Data Analytics Project

(Shree) Ravi Adhikari

2023-03-11

Section One: Business Problem

Credit Card fraud is any unauthorized use of a credit card. These can range from stolen physical cards to authorized payments using proper credentials. While most of the cardholders can avoid financial liability, the businesses end up suffering from the fraud. This project aims to provide a solution for the early detection and mitigation of such transactions.

Section Two: Data Source

The dataset used for this paper can be found at *data-flair's website*. Alternatively, the dataset is also hosted at my *github repository* for permanent access.

Section Three: Downloading, Decompressing and Importing the dataset

Subsection Three One: Downloading, and Decompressing the dataset

The tools used to download and decompress the dataset are linux commands `curl` and `7zip`.

```
# download from the remote URL
curl https://github.com/shreeraviadhikari/ANA515/raw/master/AssignmentFour/creditcard.7z

# Extract the archive
7z x creditcard.7z
```

Subsection Three Two: Importing the dataset

The following code imports and loads the csv data into `initial_dataset` variable.

```
initial_dataset = read.csv('./creditcard.csv')
```

Section Four: Data Description

The dataset has 284807 entries with 31 attributes. The variables of interest are: time, amount and classification (fraudulent transactions are marked as 1 and non-fraudulent transactions are marked as 0). Other 28 variables are cryptic and represent computed scores. They are labelled V1 to V28. The summary for variables of interest are presented below:

Variables Summary

Variable	Type	Average	Minimum	Maximum	Null Values
Time	Numeric	9.481386×10^4	0	1.72792×10^5	0
Amount	Decimal	88.3496193	0	2.569116×10^4	0
Classification	Label	0.0017275	0	1	0

The number of fraud (1) and non-fraudulent(0) transactions are shown below:

```
##
##      0      1
## 284315    492
```

Section Five: Data Preparation

Data Normalization

All the input columns except `Time` were normalized for faster and efficient computation. The step for `Amount` column is shown below:

```
initial_dataset$Amount = scale(initial_dataset$Amount)

dataset = initial_dataset[,-c(1)]
```

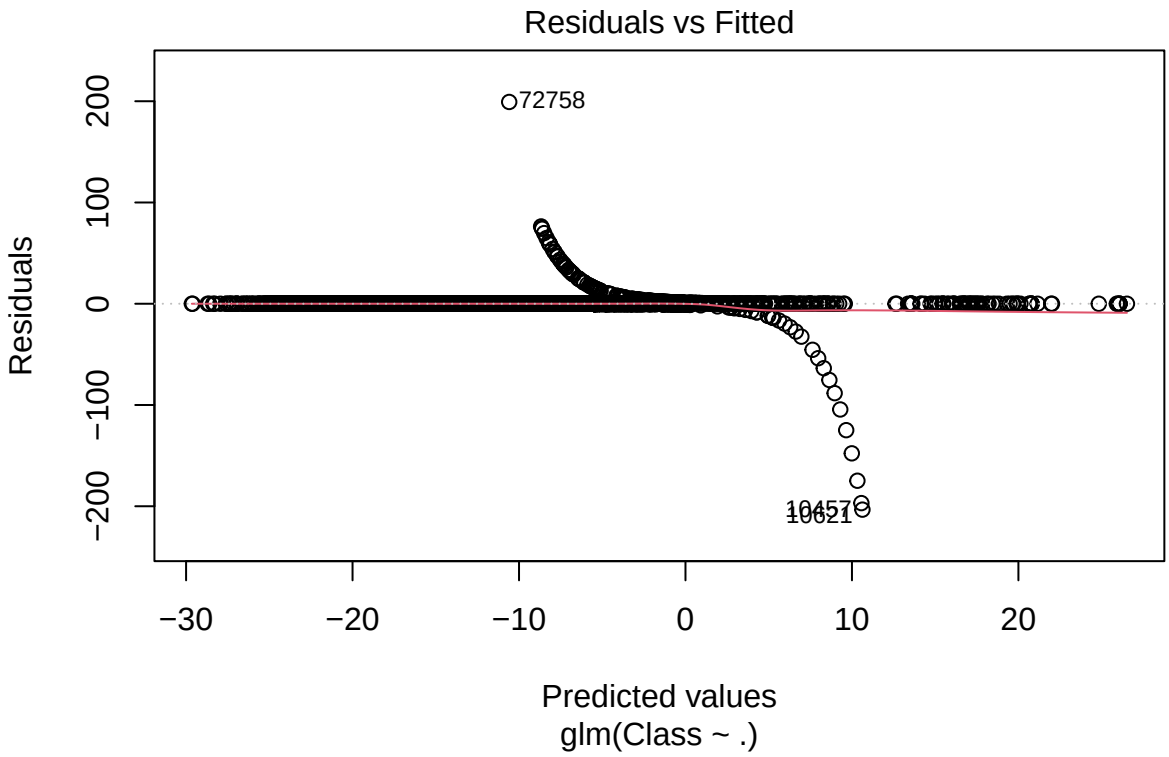
During the data preparation, all the collected values were made error free, and entries with missing values were ignored. All of the variables were converted into numeric form for dimension reduction and personal details were blinded. Some of the converted values include:

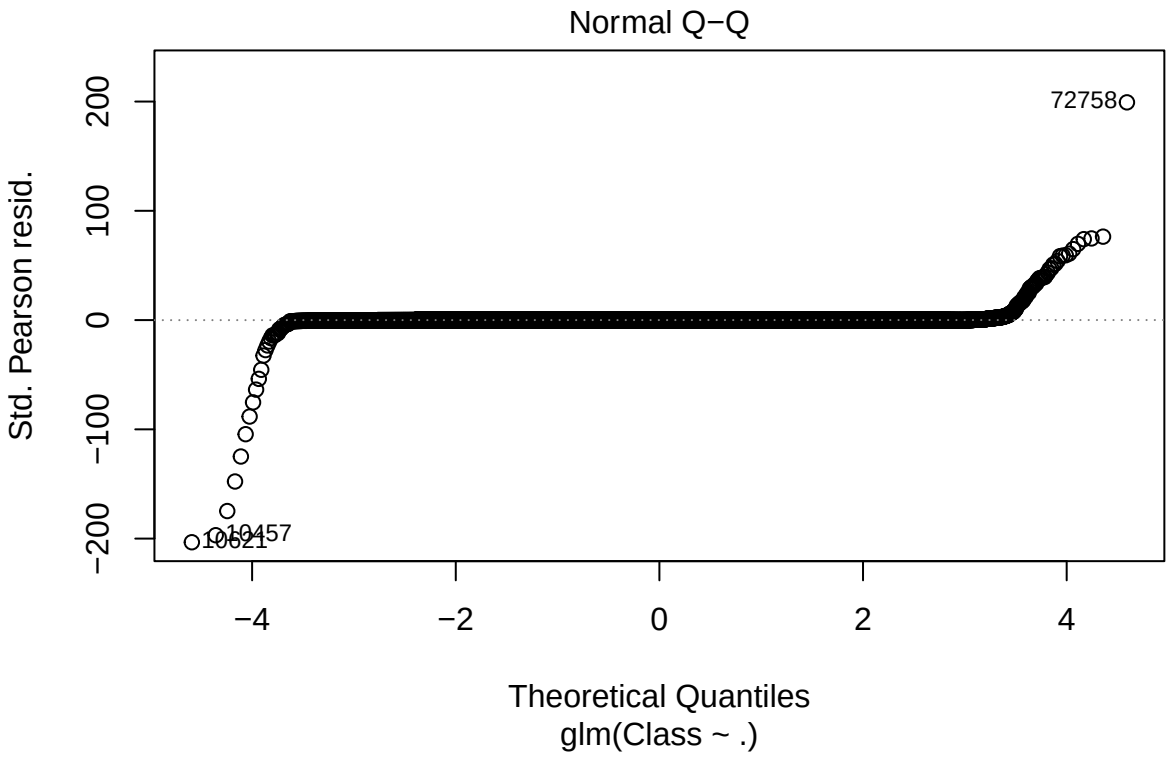
- User-Agent: The agent of transaction (for example: Physical Card, Apple Pay, Virtual transaction)
- Location: The location of transaction
- Remarks: the remarks for transaction

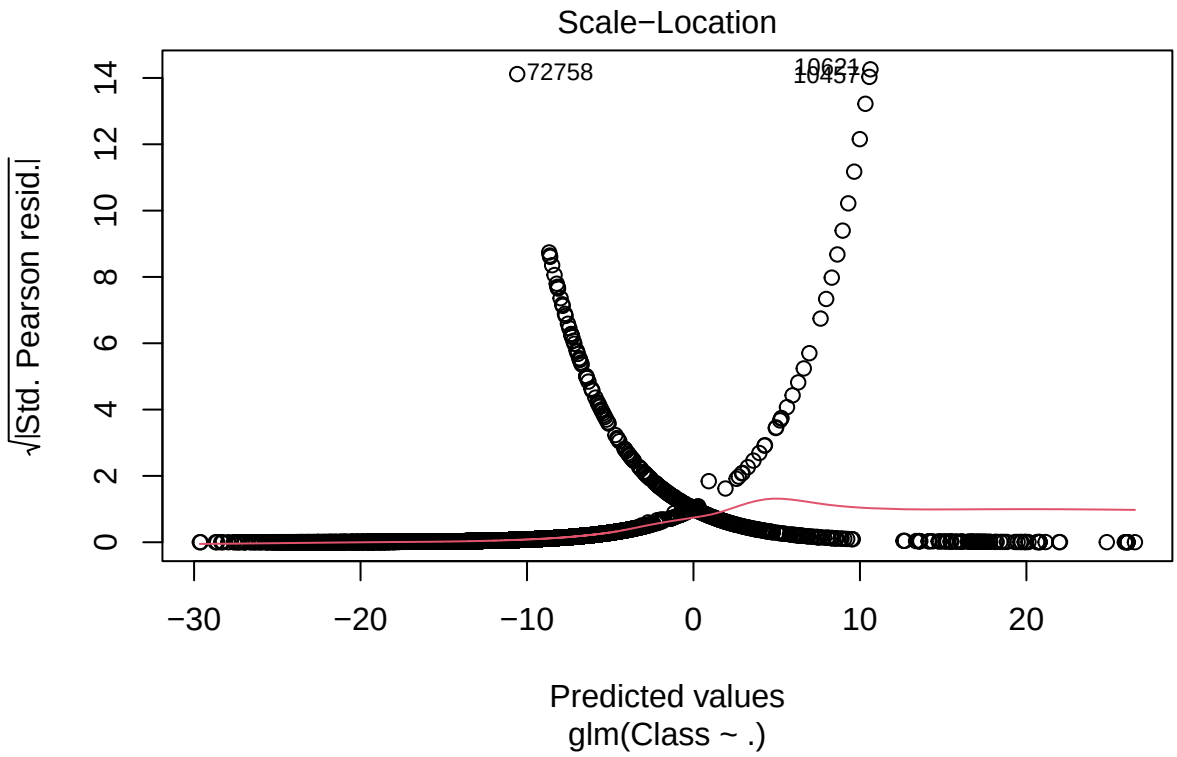
Section Six: Data Modeling

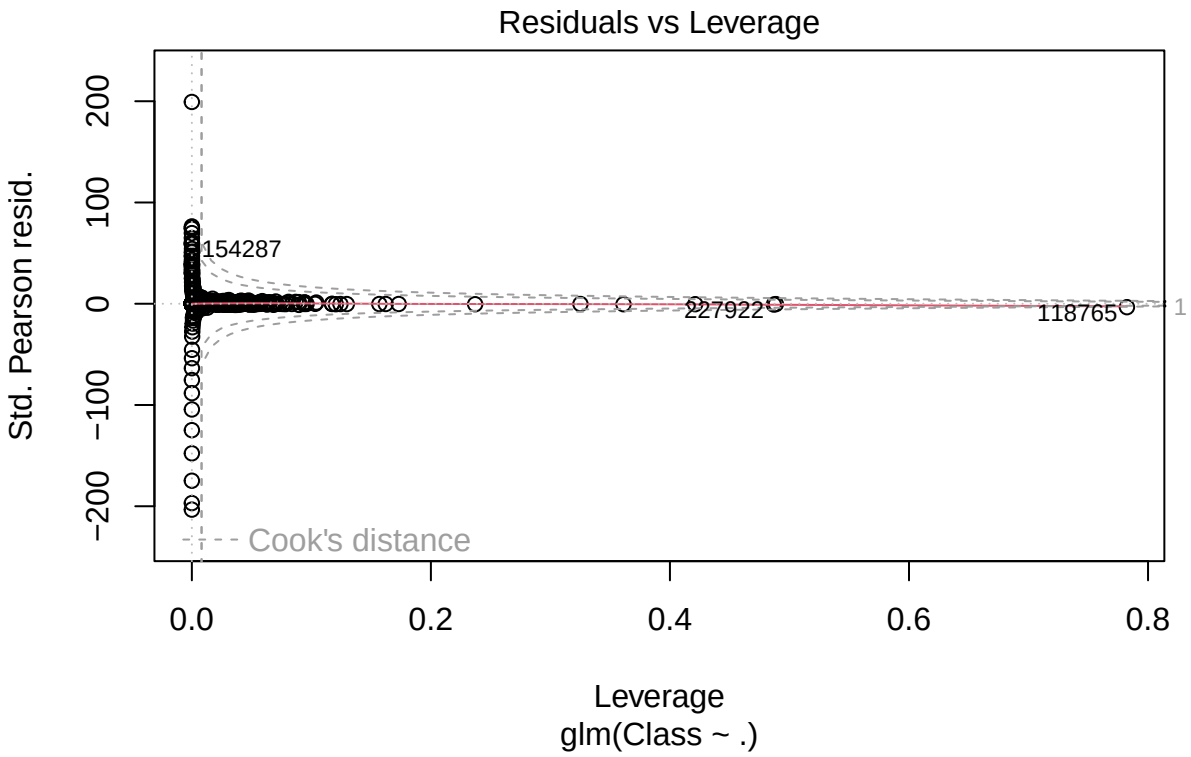
The dataset is broken down into 80:20 ratio for train and test. The first set will be used to train the model, and the later to learn how well the system performs for a different dataset.

The model selected is Logistic Regression. The Logistic Regression is one of the simple but good model for binary classification problems. The goal here is to classify the transactions as Fraudulent or Non-Fraudulent.

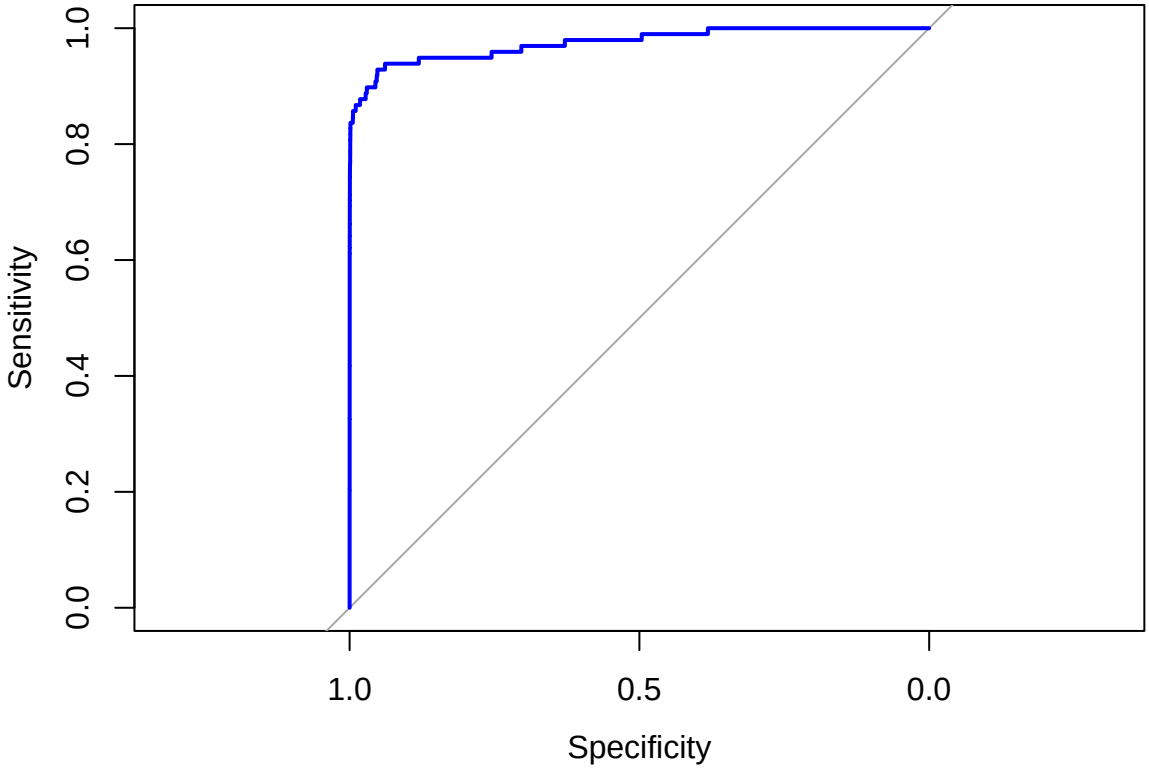

Residuals vs Fitted
72758
10457
10621
Residuals
200
100
0
−100
−200
−30
−20
−10
0
10
20
Predicted values
glm(Class ~ .)


Normal Q−Q
Std. Pearson resid.
200
100
0
−100
−200
72758
10457
10621
−4
−2
0
2
4
Theoretical Quantiles
glm(Class ~ .)


Scale−Location
72758
√|Std. Pearson resid.|
0
2
4
6
8
10
12
14
−30
−20
−10
0
10
20
Predicted values
glm(Class ~ .)


Residuals vs Leverage
Std. Pearson resid.
200
100
0
−100
−200
154287
227922
118765
1
Cook's distance
0.0
0.2
0.4
0.6
0.8
Leverage
glm(Class ~ .)

Part Seven: Results



The AUC measures the false positives and true negatives in a 2x2 grid. The AUC graph shows 97.48 % accuracy.

Part Nine: Summary

Credit Card Frauds are one of the most occurring frauds in the United States. The goal of this project (i.e. classification of fraudulent and non-fraudulent transactions) was successfully demonstrated and the resulting outputs were discussed above.

Part Ten: Recommendations

This project although successfully implements the credit card classification problem, it is nowhere close to solving the real problem. During the development of this project, few key points were realized:

1. The classification problem for Logistic Regression requires tremendous amount of data.
2. There needs to be sufficient cases for both fraudulent and non-fraudulent transactions.
3. A number of other algorithms (e.g. Decision Tree, Neural Networks, SVM) needs to be tested and compared alongside Logistic Regression for a holistic approach to solving the business needs.

Part Eleven: References

1. Machine Learning Crash Course/ Classification
2. Data Flair Credit Card Fraud Detection
3. Generalized Linear Models