Homework #8 - Report

Justin Robinette

March 19, 2019

No collaborators for any problem

Question 4: In the past couple of homework assignments you have used different classification methods to analyze the dataset you chose. For this homework, please write a summary report.

i) Introduction to the Dataset I used the Credit Approval Data Set (https://archive.ics.uci.edu/ml/datasets/Credit+Approval) from the UCI Machine Learning Repository. This data set contains information regarding credit card applications. All attribute names and the values were changed to protect confidentiality. This makes the data set more interesting because it helps to eliminate bias that one might have based on variable names and/or the values. For example, a variable titled 'CollegeStudent' that contains 'T' and 'F' values may bias the data if one felt that college students were more likely to be declined for credit cards.

As you can see below, there is also a good mix of attribute types and some missing values that were imputed.

```
##  V1            V2            V3          V4       V5           V6
##  ?: 12   ?      : 12   Min.   : 0.000   ?:  6   ? :  6   c      :137
##  a:210   22.67  :  9   1st Qu.: 1.000   l:  2   g :519   q      : 78
##  b:468   20.42  :  7   Median : 2.750   u:519   gg:  2   w      : 64
##          18.83  :  6   Mean   : 4.759   y:163   p :163   i      : 59
##          19.17  :  6   3rd Qu.: 7.207                    aa     : 54
##          20.67  :  6   Max.   :28.000                    ff     : 53
##          (Other):644                                     (Other):245
##        V7            V8          V9      V10          V11        V12
##  v      :399   Min.   : 0.000   f:329   f:395   Min.   : 0.0   f:374
##  h      :138   1st Qu.: 0.165   t:361   t:295   1st Qu.: 0.0   t:316
##  bb     : 59   Median : 1.000                   Median : 0.0
##  ff     : 57   Mean   : 2.223                   Mean   : 2.4
##  ?      :  9   3rd Qu.: 2.625                   3rd Qu.: 3.0
##  j      :  8   Max.   :28.500                   Max.   :67.0
##  (Other): 20
##  V13          V14           V15            V16
##  g:625   00000  :132   Min.   :     0.0   -:383
##  p:  8   00120  : 35   1st Qu.:     0.0   +:307
##  s: 57   00200  : 35   Median :     5.0
##          00160  : 34   Mean   :  1017.4
##          00080  : 30   3rd Qu.:   395.5
##          00100  : 30   Max.   :100000.0
##          (Other):394
```

ii) The Question to be Addressed The question that is addressed by this data set is whether the credit card application received a positive (+) or negative (-) decision.

iii) Initial Data Cleansing The first step in cleansing was to set column names. Then, the '?' values were replaced with 'NA' values. Then the 2nd and 14th columns were changed from factor values to numeric values since they are numeric. Then the response variable was changed from '+' and '-' to 'P' and 'N'.

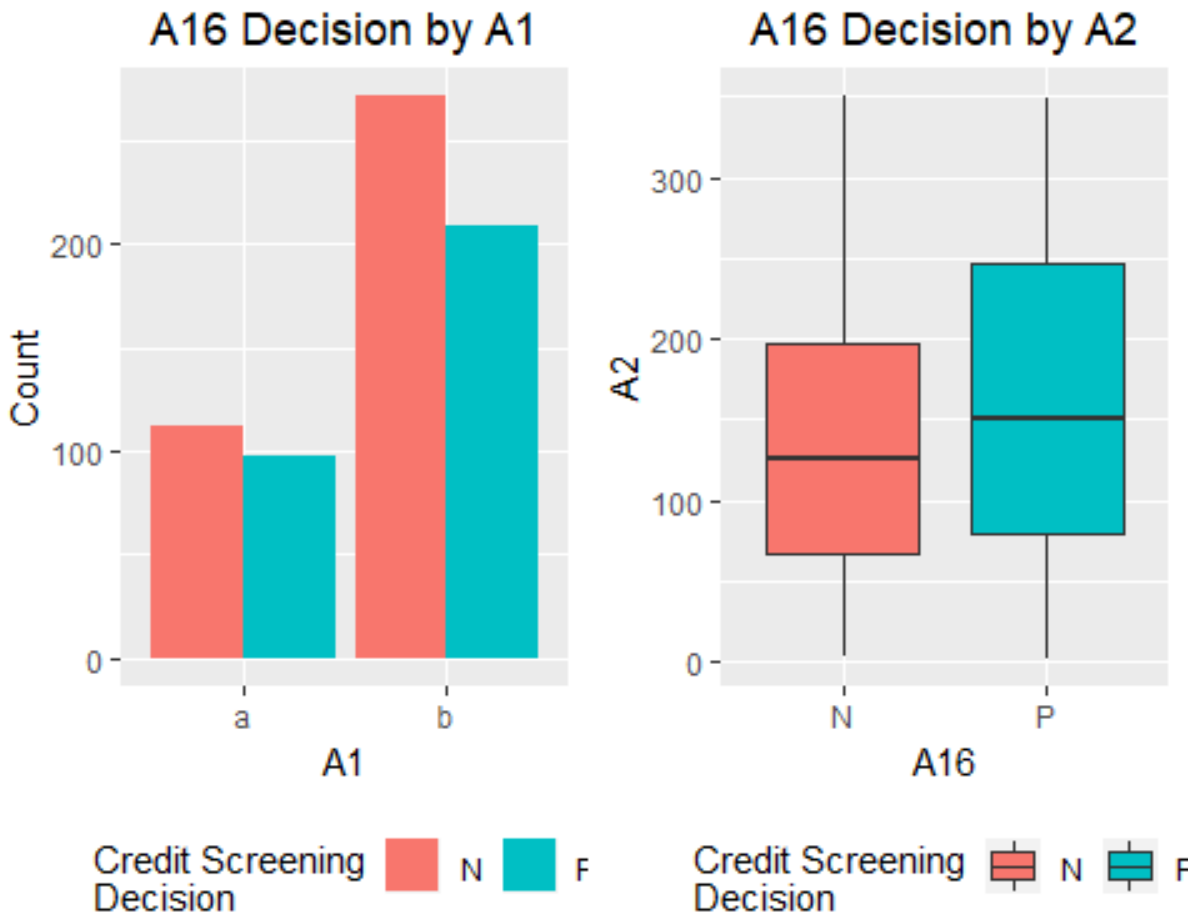
Missing factor variables were then imputed by using a function to fill 'NA' values with the most often occuring value. Missing numeric values were then replaced with the mean value for the respective variable. Lastly, so that the data set would work with Neural Network modeling, the 9th and 10th columns were changed from 't' and 'f' to '1' and '0' values.

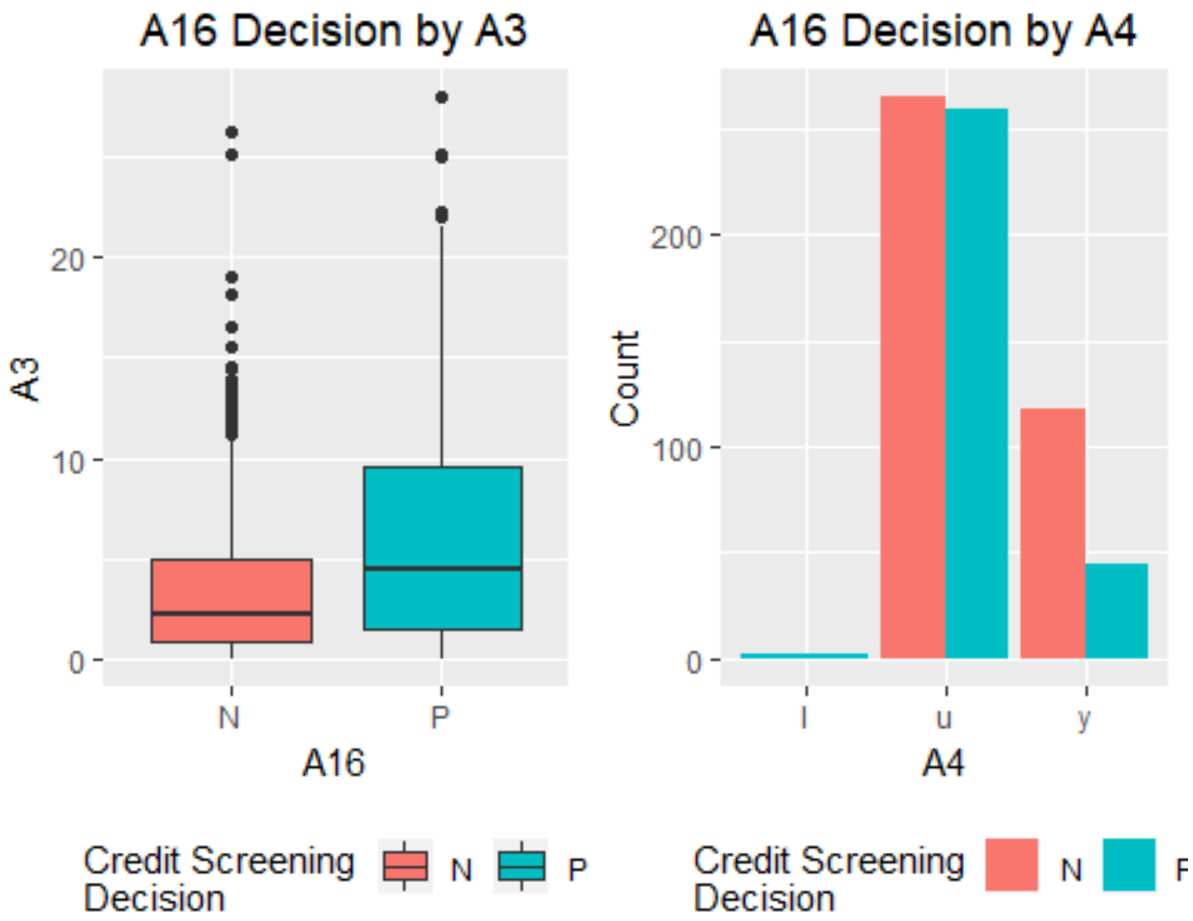
The summary below shows the results of these changes.

```
##  A1            A2              A3          A4       A5            A6
##  ?:  0   Min.   :  2.0   Min.   : 0.000   ?:  0   ? :  0   c      :146
##  a:210   1st Qu.: 73.0   1st Qu.: 1.000   l:  2   g :525   q      : 78
##  b:480   Median :135.5   Median : 2.750   u:525   gg:  2   w      : 64
##          Mean   :149.0   Mean   : 4.759   y:163   p :163   i      : 59
##          3rd Qu.:219.8   3rd Qu.: 7.207                    aa     : 54
##          Max.   :350.0   Max.   :28.000                    ff     : 53
##                                                            (Other):236
##        A7            A8               A9              A10
##  v      :408   Min.   : 0.000   Min.   :0.0000   Min.   :0.0000
##  h      :138   1st Qu.: 0.165   1st Qu.:0.0000   1st Qu.:0.0000
##  bb     : 59   Median : 1.000   Median :1.0000   Median :0.0000
##  ff     : 57   Mean   : 2.223   Mean   :0.5232   Mean   :0.4275
##  j      :  8   3rd Qu.: 2.625   3rd Qu.:1.0000   3rd Qu.:1.0000
##  z      :  8   Max.   :28.500   Max.   :1.0000   Max.   :1.0000
##  (Other): 12
##       A11       A12     A13          A14              A15
##  Min.   : 0.0   f:374   g:625   Min.   :  2.00   Min.   :     0.0
##  1st Qu.: 0.0   t:316   p:  8   1st Qu.: 25.00   1st Qu.:     0.0
##  Median : 0.0           s: 57   Median : 54.00   Median :     5.0
##  Mean   : 2.4                   Mean   : 59.27   Mean   :  1017.4
##  3rd Qu.: 3.0                   3rd Qu.: 95.00   3rd Qu.:   395.5
##  Max.   :67.0                   Max.   :171.00   Max.   :100000.0
##
##  A16
##  N:383
##  P:307
##
##
##
##
##
```

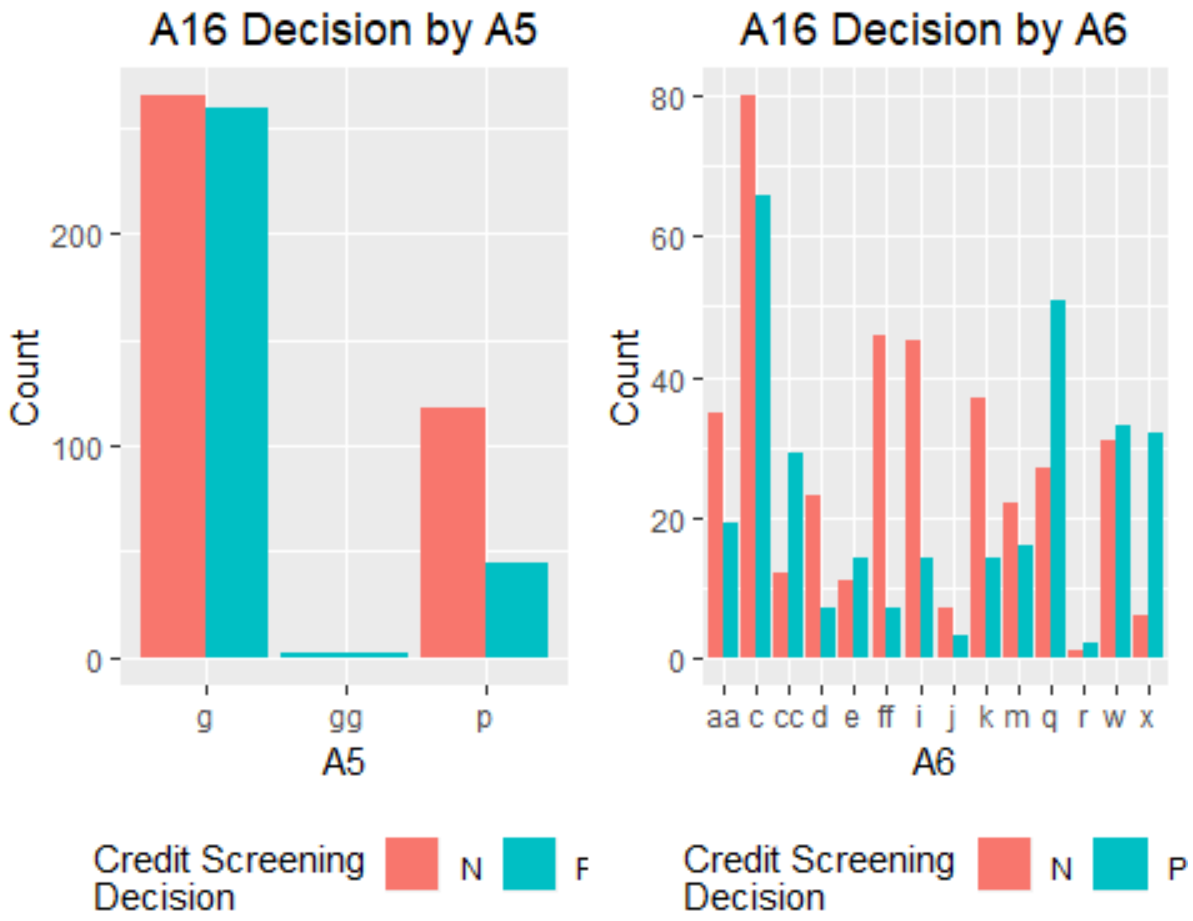
iv) Initial Descriptive Analysis Completed

First, plots were created to examine the relationship between each predictor and the response variable. The following is a summary of the plots shown below.
- A higher *A3* value indicates higher likihood toward a '+', or Positive in the credit screening
- Observations with an 'l' for *A4* receive all Positive screening results
- Observations with gg for *A5* receive all Positive screening results
- For *A6*, 'cc', 'q', 'r', 'w', and 'x' result in mostly Positive screening results
- For *A7*, 'h' and 'z' are the only values that result in mostly Positive screening results
- A higher *A8* results in more Positive screening results
- For both *A9* and *A10*, 't' values (True?) make a Positive screening result much better
- For *A11*, a higher value makes a Positive result better
- *A15* has some extreme outliers that may need to be removed for accurate modeling.

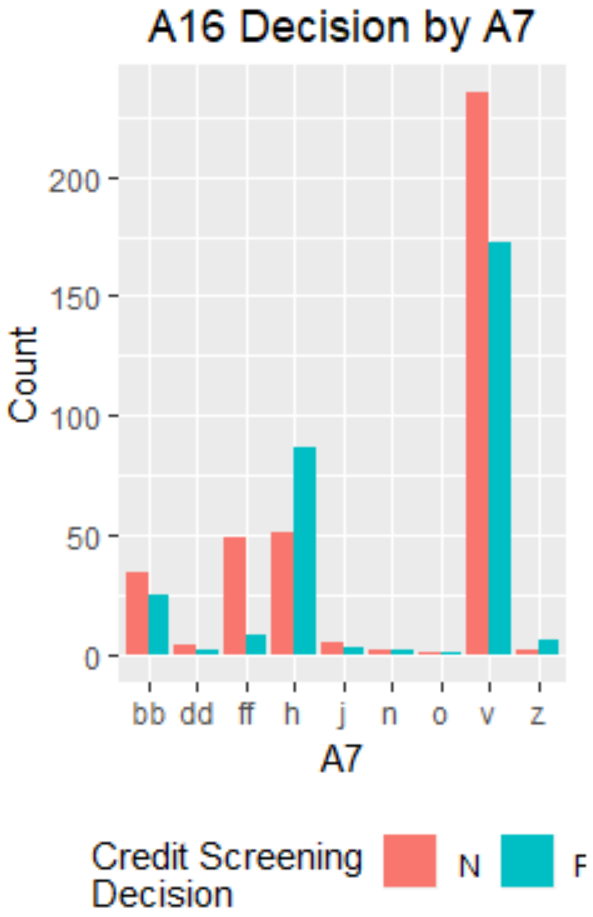


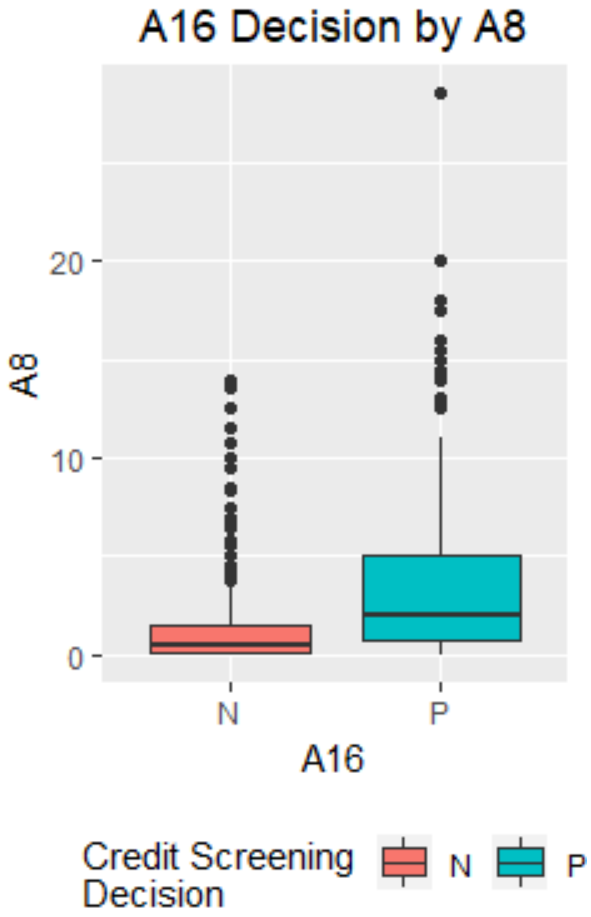

A16 Decision by A3
A3
20
10
0
N
P
A16
Credit Screening Decision
N
P

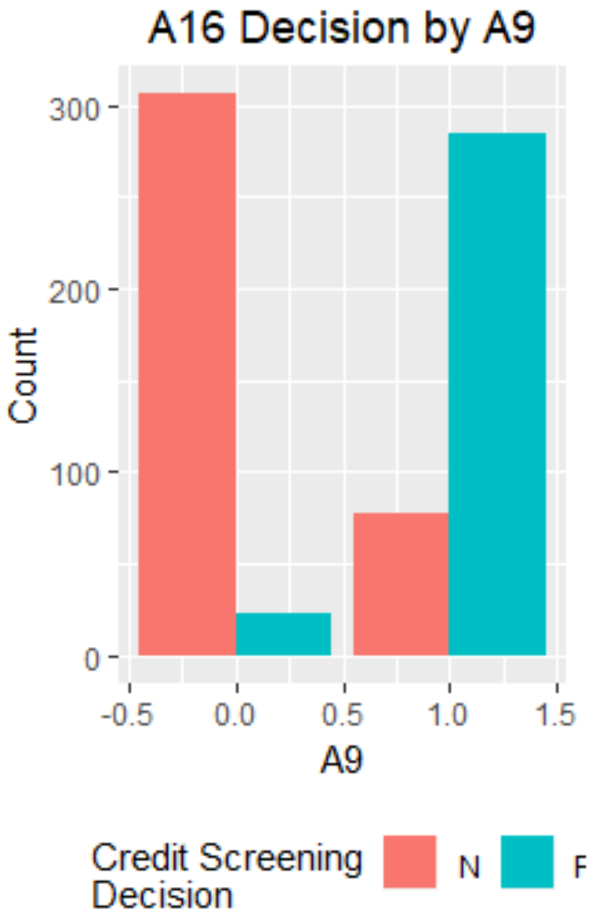
A16 Decision by A4
Count
200
100
0
l
u
y
A4
Credit Screening Decision
N
P

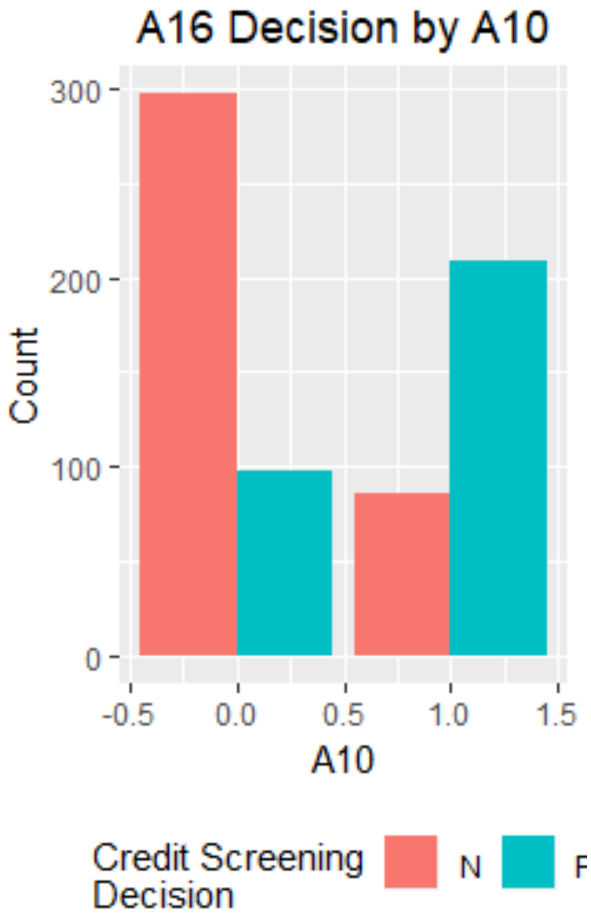

A16 Decision by A5
Count
200
100
0
g
gg
p
A5
Credit Screening Decision
N
P

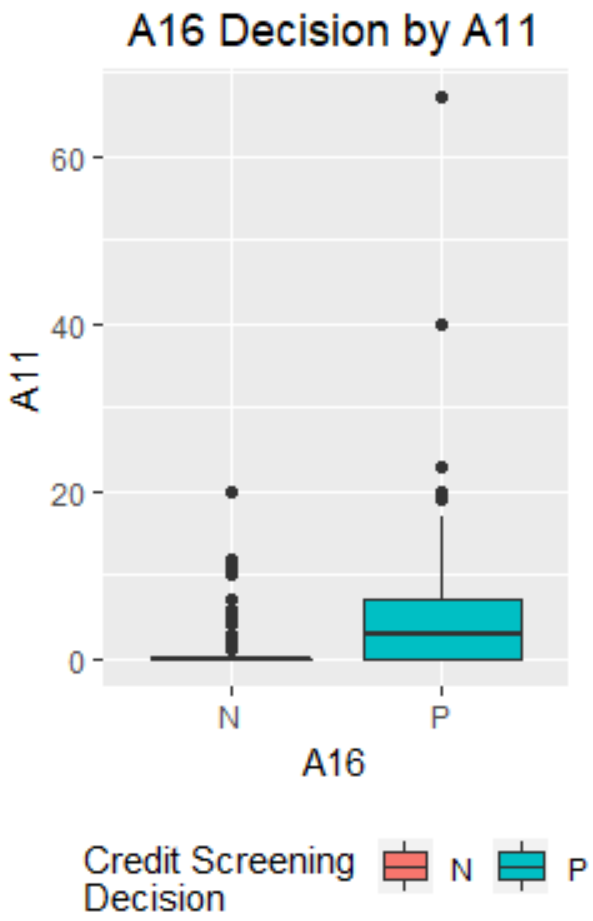
A16 Decision by A6
Count
80
60
40
20
0
aa c cc d e ff i j k m q r w x
A6
Credit Screening Decision
N
P

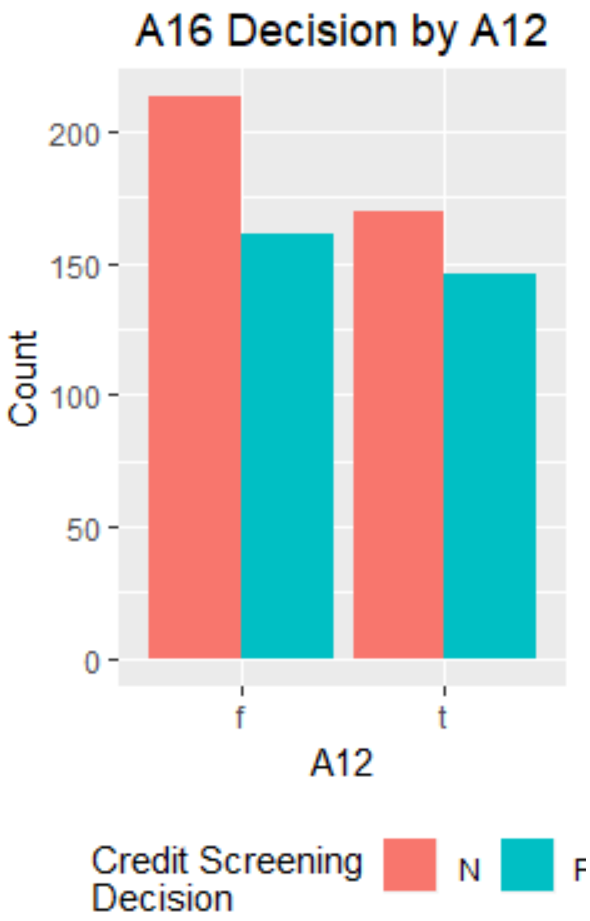

A16 Decision by A7
Count
200
150
100
50
0
bb dd ff h j n o v z
A7
Credit Screening Decision
N
F

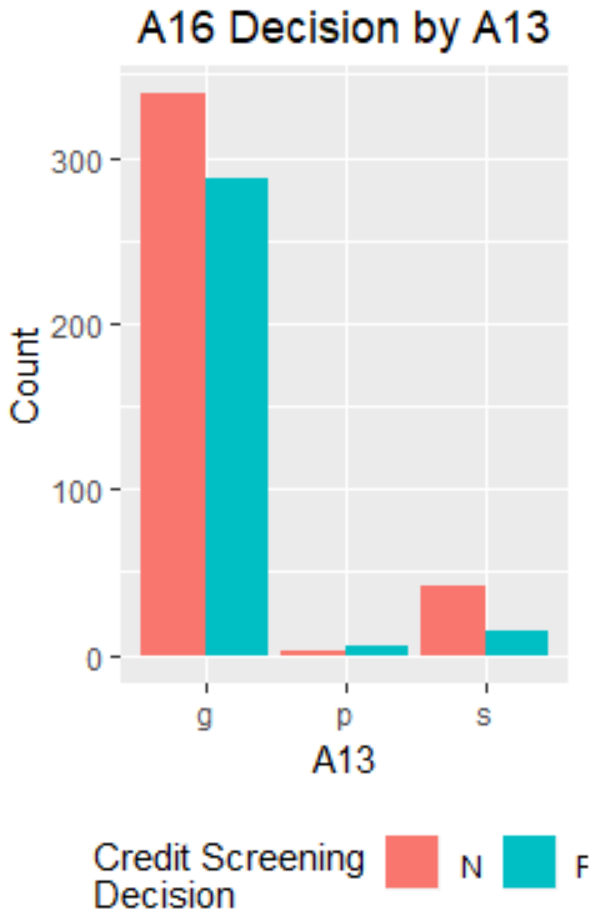

A16 Decision by A8
A8
20
10
0
N
P
A16
Credit Screening Decision
N
P

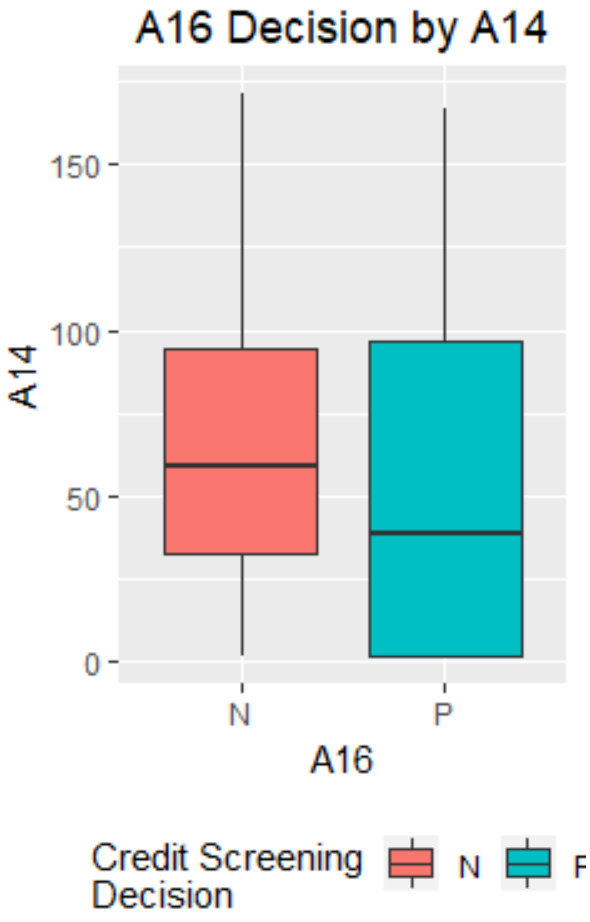

A16 Decision by A9
Count
300
200
100
0
-0.5
0.0
0.5
1.0
1.5
A9
Credit Screening Decision
N
F

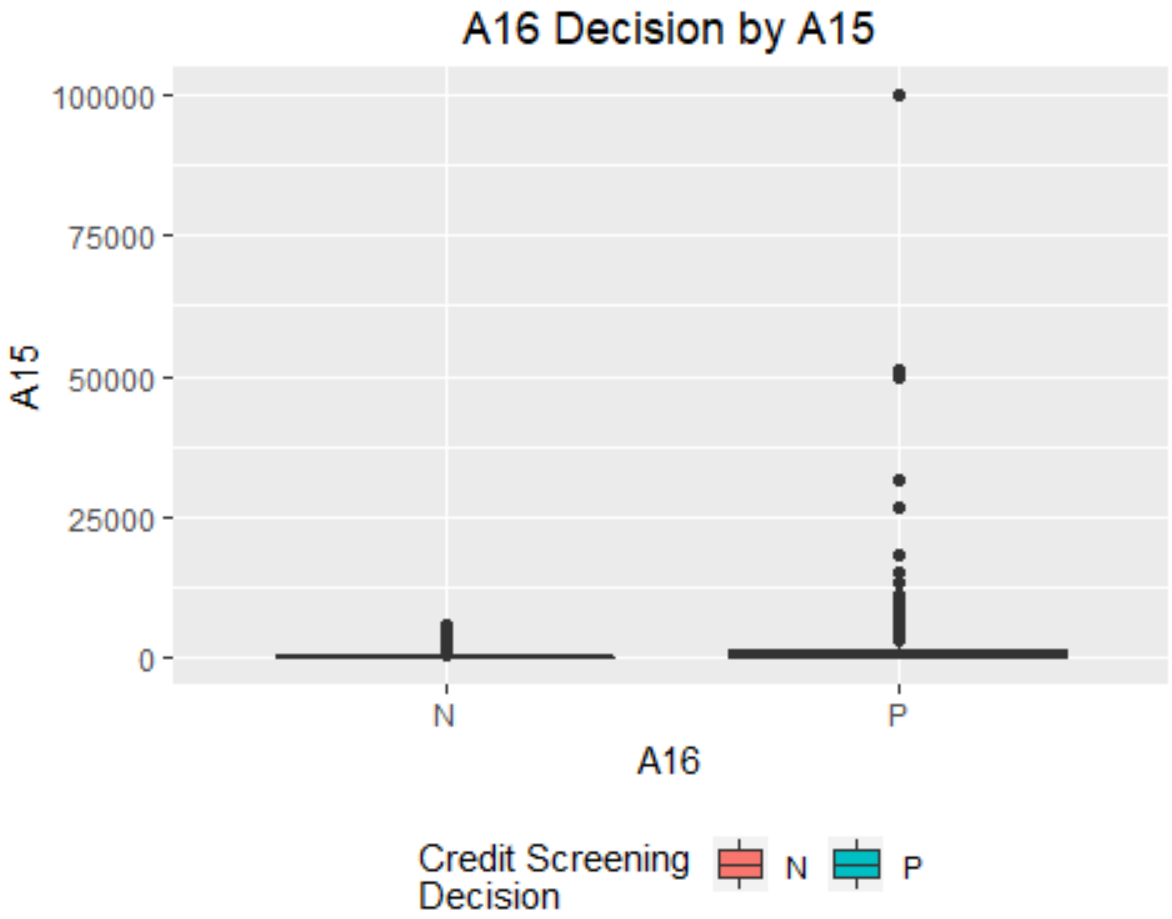

A16 Decision by A10
Count
300
200
100
0
-0.5
0.0
0.5
1.0
1.5
A10
Credit Screening Decision
N
F


A16 Decision by A11
A11
60
40
20
0
N
P
A16
Credit Screening Decision
N
P

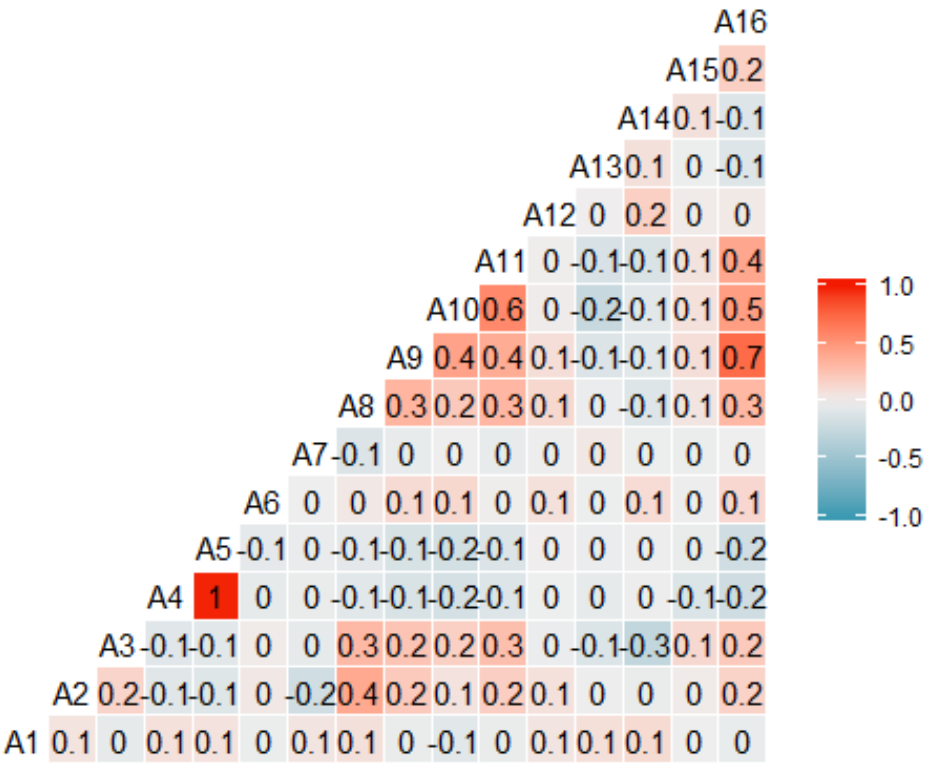

A16 Decision by A12
Count
200
150
100
50
0
f
t
A12
Credit Screening Decision
N
P


A16 Decision by A13
Count
300
200
100
0
g
p
s
A13
Credit Screening Decision
N
F


A16 Decision by A14
A14
150
100
50
0
N
P
A16
Credit Screening Decision
N
F



Next, a correlation plot was produced to further examine the correlation between the variables and the response variable (as well as correlations between predictors).



Lastly, to further examine the correlation, a correlation matrix was produced that shows the correlation between each predictor and the response variable (*A16*).

After these steps, it was determined that the variables with a correlation greater than or equal to the absolute value of 0.3 would be included in the models. The variables that meet this arbitrary threshold are: *A8*, *A9*, *A10*, and *A11*.

```
##        A1   A2   A3    A4    A5   A6 A7   A8   A9  A10  A11  A12  A13  A14
## A16 -0.03 0.16 0.21 -0.19 -0.19 0.13  0 0.31 0.72 0.46 0.41 0.03 -0.1 -0.1
##      A15 A16
## A16 0.18   1
```

v) Classification Methods Utilized

The following classification methods were utilized:
- Logistic Regression
- k-Nearest Neighbor (KNN) (k = 1, 5 and 10 were used and k = 5 had the best result)
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- MclustDA
- MclustDA with EDDA
- Neural Network

vi) Choosing the Model - Test Error / Cross Validation

For each of the methods listed above, Validation Set Approach, Leave One Out Cross-Validation (LOOCV), and 5-Fold Cross-Validation was used. The test errors of each method and each validation approach were then compiled into a table for optimal readability.

Next, LOOCV was used for the 7 different methods, as requested. Depending on the modeling method, a couple different methods were used to obtain the error rate from LOOCV. Next, the error rates were saved for reporting in the final table.

Lastly, 5-Fold Cross Validation was performed using a loop that iterates 5 times through each of the modeling methods to obtain the test error rate. The error rates were then compiled to add to the final summary table.

vii) Conclusion and Discussion

Below we see the final summary table, similar to the one that was requested. It displays the test error rate for each Method for VSA, LOOCV, and 5-Fold CV.

We can see that the overall worst error rate came from the LOOCV approach with the MclustDA method. We also notice that KNN performed relatively worse than the other methods using k=5 (which was shown to be the best k in the previous homework assignments.)

The best performing methods using VSA were Logistic Reg, LDA, and Neural Network all coming in with a test error rate of *12.0192%*. The best performing methods using the LOOCV approach were Logistic Regression and LDA - each achieving a test error rate of *14.4928% (rounded to 4 decimals)*. This test error rate is also the best test error rate present in the 5-Fold CV column of the table. Here it was achieved by Logistic Regression, LDA, and Neural Network.

On the surface, these error rates are relatively high. To relate these back to the question to be addressed from *Part ii* of this report, we are able to predict the outcome of the credit application process approximately 85-87% of the time given the predictors used.

Considering the somewhat ambiguous nature of the data, due to the masking for privacy purposes, this may not be quite as bad as it seems at first glance. Even the worst performing model between the three methods, MClust DA with LOOCV, was able to accurately predict 80% of the credit decisions based on the data.

Test Error by Validation Approach (%)

Method	VSA	LOOCV	5-Fold CV
Logistic Reg	12.0192	14.4928	14.4928
KNN	16.3462	18.6957	18.5507
LDA	12.0192	14.4928	14.4928
QDA	16.8269	17.3913	17.5362
MclustDA	16.8269	20	17.5362
MclustDA (EDDA)	16.8269	17.3913	17.5362
Neural Network	12.0192	14.6377	14.4928