes the fraudulent cases to a good accuracy and minimizes the risk of identifying the Fraudulent transactions as legitimate transactions. But on the other hand, the number of misclassified normal cases have increased considerably as compared to the logistic regression. Although this considerably affects the accuracy, this case does not cause huge impact as that of the misidentification of fraudulent ones.

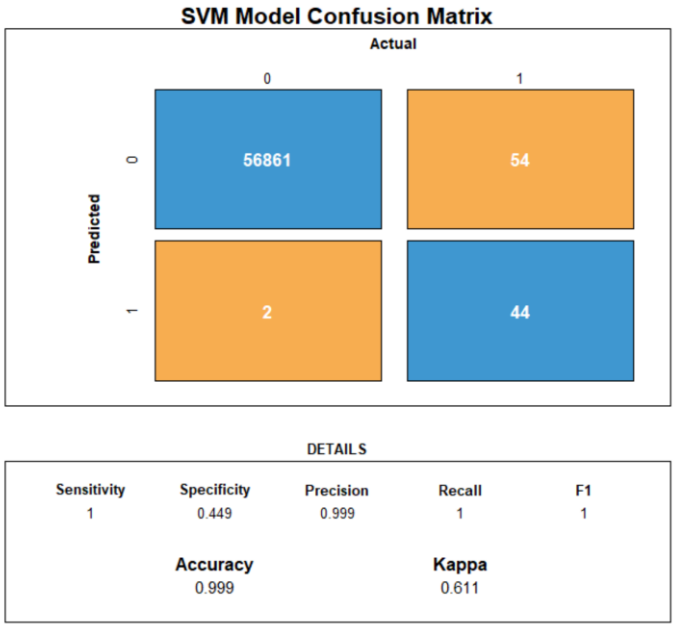


*Figure 6: Confusion Matrix of SVM Model*

## Neural Networks as a classifier:

Artificial neural networks (ANN) are computing systems that are inspired by, but not necessarily identical to, the biological neural networks that constitute animal brains. Neural networks consist of input and output layers, as well as a hidden layer consisting of units that transform the input into the output layer can use. Neural networks are excellent machine learning technique for classification and pattern recognition.

## Neural Network Architecture:

The neural network architecture is as follows

- **Input Layer:** The input layer has 30 input nodes for the initial data into the system for further processing by subsequent layers of neurons.
- **Hidden Layers:** There are 3 hidden layers in the network. Each of the 3 hidden networks has 20, 15 and 5 nodes respectively.
- **Output Layer:** The output layer sigmoid activated binary classification output of either 0 or 1.

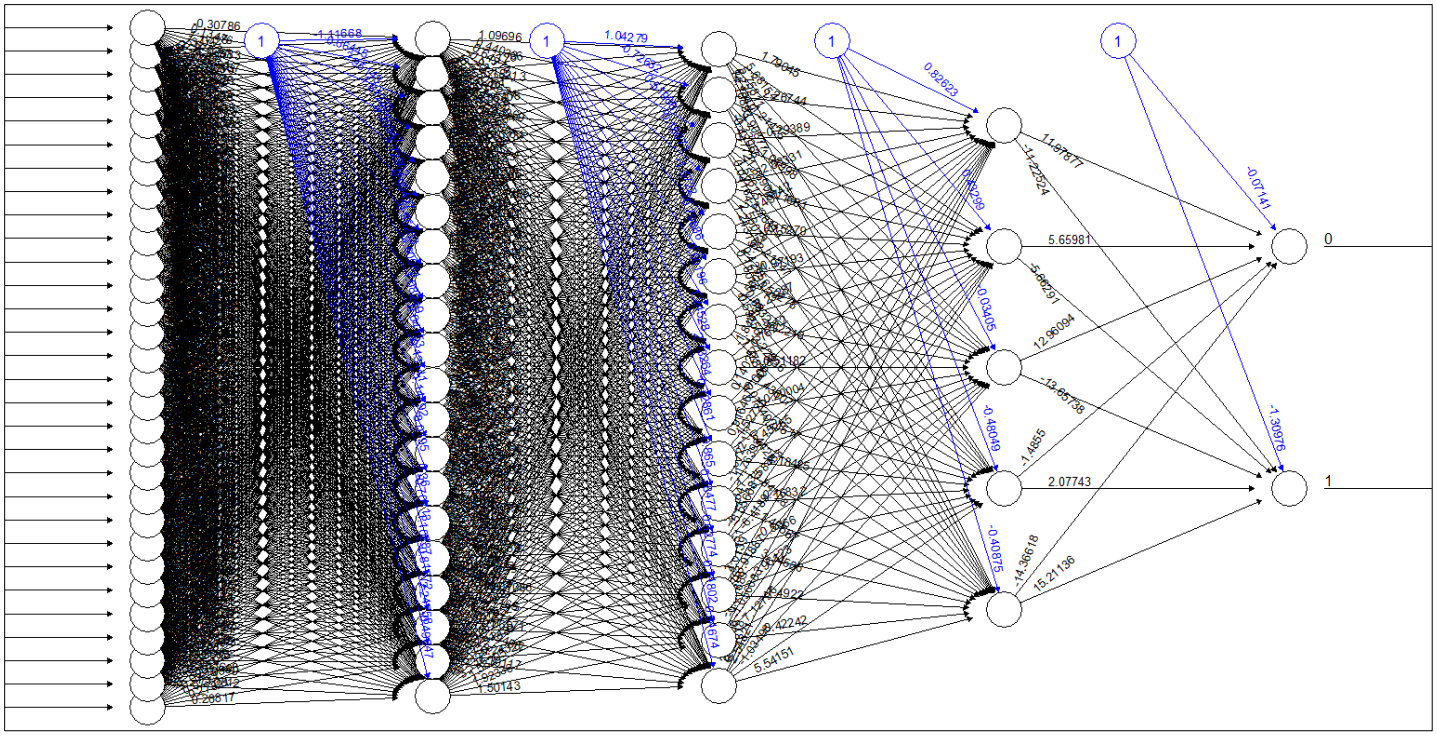


*Figure 7: Neural Network Architecture*

### *Confusion Matrix*

The overall accuracy of a Neural network is high as compared to logistic regression or SVM. Misidentification of Fraudulent and Non-Fraudulent cases have been greatly reduced. Neural Network proving to be an excellent classification algorithm.

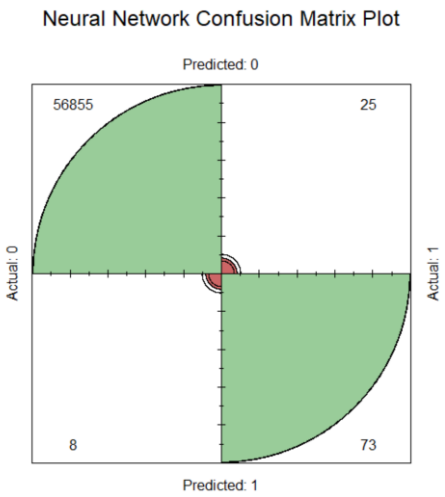


*Figure 8: Confusion Matrix plot of Neural Network*

## Other Classification Models:

There are a ton of machine learning models for classification. Few other techniques that help us classify and prevent fraudulent credit card transactions are

1. Naïve Bayes Classification Model
2. Decision Trees
3. Random Forest
4. KNN Classification