Credit Card Fraud Detection

Final Project Report

Instructor: Samantha-Jo Caetano

Yuhan Zhu

Github Link:https://github.com/juliaz1231/final-proj.git

Date: 12/22/2020

Abstract

In this project, a predictive model has been produced for credit card fraud detections. The accuracy score for the model is high enough for using. The reason to predict credit card fraud activities is to secure people's property and money.By doing so, banks can gain more trust from people.

KeyWord

Data Cleaning, Credit Card Fraud Detection, Predictive Model, Logistic Regression, Linear Regression,Accuracy Score, Precision-Recall Curve

Introduction

Big data has occupied almost the entire market in every industry, including credit card companies. They use predictive analysis to detect transactions that contain potential risks. According to the government of the United States, Credit card fraud is defined as "an unauthorized use of a credit or debit card, or similar tool, to fraudulently obtain money or property." Recognizing fraud transactions can prevent customers from charging items they did not purchased.

According to the Federal Reserve Payments study, 26.2 Billion purchases in 2012 were made by US citizens' credit card usage, with a total value of $6.1 billion unauthorized transactions. In order to prevent the loss of dollars, credit cards, eCommerce companies and banks have decided to use big data to solve their problems.

After including big data technology into the system, credit card companies were able to detect credit card fraud transactions successfully. Algorithms can produce a possibility of fraud activities by analyzing an user's habit, location, currency and other attributes. The credit card will be declined if there is a suspicious transaction.

In this project, one dataset will be used to produce a predictive analysis for potential credit card fraud. In the Methodology section. I will describe the data, and the model that was used to perform the analysis. Results of the analysis are provided in the Results section and inferences of this data along with conclusions are presented in the Conclusion section. section.

Data and Methodology

Data: The dataset is collected by a research collaboration of Worldline and the Machine Learning Group of ULB on big data mining and fraud detection. It contains only numerical input variables and they are results of PCA transformation. Features V1,V2,V3,…,V28 are obtained with PCA. "Time" and "Amount" have not been transformed."Class" represents the response variable and "1" means fraud, "0" means not fraud.

Model: An exploratory data analysis is conducted first.Data preprocessing is important for furthur analysis, such as removing missing values or outliers.In order to understand the relationships among attributes easier, a corrplot is conducted because it is a graphical representation of data that allows people understanding the dataset easier by looking at it.

In order to build a predictive model, understanding the type of models is important. The problem is to detect whether a transaction is fraud, so the outcome is binary which means it should be either "Yes" or "No". Logistic regression will be applied in the project because When the dependent variable is binary, it is appropriate to apply logistic regression analysis.

Lastly, a Hosmer-Lemeshow Test is conducted to test for logistic regression. It is a goodness of fit test and it is especially good for logistic regression.

Results

Exploratory Data Analysis:

During the process of EDA, I first imported the dataset and named it as "df". Here is the head of the dataset:

	Time <dbl>	V1 <dbl>	V2 <dbl>	V3 <dbl>	V4 <dbl>	V5 <dbl>	V6 <dbl>	V7 <dbl>
1	0	-1.3598071	-0.07278117	2.5363467	1.3781552	-0.33832077	0.46238778	0.23959855
2	0	1.1918571	0.26615071	0.1664801	0.4481541	0.06001765	-0.08236081	-0.07880298
3	1	-1.3583541	-1.34016307	1.7732093	0.3797796	-0.50319813	1.80049938	0.79146096
4	1	-0.9662717	-0.18522601	1.7929933	-0.8632913	-0.01030888	1.24720317	0.23760894
5	2	-1.1582331	0.87773675	1.5487178	0.4030339	-0.40719338	0.09592146	0.59294075
6	2	-0.4259659	0.96052304	1.1411093	-0.1682521	0.42098688	-0.02972755	0.47620095

6 rows | 1-9 of 31 columns

Figure 1

There are total of 284807 observations of 31 variables. There is 0 missing values. Because the time and the amount have not been transformed. For better model performance, I removed the time variable and changed the "amount" variable by using scale() function. The new dataset was named as "df2" and it contains 284807 observations of 30 variables.

The total numbers of "0"s in the dataset is 284315 and the total number of "1"s is 492. Here is a barplot for both classes:

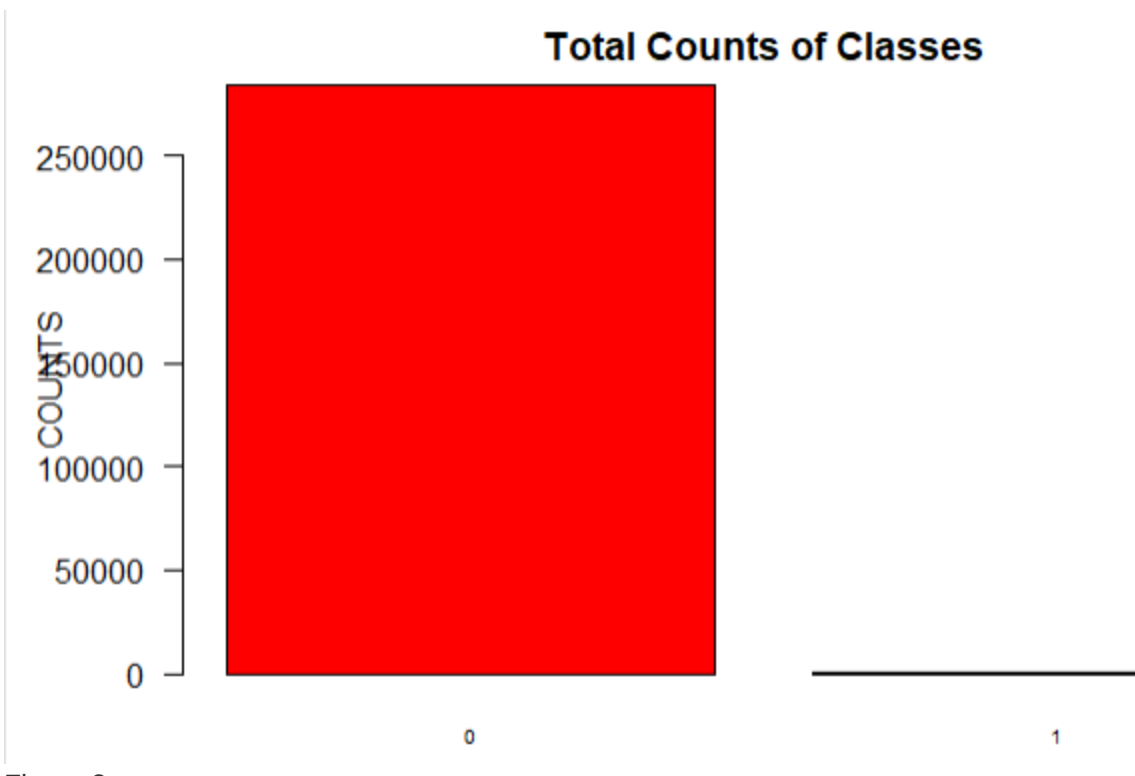

Figure 2

The next step is to find correlations among variables. A correplot was conducted by using "corrplot" function. It plots all the attributes and the plot is shown below:

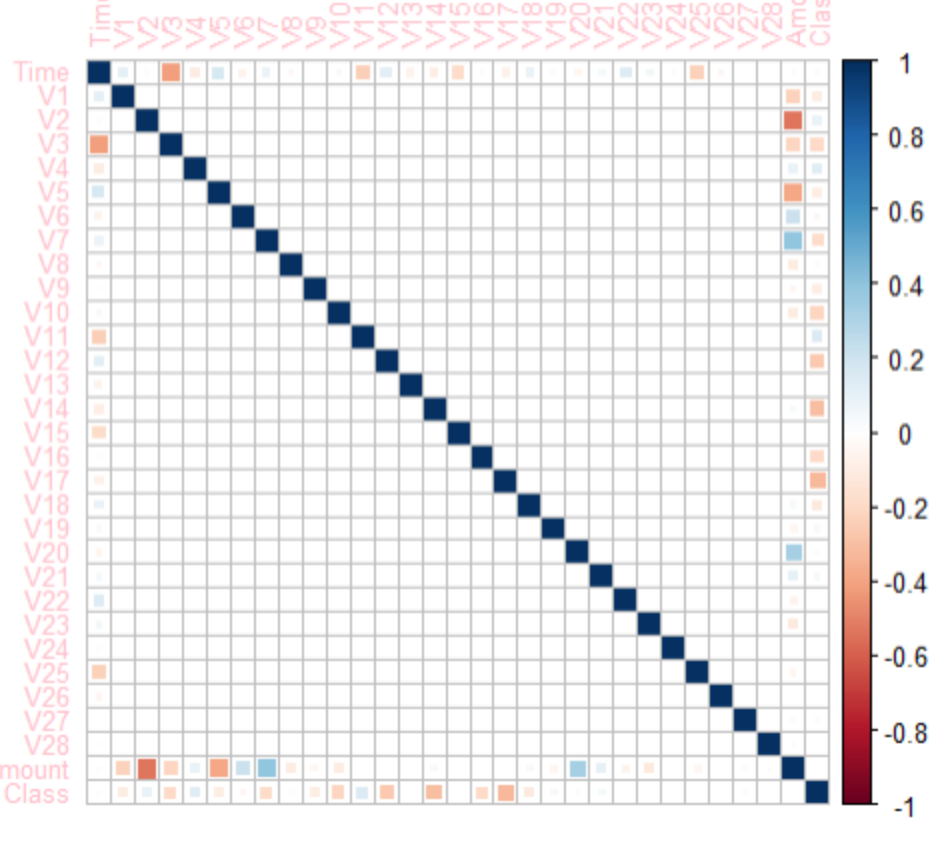

Figure 3

From the plot, I found that it is hard to tell if the correlation between variables.

Predictive analysis:

Now the data has been preprocessed. It is time to build a good predictive model. First step is to create the train and test dataset.I splitted the origional dataset by using ratio 7:3. So 70% of the dataset is training set, and 30% of the dataset is testing set.Then I generated a logistic regression model by using glm(Class~., train, family = binomial(link = "logit")). In this function, Class was the reponse variable and the rest are independent variable. Here is a summary for the model:

```
Call:
glm(formula = Class ~ ., family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.7488  -0.0284  -0.0187  -0.0120   4.6150

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.702632   0.175555 -49.572  < 2e-16 ***
V1           0.093320   0.051452   1.814 0.069720 .
V2           0.006254   0.068403   0.091 0.927150
V3           0.070974   0.057214   1.240 0.214795
V4           0.650667   0.086237   7.545 4.52e-14 ***
V5           0.066501   0.081086   0.820 0.412147
V6          -0.149830   0.099462  -1.506 0.131965
V7          -0.111501   0.081018  -1.376 0.168745
V8          -0.211723   0.037734  -5.611 2.01e-08 ***
V9          -0.136187   0.130579  -1.043 0.296972
V10         -0.729737   0.122592  -5.953 2.64e-09 ***
V11         -0.018804   0.090489  -0.208 0.835379
V12          0.014376   0.101725   0.141 0.887619
V13         -0.342201   0.095928  -3.567 0.000361 ***
V14         -0.639460   0.075174  -8.506  < 2e-16 ***
V15         -0.087865   0.101018  -0.870 0.384412
V16         -0.264923   0.144504  -1.833 0.066754 .
V17          0.045750   0.081781   0.559 0.575878
V18         -0.034233   0.147698  -0.232 0.816714
V19          0.063134   0.113935   0.554 0.579498
V20         -0.407258   0.106855  -3.811 0.000138 ***
V21          0.343893   0.068850   4.995 5.89e-07 ***
V22          0.609219   0.155898   3.908 9.31e-05 ***
V23         -0.162088   0.074803  -2.167 0.030245 *
V24          0.033355   0.182018   0.183 0.854600
V25          0.072300   0.154279   0.469 0.639333
V26         -0.070373   0.228543  -0.308 0.758143
V27         -0.864514   0.163244  -5.296 1.18e-07 ***
V28         -0.433169   0.149137  -2.905 0.003678 **
Amount       0.251559   0.116305   2.163 0.030548 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5064.6  on 199363  degrees of freedom
Residual deviance: 1522.3  on 199334  degrees of freedom
AIC: 1582.3

Number of Fisher Scoring iterations: 12
```

Figure 4

Then I plotted the model and here are four plots:

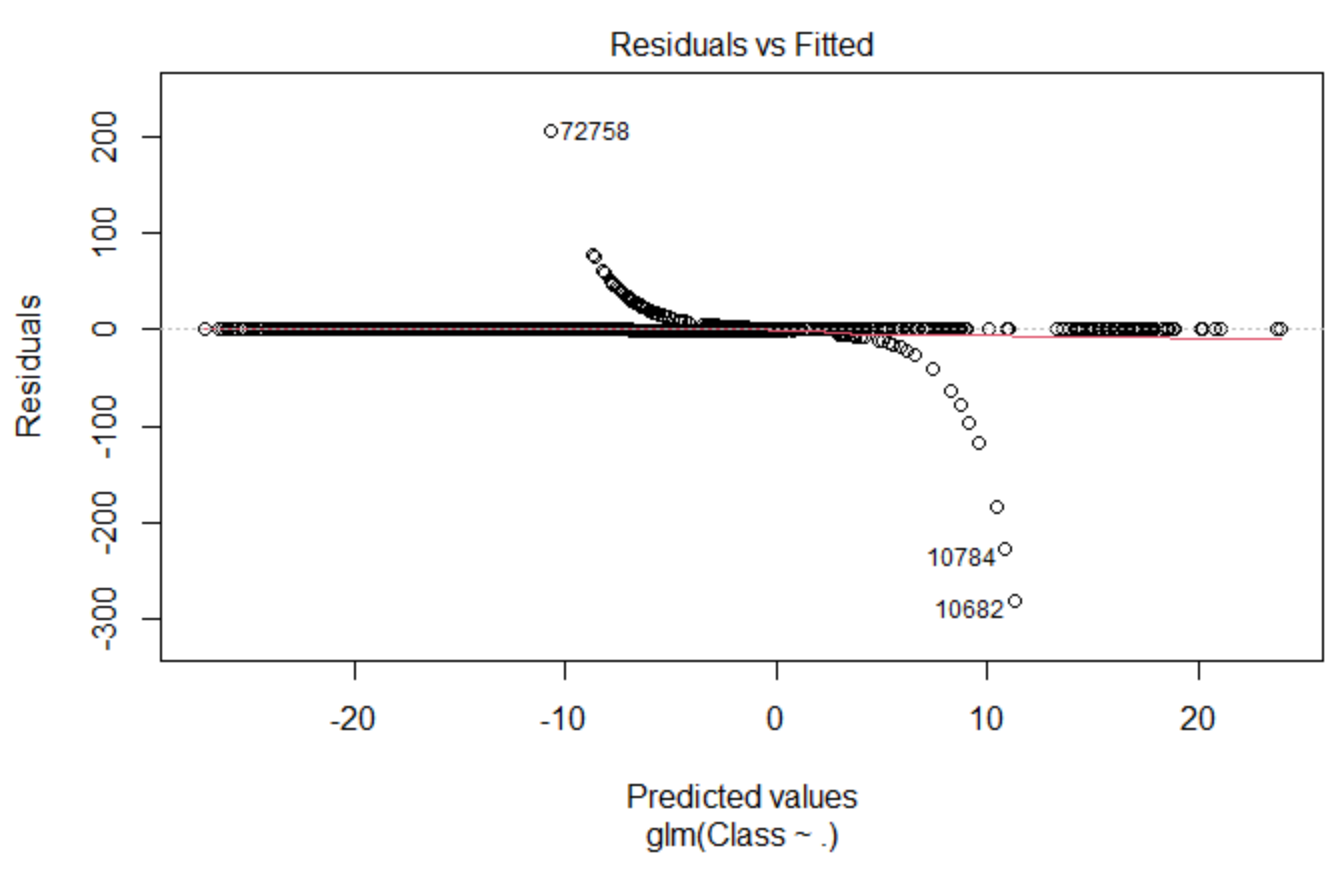

Figure 5

In the "Residuals vs fitted" plot, there are three out liers with large residual values.THe red line and the dashed line are almost perfectly attached.

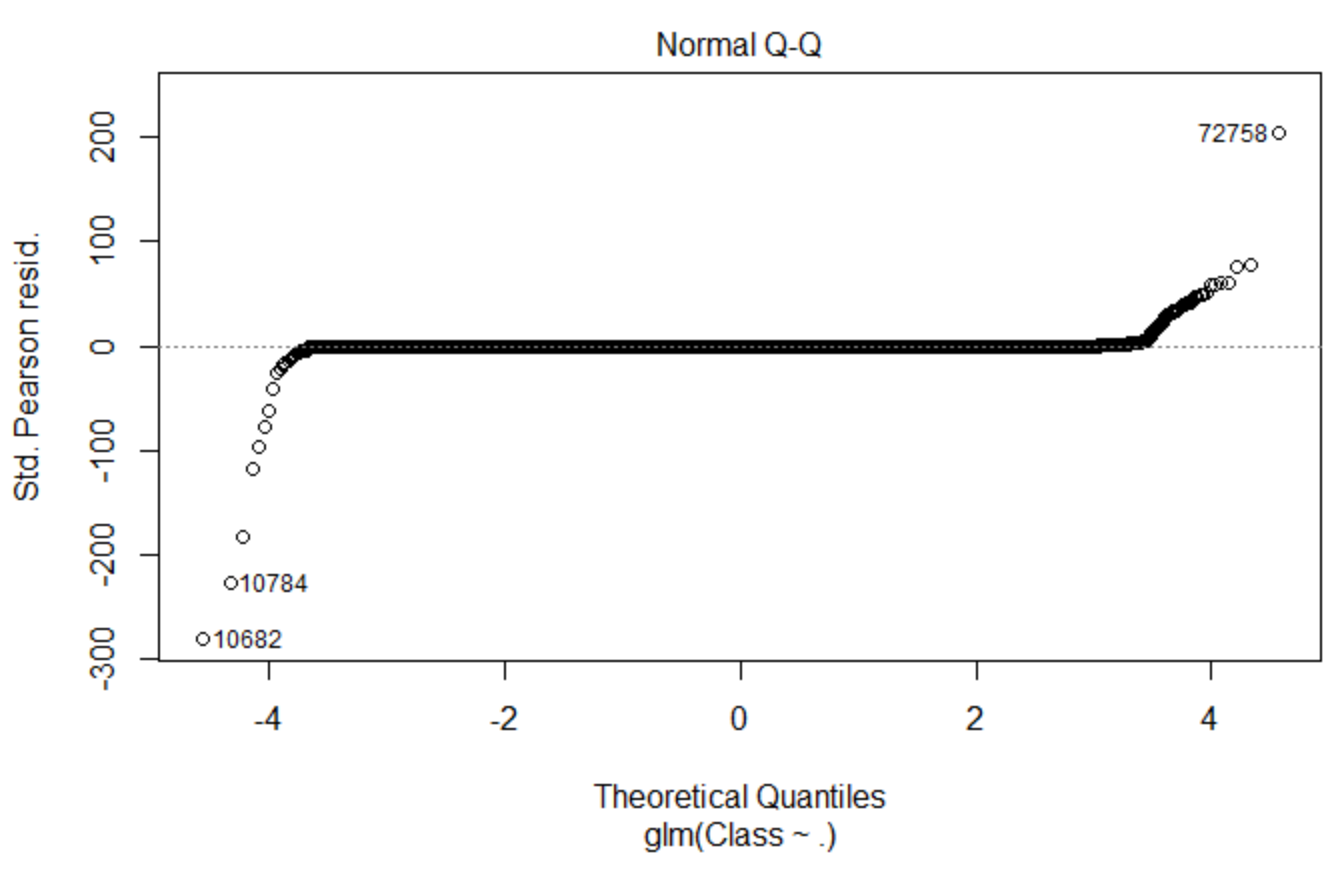

Figure 6

In the Normal Q-Q plot, it clearly shows that the distribution is not normal because it is heavily tailed at each end.

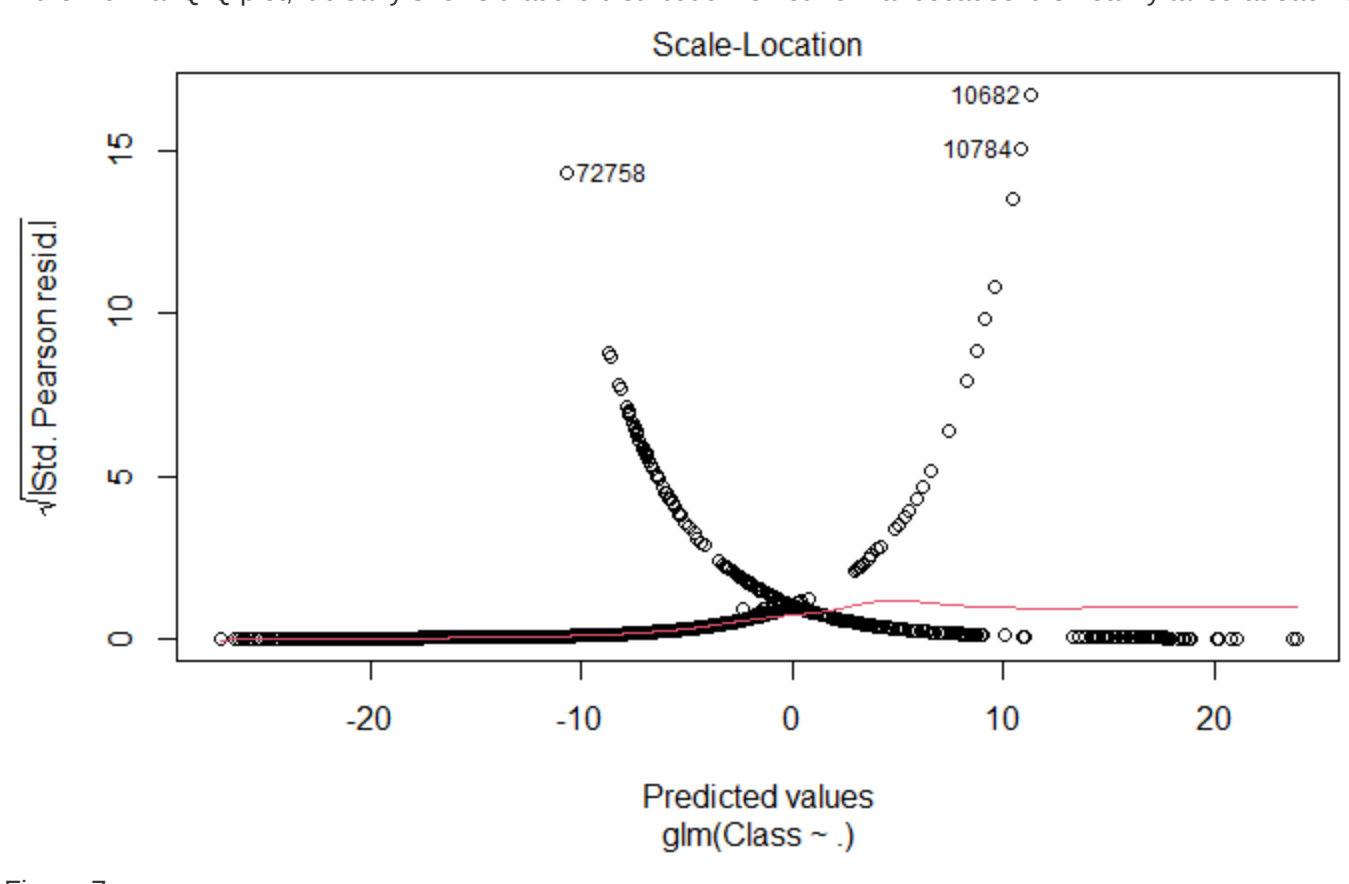

Figure 7

In the scale-location plot, the red line fits the dots pretty well and it is roughly horizontal, which indicates the spread of residuals is roughly equal at all fitted values.

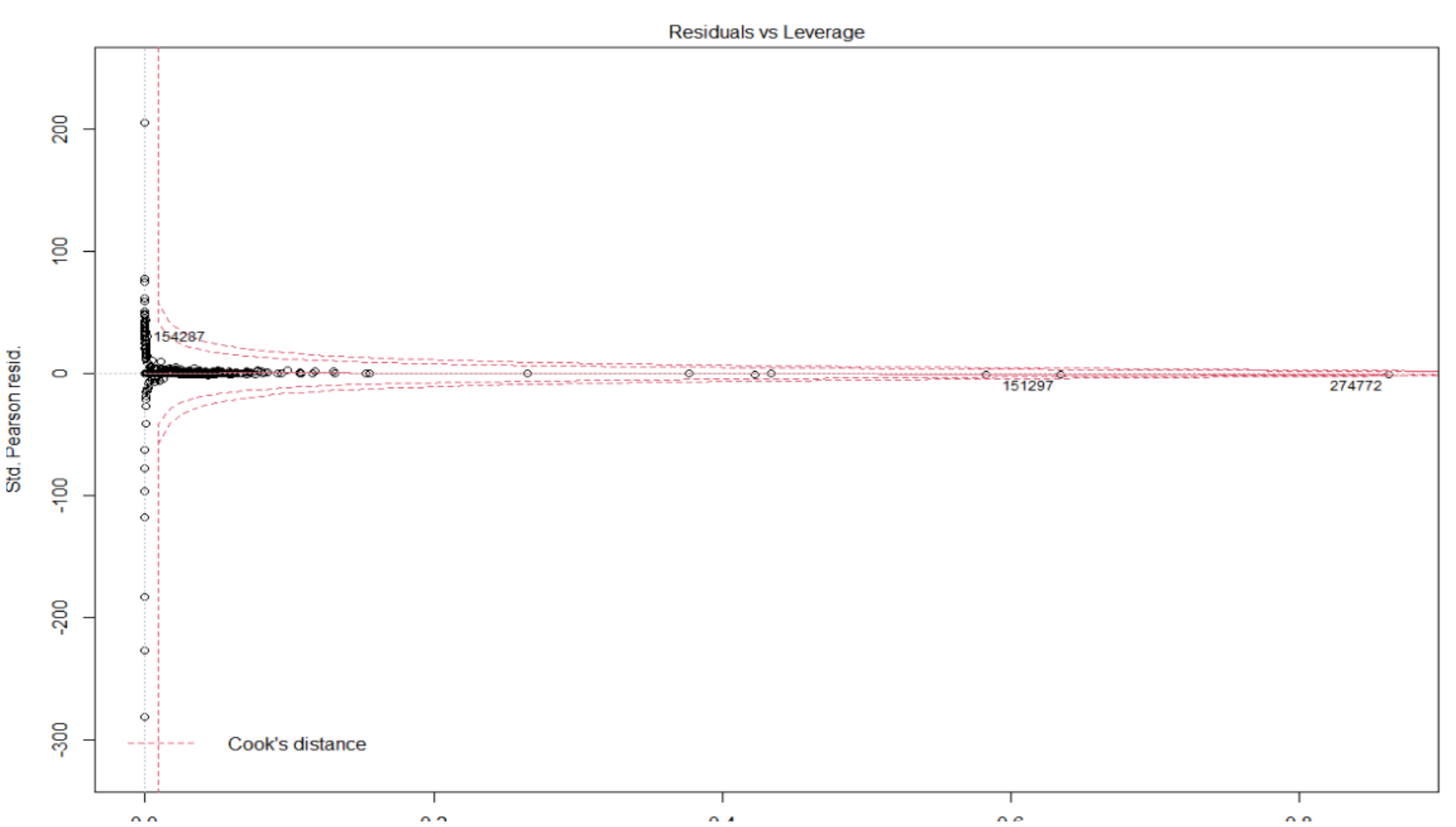

Figure 8

By looking at the residuals vs leverage plot, it is unclear to see any information because of outliers.

After creating the logistic regression model, I used it to predict y values by using testing set and named the predicted values as "predict".Here is the table of predicted value and test value:

```
  predict
        0     1
  0 85282    13
  1    69    79
```

Figure 9

I then calculated the RMSE value, which stands for root-mean-square-error. Here is its formula:

Formula

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \hat{x}_i\right)^2}{N}}$$

Figure 10

The value I got is 5.64486934348491. The model accuracy is 0.999040295869761 and the error rate of the model is 0.00107674121929239.I also created an ROC curve for the logistic regression model.ROC curve indicates the performance of a classfication model.Specificity stands for False positive rate, and sensitivity stands for true positive rate.

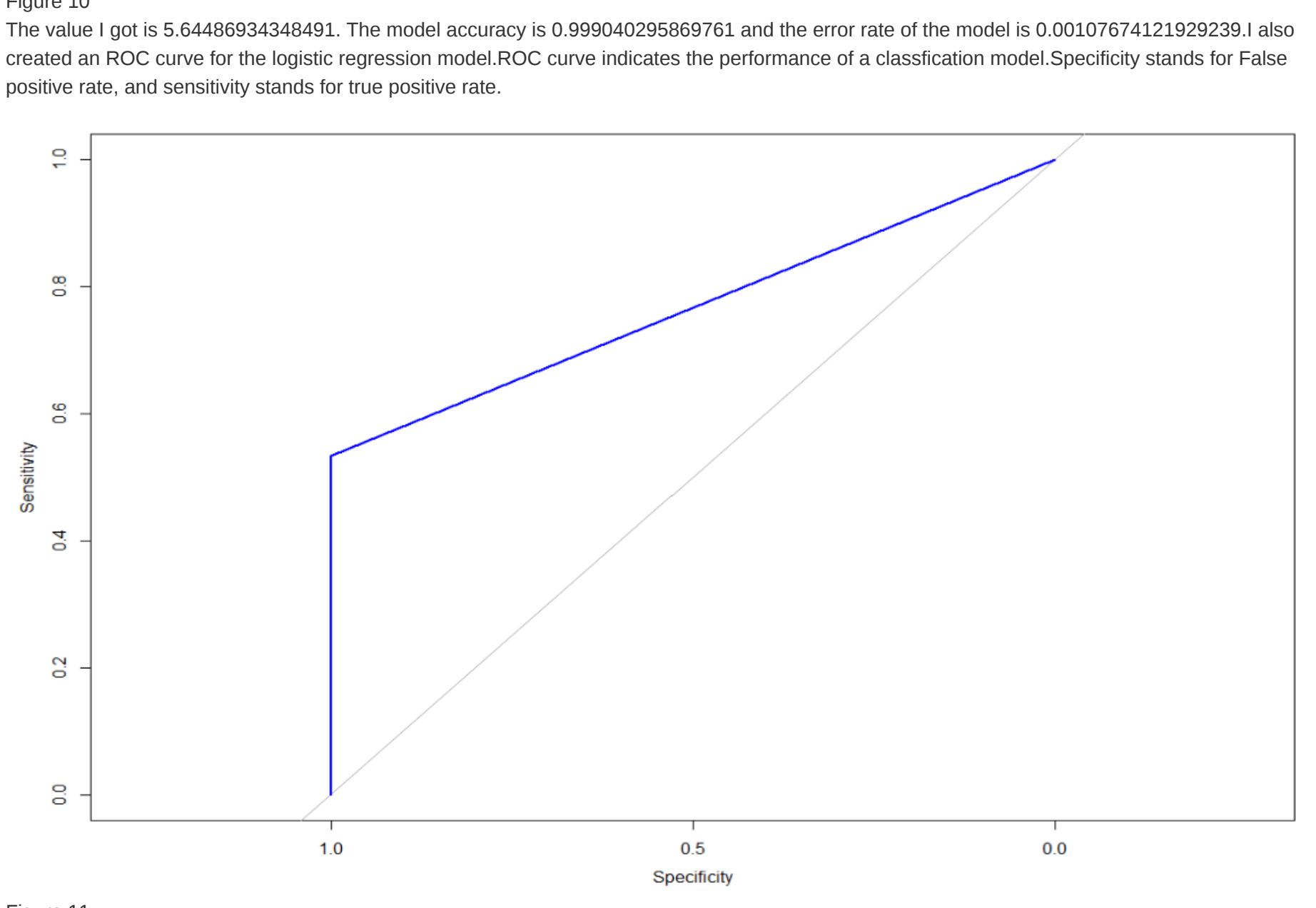

Figure 11

```
Call:
roc.default(response = test$Class, predictor = predict, plot = T,     col =
"blue")

Data: predict in 85295 controls (test$Class 0) < 148 cases (test$Class 1).
Area under the curve: 0.7668
```

Figure 12

The area under the curve, which is AUC, is 0.7668 and it shows an acceptable result.

I also used Hosmer-Lemeshow test to do a goodness of fit test for the logistic model I produced and then I ggplot the test.Here are the GOF table and ggplot:

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  logistic_regression_model$y, fitted(logistic_regression_model)
X-squared = 12.866, df = 8, p-value = 0.1166
```

Figure 13

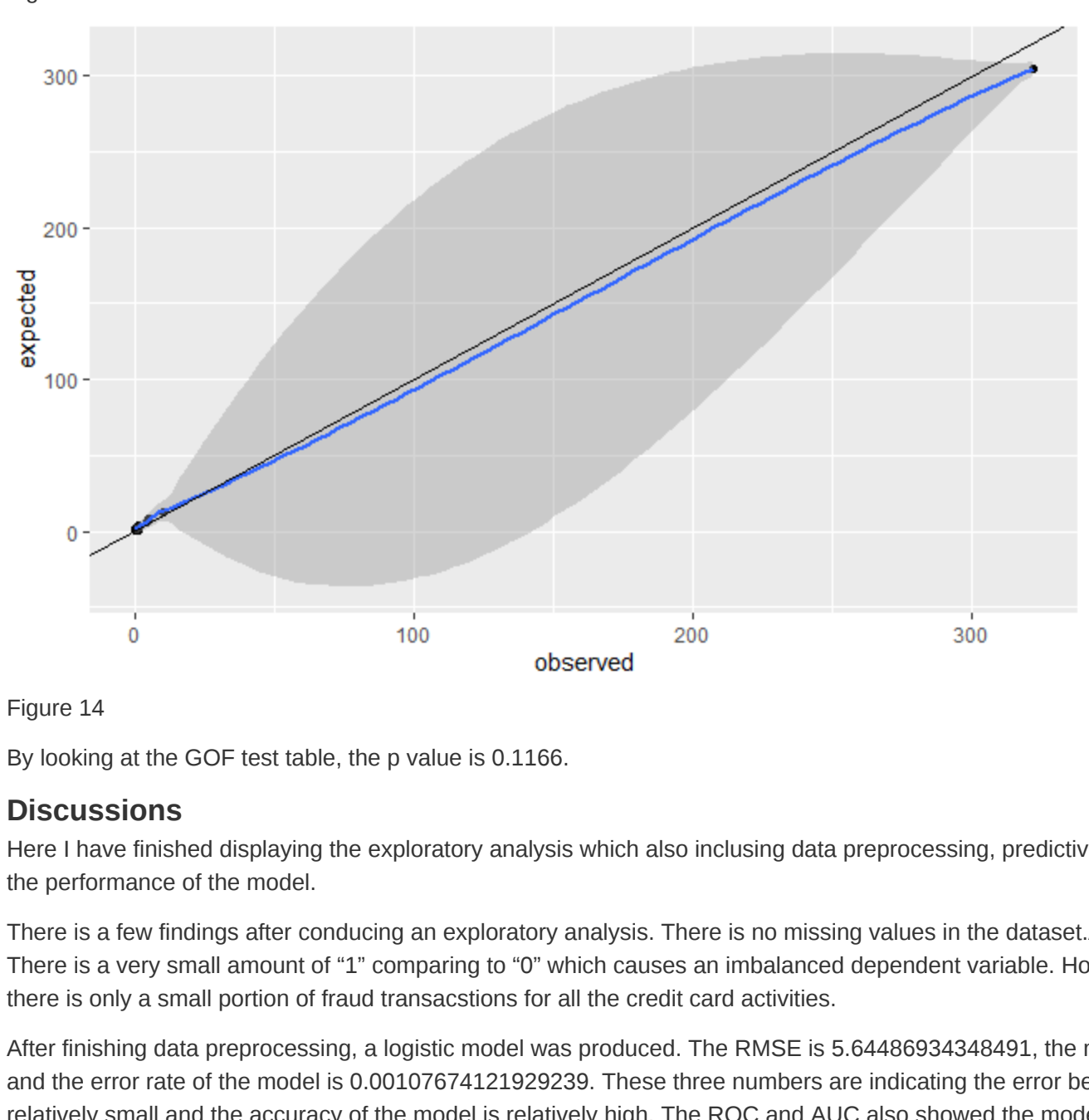

Figure 14

By looking at the GOF test table, the p value is 0.1166.

Discussions

Here I have finished displaying the exploratory analysis which also inclusing data preprocessing, predictive analysis and goodess-of-fit that shows the performance of the model.

There is a few findings after conducing an exploratory analysis. There is no missing values in the dataset.All the variables are numeric variables. There is a very small amount of "1" comparing to "0" which causes an imbalanced dependent variable. Howeverf, it also indicates a good sign that there is only a small portion of fraud transacstions for all the credit card activities.

After finishing data preprocessing, a logistic model was produced. The RMSE is 5.64486934348491, the model accuracy is 0.999040295869761 and the error rate of the model is 0.00107674121929239. These three numbers are indicating the error between predicted and tested values is relatively small and the accuracy of the model is relatively high. The ROC and AUC also showed the model is relatively good. I also used Hosmer-Lemeshow test to do a goodess-of-fit test for the logistic regression model I created. The p value is 0.1166, which is larger than the significant level 0.05. It indicates that this logistic regression model is a good model.

The weakness of the analysis is the lack of comparisons between models. Due to the lack of knowledge I have learned about big data, I didn't know what other models are good for a binary dataset. I have tried usng linear regression, postStratification; however, they both didn't work well and console kept getting errors. so I removed them in the coding process.

In the future, I would like to apply decision tree, random forest and other popular classification models for this project. I have browsed online and these algorithms are good for classification datasets.By comparing different models, I could choose the best one to be my final result.

References

Credit Card Fraud. (2016, June 15). Retrieved December 16, 2020, from https://www.fbi.gov/scams-and-safety/common-scams-and-crimes/credit-card-fraud

James, R. (2020, January 15). How Is Big Data Used To Fight Against Credit Card Fraud? Retrieved December 16, 2020, from https://becominghuman.ai/how-is-big-data-used-to-fight-against-credit-card-fraud-568fd6d63387

Stephanie. (2016, August 28). Hosmer-Lemeshow Test: Definition. Retrieved December 16, 2020, from https://www.statisticshowto.com/hosmer-lemeshow-test/

Swaminathan, S. (2019, January 18). Logistic Regression - Detailed Overview. Retrieved December 16, 2020, from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

ULB, M. (2018, March 23). Credit Card Fraud Detection. Retrieved December 16, 2020, from https://www.kaggle.com/mlg-ulb/creditcardfraud

What is Logistic Regression? (2020, March 09). Retrieved December 16, 2020, from https://www.statisticssolutions.com/what-is-logistic-regression/