HarvardX: PH125.9x Data Science: Capstone

Capstone Project Report: Credit Card Fraud Detection

Yan Song Bai

January 9, 2020

Executive Summary

The purpose of this Capstone Project is to create a credit card fraud detection system using data science and machine learning techniques to analyze transaction data of credit cards in September 2013 by European cardholders, with the dataset available at https://www.kaggle.com/mlg-ulb/creditcardfraud. The full dataset includes 284,407 transactions.

Key steps performed in this project are:

1. Download and import the dataset
2. Explore the dataset and create the train and Test sets
3. Process the data and develop data model
4. Review the model based on the Test set

The metric used for measuring the score is the Area Under Curve (AUC) and a desirable result should have an AUC at least greater than 0.85. In this analysis, the model is able to achieve an AUC of 0.9799858, indicating the of the analysis.

Methods and Analysis

Exploratory Data Analysis

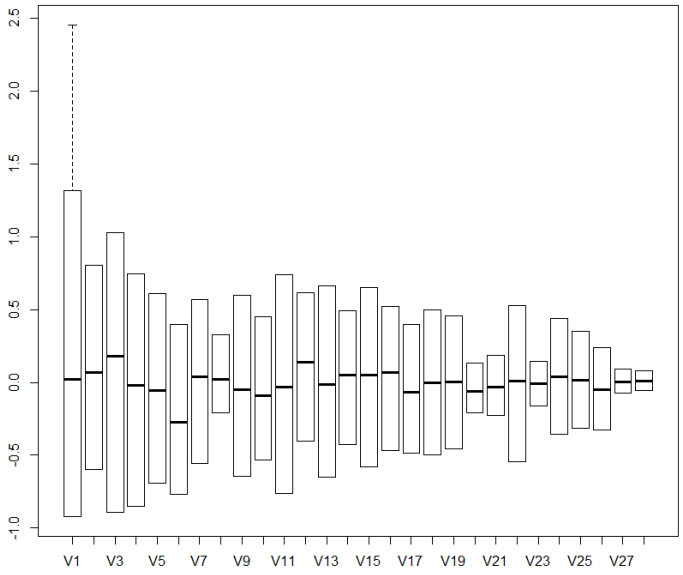
The dataset used in this analysis includes the credit card transactions during a two-day period in September 2013 by European cardholders. The dataset contains 284,407 transactions, with 30 features associated with the transaction.

Only numerical values are contained in this dataset due to PCA transformation, while the only 2 features that have not been transformed are 'Time' (i.e. the duration between the first transaction and the recorded transaction) and 'Amount' of the transactions and the rest of the features are labeled from V1 to V28 as they have low relevance to this analysis. The dataset is labeled with 'Class' and when Class has a value of 1, a positive (fraudulent) transaction is recorded, whereas a 0 value indicates regular transaction.

The first 6 rows of the dataset is as follows:

	Time	V1	V2	V3	V4	V5	V6
1	0	−1.35981	−0.07278	2.536347	1.378155	−0.33832	0.462388
2	0	1.191857	0.266151	0.16648	0.448154	0.060018	−0.08236
3	1	−1.35835	−1.34016	1.773209	0.37978	−0.5032	1.800499
4	1	−0.96627	−0.18523	1.792993	−0.86329	−0.01031	1.247203
5	2	−1.15823	0.877737	1.548718	0.403034	−0.40719	0.095921
6	2	−0.42597	0.960523	1.141109	−0.16825	0.420987	−0.02973
	V7	V8	V9	V10	V11	V12	
1	0.239599	0.098698	0.363787	0.090794	−0.5516	−0.6178	
2	−0.0788	0.085102	−0.25543	−0.16697	1.612727	1.065235	
3	0.791461	0.247676	−1.51465	0.207643	0.624502	0.066084	
4	0.237609	0.377436	−1.38702	−0.05495	−0.22649	0.178228	
5	0.592941	−0.27053	0.817739	0.753074	−0.82284	0.538196	
6	0.476201	0.260314	−0.56867	−0.37141	1.341262	0.359894	
	V13	V14	V15	V16	V17	V18	
1	−0.99139	−0.31117	1.468177	−0.4704	0.207971	0.025791	
2	0.489095	−0.14377	0.635558	0.463917	−0.1148	−0.18336	
3	0.717293	−0.16595	2.345865	−2.89008	1.109969	−0.12136	
4	0.507757	−0.28792	−0.63142	−1.05965	−0.68409	1.965775	
5	1.345852	−1.11967	0.175121	−0.45145	−0.23703	−0.03819	
6	−0.35809	−0.13713	0.517617	0.401726	−0.05813	0.068653	
	V19	V20	V21	V22	V23	V24	
1	0.403993	0.251412	−0.01831	0.277838	−0.11047	0.066928	
2	−0.14578	−0.06908	−0.22578	−0.63867	0.101288	−0.33985	
3	−2.26186	0.52498	0.247998	0.771679	0.909412	−0.68928	
4	−1.23262	−0.20804	−0.1083	0.005274	−0.19032	−1.17558	
5	0.803487	0.408542	−0.00943	0.798278	−0.13746	0.141267	
6	−0.03319	0.084968	−0.20825	−0.55982	−0.0264	−0.37143	
	V25	V26	V27	V28	Amount	Class	
1	0.128539	−0.18911	0.133558	−0.02105	149.62	0	
2	0.16717	0.125895	−0.00898	0.014724	2.69	0	
3	−0.32764	−0.1391	−0.05535	−0.05975	378.66	0	
4	0.647376	−0.22193	0.062723	0.061458	123.5	0	
5	−0.20601	0.502292	0.219422	0.215153	69.99	0	
6	−0.23279	0.105915	0.253844	0.08108	3.67	0	

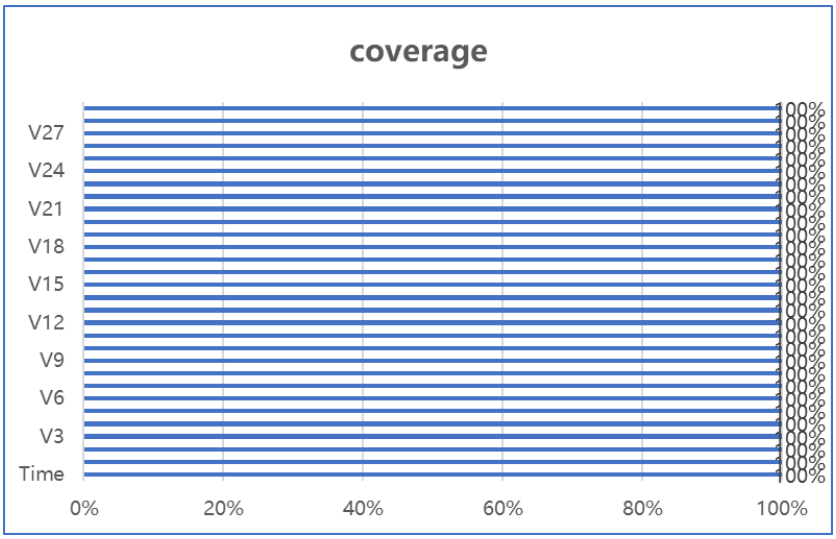
Data outlier is also identified in this analysis. It can be seen in the following graphy that there are no significant outliers in the dataset.



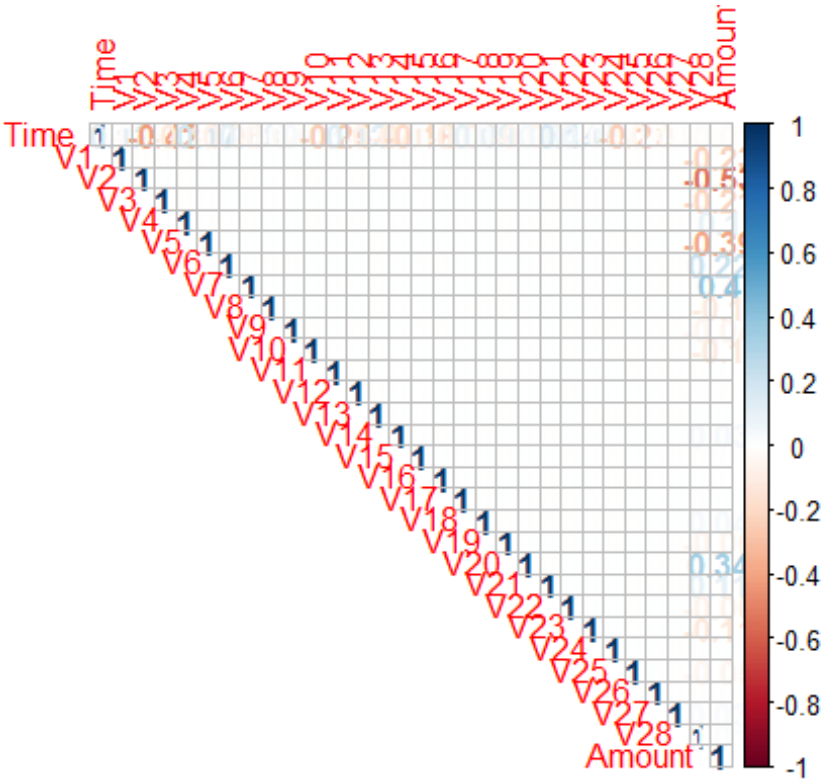
The dataset is arbitrarily separated into training and testing sets. Training set contains 70% of the data while testing set contains 30%. The distribution of the transaction data is shown as below. It can be seen that the dataset is imbalanced.

Label	Total sample	Train	Test
0 (Normal)	284315	199145	85170
1 (Fraud)	492	342	150

As missing values may cause high degree of uncertainties and instabilities in data analysis and modelling, it is important to determine the degree of data coverage and data quality of the dataset. As the chart below suggests, the dataset has no missing values and has full coverage.



Correlation between each variables in the dataset are examined, as shown in the following graph. It can be seen that the variables are not closely related to each other and multicollinearity is not a major concern in the following analysis.



Data Processing

Because of the continuity of data, the ChiMerge method is used in order to make discretized data easier to process and stabilize modelling. The dplyr package is mainly used in this part of the analysis. The whole dataset is divided into 100 intervals and Chi-square values are calculated to merge the two adjacent intervals with the lowest Chi-square values until all pairs have Chi-square values above the threshold value. Each interval must contain positive (fraud) and negative (normal) data. The ChiMerge results are shown as follows.

```
##     df_name        Var1 Var2   Freq
## 1      Time     <=42500    0   7799
## 2      Time 42500~83200    0  37858
## 3      Time 83200~10900    0  93623
## 4      Time      >10900    0 145035
## 5      Time     <=42500    1     25
## 6      Time 42500~83200    1    118
## 7      Time 83200~10900    1    121
## 8      Time      >10900    1    228
## 9        V1        <=-4    0   7841
## 10       V1       -4~-1    0  57237
## 11       V1        -1~1    0 102647
```

```
## 12       V1          >1    0 116590
## 13       V1       <=-4     1    181
## 14       V1      -4~-1     1    159
## 15       V1       -1~1     1     99
## 16       V1         >1     1     53
## 17       V2       <=-2     0  13850
## 18       V2       -2~1     0 217937
## 19       V2        1~2     0  41508
## 20       V2         >2     0  11020
## 21       V2       <=-2     1     18
## 22       V2       -2~1     1     83
## 23       V2        1~2     1     94
## 24       V2         >2     1    297
## 25       V3       <=-4     0   2503
## 26       V3      -4~-2     0  19596
## 27       V3       -2~0     0 105604
## 28       V3         >0     0 156612
## 29       V3       <=-4     1    299
## 30       V3      -4~-2     1     82
## 31       V3       -2~0     1     79
## 32       V3         >0     1     32
## 33       V4        <=1     0 228598
## 34       V4        1~2     0  36924
## 35       V4        2~4     0  15078
## 36       V4         >4     0   3715
## 37       V4        <=1     1     46
## 38       V4        1~2     1     50
## 39       V4        2~4     1    139
## 40       V4         >4     1    257
## 41       V5       <=-3     0   2803
## 42       V5     -3~-1.6    0  14105
## 43       V5    -1.6~0.6    0 195274
## 44       V5        >0.6    0  72133
## 45       V5       <=-3     1    170
## 46       V5     -3~-1.6    1     71
## 47       V5    -1.6~0.6    1    153
## 48       V5        >0.6    1     98
## 49       V6       <=-2     0   3344
## 50       V6      -2~-1     0  40619
## 51       V6       -1~0     0 132256
## 52       V6         >0     0 108096
## 53       V6       <=-2     1    177
## 54       V6      -2~-1     1    126
## 55       V6       -1~0     1    111
## 56       V6         >0     1     78
## 57       V7       <=-3     0   2616
## 58       V7     -3~-1.5    0   9735
## 59       V7     -1.5~-1    0  20343
## 60       V7         >-1    0 251621
## 61       V7       <=-3     1    249
```

```
## 62       V7     -3~-1.5    1     96
## 63       V7   -1.5~-1     1     21
## 64       V7        >-1     1    126
## 65       V8         <=0    0 135372
## 66       V8         0~1    0 133068
## 67       V8         1~2    0  12943
## 68       V8          >2    0   2932
## 69       V8         <=0    1    149
## 70       V8         0~1    1    149
## 71       V8         1~2    1     84
## 72       V8          >2    1    110
## 73       V9       <=-2     0   8723
## 74       V9      -2~-1     0  34610
## 75       V9   -1~-0.5     0  41946
## 76       V9       >0.5    0 199036
## 77       V9       <=-2    1    257
## 78       V9      -2~-1     1     99
## 79       V9   -1~-0.5     1     38
## 80       V9       >0.5    1     98
## 81      V10       <=-2    0   3125
## 82      V10      -2~-1     0  27113
## 83      V10       -1~0    0 129229
## 84      V10         >0    0 124848
## 85      V10       <=-2    1    395
## 86      V10      -2~-1     1     24
## 87      V10       -1~0    1     36
## 88      V10         >0    1     37
## 89      V11        <=0    0 145600
## 90      V11      0~1.8    0 130437
## 91      V11   1.8~2.2     0   5137
## 92      V11       >2.2    0   3141
## 93      V11        <=0    1     32
## 94      V11      0~1.8    1     75
## 95      V11   1.8~2.2     1     34
## 96      V11       >2.2    1    351
## 97      V12       <=-3    0   2834
## 98      V12      -3~-2    0  10547
## 99      V12       -2~0    0 107241
## 100     V12         >0    0 163693
## 101     V12       <=-3    1    368
## 102     V12      -3~-2    1     38
## 103     V12       -2~0    1     61
## 104     V12         >0    1     25
## 105     V13       <=-1    0  44496
## 106     V13       -1~0    0  99378
## 107     V13        0~1    0  96554
## 108     V13         >1    0  43887
## 109     V13       <=-1    1    120
## 110     V13       -1~0    1    129
## 111     V13        0~1    1    163
```

```
## 112     V13          >1    1     80
## 113     V14        <=-3    0   1983
## 114     V14     -3~-0.4    0  71848
## 115     V14    -0.4~0.3    0 110600
## 116     V14        >0.3    0  99884
## 117     V14        <=-3    1    405
## 118     V14     -3~-0.4    1     58
## 119     V14    -0.4~0.3    1     13
## 120     V14        >0.3    1     16
## 121     V15        <=-1    0  38166
## 122     V15        -1~0    0  97805
## 123     V15         0~1    0 109653
## 124     V15          >1    0  38691
## 125     V15        <=-1    1     85
## 126     V15        -1~0    1    178
## 127     V15         0~1    1    164
## 128     V15          >1    1     65
## 129     V16        <=-2    0   5688
## 130     V16        -2~0    0 126452
## 131     V16         0~1    0 124237
## 132     V16          >1    0  27938
## 133     V16        <=-2    1    340
## 134     V16        -2~0    1     72
## 135     V16         0~1    1     46
## 136     V16          >1    1     34
## 137     V17        <=-1    0  12966
## 138     V17        -1~0    0 140274
## 139     V17         0~1    0 107178
## 140     V17          >1    0  23897
## 141     V17        <=-1    1    374
## 142     V17        -1~0    1     21
## 143     V17         0~1    1     40
## 144     V17          >1    1     57
## 145     V18        <=-2    0   3748
## 146     V18       -2~-1    0  23187
## 147     V18        -1~1    0 227825
## 148     V18          >1    0  29555
## 149     V18        <=-2    1    223
## 150     V18       -2~-1    1     68
## 151     V18        -1~1    1    150
## 152     V18          >1    1     51
## 153     V19         <=0    0 141653
## 154     V19         0~1    0 116900
## 155     V19         1~2    0  21771
## 156     V19          >2    0   3991
## 157     V19         <=0    1    151
## 158     V19         0~1    1    145
## 159     V19         1~2    1     97
## 160     V19          >2    1     99
## 161     V20      <=-0.4    0  27143
```

```
## 162     V20      -0.4~0    0 141810
## 163     V20       0~0.2    0  59491
## 164     V20         >0.2    0  55871
## 165     V20      <=-0.4    1     80
## 166     V20      -0.4~0    1     97
## 167     V20       0~0.2    1     42
## 168     V20         >0.2    1    273
## 169     V21         <=-1    0   5366
## 170     V21      -1~0.3    0 238037
## 171     V21     0.3~0.5    0  24902
## 172     V21         >0.5    0  16010
## 173     V21         <=-1    1     42
## 174     V21      -1~0.3    1    128
## 175     V21     0.3~0.5    1     48
## 176     V21         >0.5    1    274
## 177     V22         <=-1    0  17963
## 178     V22         -1~0    0 123201
## 179     V22          0~2    0 142577
## 180     V22           >2    0    574
## 181     V22         <=-1    1     45
## 182     V22         -1~0    1    193
## 183     V22          0~2    1    238
## 184     V22           >2    1     16
## 185     V23      <=-0.3    0  33345
## 186     V23      -0.3~0    0 114774
## 187     V23       0~0.4    0 116403
## 188     V23         >0.4    0  19793
## 189     V23      <=-0.3    1    143
## 190     V23      -0.3~0    1    138
## 191     V23       0~0.4    1    103
## 192     V23         >0.4    1    108
## 193     V24      <=-0.6    0  41843
## 194     V24   -0.6~-0.1    0  60305
## 195     V24    -0.1~0.5    0 120398
## 196     V24         >0.5    0  61769
## 197     V24      <=-0.6    1     84
## 198     V24   -0.6~-0.1    1    150
## 199     V24    -0.1~0.5    1    203
## 200     V24         >0.5    1     55
## 201     V25         <=-1    0   8745
## 202     V25         -1~0    0 130343
## 203     V25          0~1    0 139786
## 204     V25           >1    0   5441
## 205     V25         <=-1    1     41
## 206     V25         -1~0    1    173
## 207     V25          0~1    1    235
## 208     V25           >1    1     43
## 209     V26      <=-0.3    0  78367
## 210     V26      -0.3~0    0  73109
## 211     V26       0~0.3    0  69351
```

```
## 212     V26        >0.3    0  63488
## 213     V26      <=-0.3    1    101
## 214     V26      -0.3~0    1    143
## 215     V26       0~0.3    1     99
## 216     V26        >0.3    1    149
## 217     V27        <=-1    0   4089
## 218     V27        -1~0    0 136781
## 219     V27         0~1    0 141037
## 220     V27          >1    0   2408
## 221     V27        <=-1    1     61
## 222     V27        -1~0    1     73
## 223     V27         0~1    1    256
## 224     V27          >1    1    102
## 225     V28      <=-0.1    0  37748
## 226     V28    -0.1~0.2    0 222935
## 227     V28     0.2~0.4    0  18856
## 228     V28        >0.4    0   4776
## 229     V28      <=-0.1    1    126
## 230     V28    -0.1~0.2    1    147
## 231     V28     0.2~0.4    1    109
## 232     V28        >0.4    1    110
## 233  Amount          <=    0  30311
## 234  Amount        1~10    0  69704
## 235  Amount       10~96    0 124716
## 236  Amount         >96    0  59584
## 237  Amount          <=    1    181
## 238  Amount        1~10    1     68
## 239  Amount       10~96    1     82
## 240  Amount         >96    1    161
```

Information Value (IV) is utilized as a metric to examine the distinctiveness of variables and is used as filter towards the variables prior to modelling.

$$IV_i = (p(y=1)_i - P(y=0)_i) * WOE_i$$

Whereas Weight of Evidence (woe) represents the transformation towards the original variables in which continuous variables are discretized and each resulting interval has a corresponding value calculated; it is determined by the percentage of positive (fraud) data divided by the percentage of negative (normal) data within the interval.

$$WOE_i = ln\frac{p(y=1_i)}{p(y=0)_I}$$

In this analysis, the IV values of the data is determined as follows. Variables with IV higher than 1 are kept while the rest are filtered. 18 variables are kept as a result of this operation. Regarding the imbalance of the dataset, no particular action is performed since the presence of the 18 variables with IV higher than 1 indicates that the dataset is relatively distinctive.

```
##         feature         IV
## V14         V14 5.73056320
## V12         V12 4.77211529
## V10         V10 4.76780691
## V11         V11 4.25588245
## V3           V3 3.83718438
## V4           V4 3.79811396
## V17         V17 3.57725022
## V16         V16 3.24859276
## V7           V7 3.06602480
## V2           V2 2.47157350
## V9           V9 2.10875807
## V18         V18 2.06638119
## V21         V21 1.92303083
## V6           V6 1.62367146
## V5           V5 1.60599284
## V1           V1 1.43691543
## V28         V28 1.27175352
## V27         V27 1.26451088
## V8           V8 0.97191141
## V20         V20 0.79876647
## V19         V19 0.74391114
## Amount   Amount 0.70055422
## V23         V23 0.50852704
## V25         V25 0.18482111
## V24         V24 0.10751990
## Time       Time 0.10600724
## V22         V22 0.09996072
## V13         V13 0.06452457
## V26         V26 0.05729099
## V15         V15 0.01836866
```

Data Modeling

Based on the exploration and processing of the dataset, the logistic regression model is utilized in this analysis as a classifier of credit card transaction data. When the output is greater than 0.5, a positive (fraud) transaction is detected; when the output is smaller than 0.5, a negative (normal) transaction is detected.

$$h_\theta(x) = \frac{1}{1 + e^{\theta^T} x}$$

Area Under Curve (AUC), defined as the area under the ROC curve, is used to assess the performance of the model. A higher AUC indicates a better performance of the classifier model. In this analysis, the AUC is determined as follows.

```
##
## Call:
## glm(formula = f, family = binomial, data = traindata)
##
```

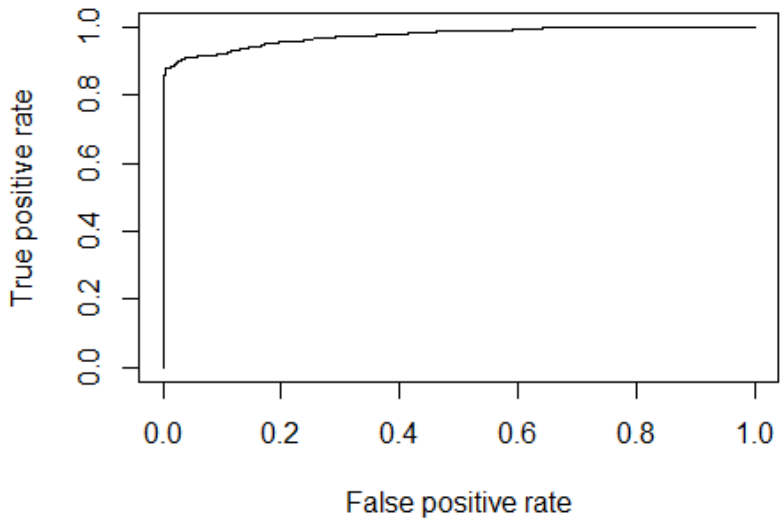
```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1603  -0.0189  -0.0109  -0.0068   4.6246
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.38664    1.21108  -0.319 0.749534
## V1-4~-1      0.06407    0.44163   0.145 0.884650
## V1-1~1       0.11433    0.52062   0.220 0.826179
## V1>1         0.52568    0.57687   0.911 0.362160
## V2-2~1       0.91327    0.62396   1.464 0.143283
## V21~2        0.80497    0.64257   1.253 0.210305
## V2>2        -1.01406    0.67391  -1.505 0.132392
## V3-4~-2     -0.43698    0.47304  -0.924 0.355611
## V3-2~0      -0.83634    0.50623  -1.652 0.098513 .
## V3>0        -1.55375    0.58054  -2.676 0.007442 **
## V41~2        1.12786    0.33787   3.338 0.000844 ***
## V42~4        3.33922    0.33447   9.984  < 2e-16 ***
## V4>4         3.69136    0.41557   8.883  < 2e-16 ***
## V5-3~-1.6    0.83800    0.58902   1.423 0.154824
## V5-1.6~0.6   0.79190    0.60162   1.316 0.188081
## V5>0.6       0.68476    0.62302   1.099 0.271719
## V6-2~-1      0.71260    0.47840   1.490 0.136341
## V6-1~0       1.62262    0.48916   3.317 0.000909 ***
## V6>0         0.84818    0.47425   1.788 0.073699 .
## V7-3~-1.5   -1.02956    0.63781  -1.614 0.106481
## V7-1.5~-1   -2.05572    0.76185  -2.698 0.006969 **
## V7>-1       -1.83899    0.61566  -2.987 0.002817 **
## V9-2~-1      0.22352    0.46363   0.482 0.629727
## V9-1~-0.5    0.07743    0.50290   0.154 0.877629
## V9>0.5       0.20994    0.47187   0.445 0.656382
## V10-2~-1    -2.46663    0.43159  -5.715 1.10e-08 ***
## V10-1~0     -2.41863    0.42951  -5.631 1.79e-08 ***
## V10>0       -3.02774    0.45220  -6.696 2.15e-11 ***
## V110~1.8     0.90788    0.29462   3.081 0.002060 **
## V111.8~2.2   2.17719    0.46109   4.722 2.34e-06 ***
## V11>2.2      1.48593    0.44870   3.312 0.000927 ***
## V12-3~-2    -0.51740    0.49267  -1.050 0.293626
## V12-2~0     -1.23831    0.43957  -2.817 0.004846 **
## V12>0       -2.53332    0.49389  -5.129 2.91e-07 ***
## V14-3~-0.4  -2.09689    0.36163  -5.798 6.69e-09 ***
## V14-0.4~0.3 -3.68073    0.51066  -7.208 5.68e-13 ***
## V14>0.3     -4.22879    0.49158  -8.602  < 2e-16 ***
## V16-2~0     -0.75334    0.58166  -1.295 0.195269
## V160~1      -1.21088    0.61047  -1.984 0.047309 *
## V16>1       -1.75022    0.63878  -2.740 0.006145 **
## V17-1~0     -0.58417    0.54068  -1.080 0.279949
## V170~1      -0.46602    0.49889  -0.934 0.350250
## V17>1       -0.82708    0.49416  -1.674 0.094187 .
## V18-2~-1    -0.30220    0.64285  -0.470 0.638293
```

```
## V18-1~1      0.35548    0.61424   0.579 0.562767
## V18>1        0.38681    0.69109   0.560 0.575677
## V21-1~0.3   -1.13245    0.46603  -2.430 0.015100 *
## V210.3~0.5   0.09824    0.56005   0.175 0.860756
## V21>0.5      0.22186    0.46805   0.474 0.635502
## V27-1~0      0.34209    0.58498   0.585 0.558689
## V270~1       0.70604    0.59081   1.195 0.232069
## V27>1       -2.26346    0.71075  -3.185 0.001450 **
## V28-0.1~0.2 -0.47859    0.34017  -1.407 0.159457
## V280.2~0.4  -1.10878    0.43438  -2.553 0.010693 *
## V28>0.4     -0.88759    0.43620  -2.035 0.041868 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5039.6  on 199486  degrees of freedom
## Residual deviance: 1160.1  on 199432  degrees of freedom
## AIC: 1270.1
##
## Number of Fisher Scoring iterations: 12
```

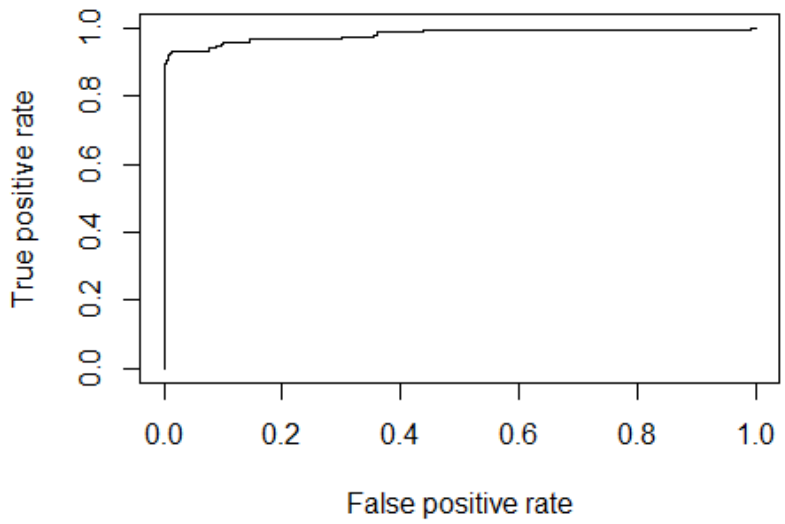
Results

For the purpose of this project, the final AUC should be greater than 0.85. The AUC results obtained from the logistic model are 0.9753151 for the training set and 0.9799858 for the test set, thus meeting the AUC requirement.

```
Train AUC: 0.9753151
```



```
Test AUC: 0.9799858
```



Conclusion

In this project, a credit card fraud detection system has been developed. The system is based on the variations among given variables. Finally a logistic regression model is developed based on data processing to classify data transaction data. An AUC of 0.9799858 has been achieved on the test set.

However, many other differentiating factors may also be considered to further tune the model and improve the accuracy. The dataset is unbalanced and the issue can be solved by under sampling with 1:1 ratio of positive and negative samples to train the model. Also, other algorithms and models can be used to further compare the performance of various models, such as GBM, KNN, Random Forest and lightGBM, etc.