Homework #5

Justin Robinette

February 19, 2019

No collaborators for any problem

Question 4.7.6, pg 170: Suppose we collect data for a group of students in a statistics class with variables **X1 = Hours Studied**, **X2 = Undergrad GPA**, and **Y = Receive an A**. We fit a logistic regression and produce estimated coefficient, $\beta_0 = -6$, $\beta_1 = 0.05$ and $\beta_2 = 1$.

Part A: Estimate the probability that a student who studies for 40h and has an undergrad GPA of 3.5 gets an A in this class.

Results: The probability can be calculated as:

$$p(x) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}$$

Plug in the beta values:

$$p(X) = \frac{e^{-6+0.05\times 40+1\times 3.5}}{1+e^{-6+0.05\times 40+1\times 3.5}}$$

Solve:

```
solution <- exp(-6+0.05*40+1*3.5)/(1+exp(-6+0.05*40+1*3.5))
paste("The probability that this student would receive an A in the class is:",r
ound(solution,5))

## [1] "The probability that this student would receive an A in the class is: 0
.37754"
```

Part B: How many hours would the student in *Part A* need to study to have a 50% chance of getting an A in the class?

Results: Set the equation equal to 0.5:

$$0.5 = \frac{e^{-6+0.05\times40+1\times3.5}}{1+e^{-6+0.05\times40+1\times3.5}}$$

Which becomes equal to:

$$log(\frac{0.5}{1-0.5}) = -6 + 0.05X_1 + 1 \times 3.5$$

Solve:

```
solution <- (log(0.5/(1-0.5)) + 6 - 3.5*1)/0.05
paste("In order to have a .5 probability of getting an A, the student would nee
d to study",solution,"hours.")
```

```
## [1] "In order to have a .5 probability of getting an A, the student would ne
ed to study 50 hours."
```

Question 4.7.7, pg 170: Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on *X*, which equals last year's percent profit. We examine a large number of companies and discover that the mean value of *X* for companies that issued a dividend was **X = 10**, while the mean for those that didn't was **X = 0**. In addition, the variance of *X* for these two sets of companies was $\sigma^2 = 36$. Finally, 80% of companies issued dividends. Assuming that *X* follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was **X = 4** last year.

Hint: Recall that the density function for a normal random variable is:

$$f(x) = \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-(x-\mu)^2/2\sigma^2}$$

. You will need to use the Bayes' theorem.

Results:

$$p(4) = \frac{0.8e^{-(1/72)(4-10)^2}}{0.8e^{-(1/72)(4-10)^2} + 0.2e^{-(1/72)(4-0)^2}}$$

Solve

```
solution <- (0.8*exp(-1/(2*36)*(4-10)^2))/(0.8*exp(-1/(2*36)*(4-10)^2)+(1-0.8)*
exp(-1/(2*36)*(4-0)^2))
paste("The probability that this company will issue a dividend this year is",ro
und(solution,5))
```

```
## [1] "The probability that this company will issue a dividend this year is 0.
75185"
```

Exercise #3: Continue Homeworks 3 & 4 using the **Weekly** dataset from 4.7.10. Fit a model (using the predictors chosen from previous homework) for classification using MclustDA function from the mclust-package.

i: Do a summary of your model.

- What is the best model selected by BIC? Report the Model Name and the BIC.(https://www.rdocumentation.org/packages/mclust/versions/5.4/topics/mclustModelNames)
- What is the training error? What is the test error?
- Report the True Positive Rate and the True Negative Rate.

Results: Here, I loaded the dataset and split into training and test, as we've done in the prior assignments. I also used the same predictor as I did in previous assignments which is **Lag2**. I then used MclustDA and reported the best model*(V - univariate, variable variance)* and it's BIC *(-4,327.8)*, according to the entire summary.

Then I used this model to predict the Weekly test data set **Direction**. I reported both the training *(0.4416244)* and test error *(0.4519231)* rates as well as the True Positive *(85.2459%)* and True Negative *(11.62791%)* rates on the test set. These are summarized in tables below and will be used for future comparisons.

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## MclustDA model summary:
##
##  log.likelihood   n df       BIC
##       -2129.439 985 10 -4327.804
##
## Classes   n Model G
##    Down 441     V 2
##    Up   544     V 2
##
## Training classification summary:
##
##       Predicted
## Class  Down  Up
##   Down   76 365
##   Up     70 474
##
## Training error = 0.4416244
```

Best Model Selected by BIC

	Best Model & Type	BIC
model	V	-4327.80411104579
type	univariate, unequal variance	-4327.80411104579

Training and Test Error of Best Model

Training Error	Test Error
0.4416244	0.4519231

True Positive/Negative Rates on Test Data

True Positive (%)	True Negative (%)
85.2459	11.62791

ii: Specify modelType = “EDDA” and run MclustDA again. Do a summary of your model.

- What is the best model by BIC?
- Find the training and test error rates.
- Report the True Positive and True Negative Rate.

Results: For this exercise, I also used the same predictor as I did in previous assignments which is **Lag2**. I then used EDDA as my *modelType* in the MclustDA function and reported the best model*(E - univariate, equal variance)* and it’s BIC *(-4,429.2)*, according to the entire summary.

Then I used this model to predict the Weekly test data set **Direction**. I reported both the training *(0.4456853)* and test error *(0.375)* rates as well as the True Positive *(91.80328%)* and True Negative *(20.93023%)* rates on the test set. These are summarized in tables below and will be used for future comparisons.

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## EDDA model summary:
##
##  log.likelihood   n df       BIC
##       -2204.237 985  3 -4429.152
##
## Classes   n Model G
##    Down 441     E 1
##    Up   544     E 1
##
## Training classification summary:
##
##       Predicted
## Class  Down  Up
##   Down   22 419
##   Up     20 524
##
## Training error = 0.4456853
```

Best Model Selected by BIC

	Best Model & Type	BIC
model	E	-4429.15236559837
type	univariate, equal variance	-4429.15236559837

Training and Test Error of Best Model

Training Error	Test Error
0.4456853	0.375

True Positive/Negative Rates on Test Data

True Positive (%)	True Negative (%)
91.80328	20.93023

iii: Compare the results with Homeworks 3 & 4. Which method performed the best? Justify your answer. *Here you need to list the previous methods and their corresponding rates.*

Results: First I brought in my code from assignments 3 & 4 to derive the previously reported accuracies. Then, I printed a table comparing the GLM, LDA, QDA, KNN, MclustDA, and EDDA methods. The best performing methods are the GLM, LDA and EDDA methods, all coming in at an accuracy rate of **62.5%**.

To attempt to differentiate between the models to select a "best" model, I compared the True Positive and True Negative rates of the 3 models. Here, as can be seen in the table printed below, the models perform equally well in both categories.

Based on this, I would conclude, for this data set, the 3 models can be seen as equally successful given the training and test set breakdowns we were asked to use.

Accuracy on Test Data Set by Method

GLM Accuracy (%)	LDA Accuracy (%)	QDA Accuracy (%)	KNN Accuracy (%)	MclustDA Accuracy (%)	EDDA Accuracy (%)
62.5	62.5	58.65385	50	54.80769	62.5

```
##          Model Measure Accuracy %
## 1:  glm.true.positive   91.80328
## 2:  glm.true.negative   20.93023
## 3:  lda.true.positive   91.80328
## 4:  lda.true.negative   20.93023
## 5: EDDA.true.positive   91.80328
## 6: EDDA.true.negative   20.93023
```

Exercise #4: Continue from Homeworks 3 & 4 using the **Auto** dataset from 4.7.11. Fit a classification model (using the predictors chosen for previous homework) using MclustDA function from the mclust-package. Use the same training and test set from previous homework assignments.

i: Do a summary of your model.

- What is the best model selected by BIC? Report the model name and BIC.
- What is the training error? What is the test error?
- Report the True Positive Rate and the True Negative Rate.

Results: Here, I loaded the dataset and split into training and test - removing the predictors that have not been used on previous assignments. I used the same predictors as I did in previous assignments. These are *cylinders*, *weight*, *displacement*, *horsepower*, and *year*. I then used MclustDA and reported the best model*(EEV - ellipsoidal, equal volume and shape)* and it's BIC *(-10847.8)*, according to the entire summary.

Then I used this model to predict the Weekly test data set **mpg01**. I reported both the training *(0.0544218)* and test error *(0.0816327)* rates as well as the True Positive *(93.75%)* and True Negative *(90%)* rates on the test set. These are summarized in tables below and will be used for future comparisons.

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## MclustDA model summary:
##
##  log.likelihood   n  df       BIC
##       -4901.032 294 184 -10847.84
##
## Classes   n Model G
##       0 146   EEV 6
##       1 148   EEV 5
##
## Training classification summary:
##
##      Predicted
## Class   0   1
##     0 135  11
##     1   5 143
##
## Training error = 0.05442177
```

Best Model Selected by BIC

	Best Model & Type	BIC
model	EEV	-10847.8425983576
type	ellipsoidal, equal volume and shape	-10847.8425983576

Training and Test Error of Best Model

Training Error	Test Error
0.0544218	0.0816327

True Positive/Negative Rates on Test Data

True Positive (%)	True Negative (%)
93.75	90

ii: Specify modelType = "EDDA" and run the MclustDA again. Do a summary of your model.

- What is the best model selected by BIC?
- Find the training and test error rates.
- Report the True Positive and True Negative Rate.

Results: Here, I used the same predictors as I did in previous assignments. These are *cylinders*, *weight*, *displacement*, *horsepower*, and *year*. I then used **EDDA** modelType from the Mclust function and reported the best model*(VVV - ellipsoidal, varying volume, shape, and orientation)* and it's BIC *(-12129.9)*, according to the entire summary.

Then I used this model to predict the Weekly test data set **mpg01**. I reported both the training *(0.1054422)* and test error *(0.0816327)* rates as well as the True Positive *(95.83333%)* and True Negative *(88%)* rates on the test set. These are summarized in tables below and will be used for future comparisons.

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## EDDA model summary:
##
##  log.likelihood   n df       BIC
##        -5951.26 294 40 -12129.86
##
## Classes   n Model G
##       0 146   VVV 1
##       1 148   VVV 1
##
## Training classification summary:
##
##      Predicted
## Class   0   1
##     0 128  18
##     1  13 135
##
## Training error = 0.1054422
```

Best Model Selected by BIC

	Best Model & Type	BIC
model	VVV	-12129.8637527857
type	ellipsoidal, varying volume, shape, and orientation	-12129.8637527857

Training and Test Error of Best Model

Training Error	Test Error
0.1054422	0.0816327

True Positive/Negative Rates on Test Data

True Positive (%)	True Negative (%)
95.83333	88

iii: Compare the results with Homeworks 3 & 4. Which method performed the best? Justify your answer. *Here you need to list the previous methods and their corresponding rates.*

Results: First I brought in my code from assignments 3 & 4 to derive the previously reported accuracies. Then, I printed a table comparing the GLM, LDA, QDA, KNN, MclustDA, and EDDA methods. I used the KNN with K=5 from Assignment 4 because that was the best performing KNN model. The best performing method is the Generalized Linear Model with an accuracy of **92.85714%**. The QDA, MclustDA and EDDA models are the next best performing models at **91.83673**. As with the last assignment, the QDA is slightly better than the LDA for this data set and the KNN method was least successful in predicting **mpg01**.

Based on the table presented below, I can conclude that, for this data set, the Generalized Linear Model is superior.

Accuracy on Test Data Set by Method

GLM Accuracy (%)	LDA Accuracy (%)	QDA Accuracy (%)	KNN Accuracy (%)	MclustDA Accuracy (%)	EDDA Accuracy (%)
92.85714	90.81633	91.83673	89.79592	91.83673	91.83673

Exercise 5: Read the paper "Who Wrote Ronald Reagan's Radio Addresses" posted on D2L. Write a one page (no more, no less) summary. *You may use 1.5 or double spacing*

Results:

The purpose of the study was to examine the authorship of Ronald Reagan's radio broadcasts, during his campaign for US Presidency which took place from 1975 and 1979. Of the over 1000 radio broadcasts given, there was some question as to the authorship for 312 of them. For the remaining radio addresses, there exists Reagan's original drafts, which eliminates the doubt of authorship. The study used semantics and non-contextual word choices as features in their data analysis.

The data was comprised of the texts from all radio addresses, as well as several newspaper columns which are known to have been drafted by Peter Hannaford. A similar study to this was done by Augustus De Morgan in his *Budget of Paradoxes* where they noticed the possibility to identify authorship by examining the average length of words used in the composition. This study took on 4 parts. In part 1, they learned how to discriminate between the writing styles of Reagan and his collaborators. In part 2, they use exploratory methods to identify some features that would differentiate Reagan's style from that of his collaborators. In part 3, they presented a full Bayesian approach allowing them to estimate the posterior odds of authorship. Lastly, in part 4, they summarized their approach looking at the comparison of predictions by the "best" machine learning methods.

Through feature selection, they were able to capture the elements of Ronald Reagan's style, as a writer, that would assist in predicting authorship. They used three types of features: words, n-grams, and semantics. Words are self-explanatory. N-grams were the ordered sequences of the adjacent words to the word used and semantic features relate to patterns of composition.

For words, they focused on 267 of the most frequent 3000 words used by Reagan and Hannaford. They then used a technique of categorization called SMART that removed words that were not considered useful. In the end, they derived 62 key words to use as features. For semantics, they used 21 semantic features that had been discussed in a paper by Collins and Kaufer (2001). Using the concept of information gain, they were able to select the words with the highest information gain ratio scores.

In the end, the "goodness-of-fit study indicated that the Negative-Binomial model was appropriate for word counts and semantic features counts data". They based their word selection scheme and the likelihood of the data upon this model. They also chose 21 sets of constants based on two smaller sets of studies that used 90 and 120 words from speeches drafted by Reagan and the other collaborators. The fully Bayesian Negative Binomial model was very consistent - both with the 21 sets of constants and in terms of predicting the 312 speeches for which authorship was unknown. They segregated 1975 from the other years of 1976-1979 and still were able to obtain consistently accurate predictions on speeches over a variety of topics.

One interesting "shortcut" used was the assumption that words were independent from one another. The authors conceded that although removing the presence of syntax is not true in reality, focusing only "on high frequency, non-contextual words" produced a reasonable initial approximation.

Exercise 6: Last homework you chose a dataset from (https://archive.ics.uci.edu/ml/datasets.html). Please do some initial exploration of the dataset. Please report the analysis you did and discuss the challenges with analyzing the data. *Any plots for this question need to be done using only GGplot2-based plots.*

Results: I chose the credit-screening data set from the website above. (https://archive.ics.uci.edu/ml/datasets/Credit+Approval). This data set concerns credit card applications, as the website says. Looking at the summary below, we can see that there are some missing values that may require imputation. These are denoted by **?** in the data set. A1, for example, has 12 missing, or **?** values.

The dependent variable is the factor **A16**. A '+' indicates positive screening (approval?) where as a '-' indicates a negative screening (decline?) on the credit application.

What I found most intriguing about this data set was that the variable names have been removed to protect the confidentiality of the applicants. I had not thought about the reality that this is probably done frequently in some industries - healthcare and finance being two that I am very interested in. I think this, on the surface, makes the data seem more "daunting". In reality though, I think it could be useful in removing biases that we have. One binary factor independent variable, for example, is **A9**. It lists a 't' and an 'f'. Maybe that variable actually represents "College Student", "Previous Bankruptcy", "Income > 150k", etc. This would inherently bias my initial exploration.

My first step in working with this data set would be to attempt to impute values based on the other values in the variable. In doing so, I would change all '?' values to 'NA'. Next I would begin the imputation process and determine if any observations should be removed (too many NAs, for example). Then I could begin the process of looking for correlations and building models.

```
##   A1    A2   A3 A4 A5 A6 A7   A8 A9 A10 A11 A12 A13   A14 A15 A16
## 1  b 30.83 0.00  u  g  w  v 1.25  t   t   1   f   g 00202   0   +
## 2  a 58.67 4.46  u  g  q  h 3.04  t   t   6   f   g 00043 560   +
## 3  a 24.50 0.50  u  g  q  h 1.50  t   f   0   f   g 00280 824   +

##  A1            A2            A3          A4       A5             A6
##  ?: 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
##        A7            A8          A9      A10         A11       A12
##  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
##  A13          A14          A15           A16
##  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
```

The below summary shows that now, after the change, all *‘?’* values are *NA*.

```
##     A1            A2            A3            A4         A5
##  ?   :  0   22.67  :  9   Min.   : 0.000   ?   :  0   ?   :  0
##  a   :210   20.42  :  7   1st Qu.: 1.000   l   :  2   g   :519
##  b   :468   18.83  :  6   Median : 2.750   u   :519   gg  :  2
##  NA's: 12   19.17  :  6   Mean   : 4.759   y   :163   p   :163
##             20.67  :  6   3rd Qu.: 7.207   NA's:  6   NA's:  6
##             (Other):644   Max.   :28.000
##             NA's   : 12
##        A6           A7            A8         A9      A10
##  c      :137   v      :399   Min.   : 0.000   f:329   f:395
##  q      : 78   h      :138   1st Qu.: 0.165   t:361   t:295
##  w      : 64   bb     : 59   Median : 1.000
##  i      : 59   ff     : 57   Mean   : 2.223
##  aa     : 54   j      :  8   3rd Qu.: 2.625
##  (Other):289   (Other): 20   Max.   :28.500
##  NA's   :  9   NA's   :  9
##       A11        A12     A13          A14           A15           A16
##  Min.   : 0.0   f:374   g:625   00000  :132   Min.   :     0.0   -:383
##  1st Qu.: 0.0   t:316   p:  8   00120  : 35   1st Qu.:     0.0   +:307
##  Median : 0.0           s: 57   00200  : 35   Median :     5.0
##  Mean   : 2.4                   00160  : 34   Mean   :  1017.4
##  3rd Qu.: 3.0                   00080  : 30   3rd Qu.:   395.5
##  Max.   :67.0                   (Other):411   Max.   :100000.0
##                                 NA's   : 13
```

Now I want to correct a couple of variables - A2 is listed as a factor but is numeric, as is A14. A16 (the dependent variable) is listed as ‘+’ or ‘-’. I’m making the assumption that ‘+’ means a ‘Positive’ decision on credit screening and a ‘-’ represents a ‘Negative’ decision on the credit screening. These are represented by ‘P’ and ‘N’ values. Once these are changed, I am printing the header to reflect the changes before I move into imputation.

```
##   A1  A2   A3 A4 A5 A6 A7   A8 A9 A10 A11 A12 A13 A14 A15 A16
## 1  b 158 0.00  u  g  w  v 1.25  t   t   1   f   g  70   0   P
## 2  a 330 4.46  u  g  q  h 3.04  t   t   6   f   g  13 560   P
## 3  a  91 0.50  u  g  q  h 1.50  t   f   0   f   g  98 824   P
```

Our last step before visualization will be to impute the missing values (NAs) with the most often occuring value in the dataset.

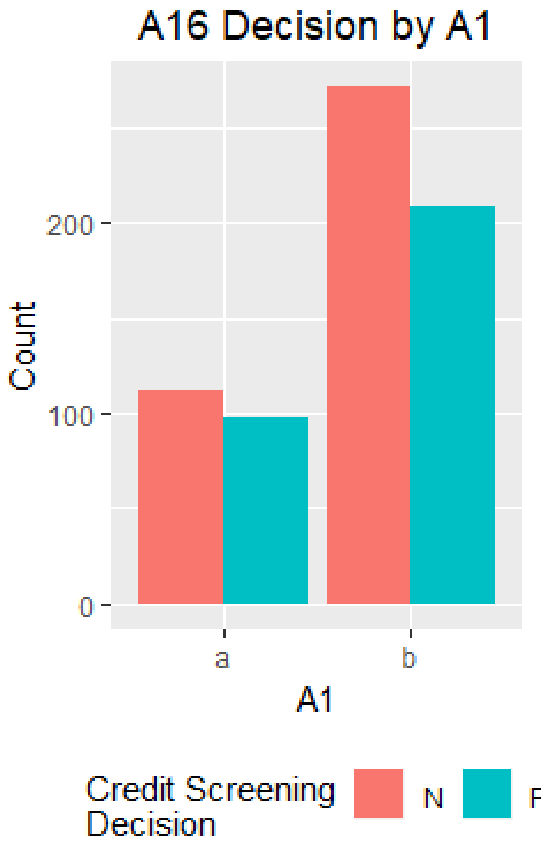
Now we can print the updated summary to show that the NAs have been replaced with the most often occuring variable value for the factors and the mean values for the numeric variables. Now that the NAs are gone, I will do some analysis.

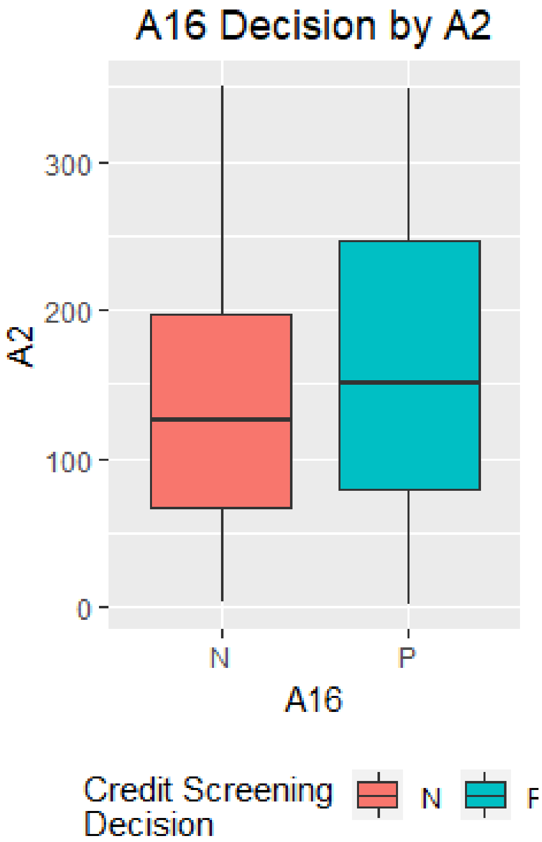
```
##  A1            A2              A3         A4       A5            A6
##  ?:  0   Min.   :  2.0   Min.   : 0.000   ?:  0   ? :  0   c      :146
##  a:210   1st Qu.: 73.0   1st Qu.: 1.000   l:  2   g :525   q      : 78
##  b:480   Median :135.5   Median : 2.750   u:525   gg:  2   w      : 64
##          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          A11       A12
##  v      :408   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
##  j      :  8   3rd Qu.: 2.625                   3rd Qu.: 3.0
##  z      :  8   Max.   :28.500                   Max.   :67.0
##  (Other): 12
##  A13          A14              A15           A16
##  g:625   Min.   :  2.00   Min.   :     0.0   N:383
##  p:  8   1st Qu.: 25.00   1st Qu.:     0.0   P:307
##  s: 57   Median : 54.00   Median :     5.0
##          Mean   : 59.27   Mean   :  1017.4
##          3rd Qu.: 95.00   3rd Qu.:   395.5
##          Max.   :171.00   Max.   :100000.0
##
```

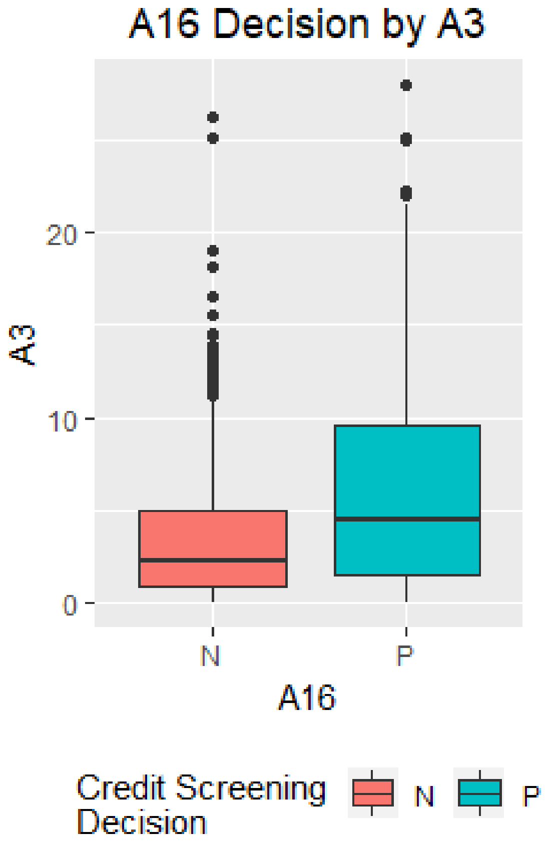
First in my exploration, I plotted the relationship for each variable between that of the dependent variable (A16). For numeric predictors, I used box plots and for factor variables I used histograms.

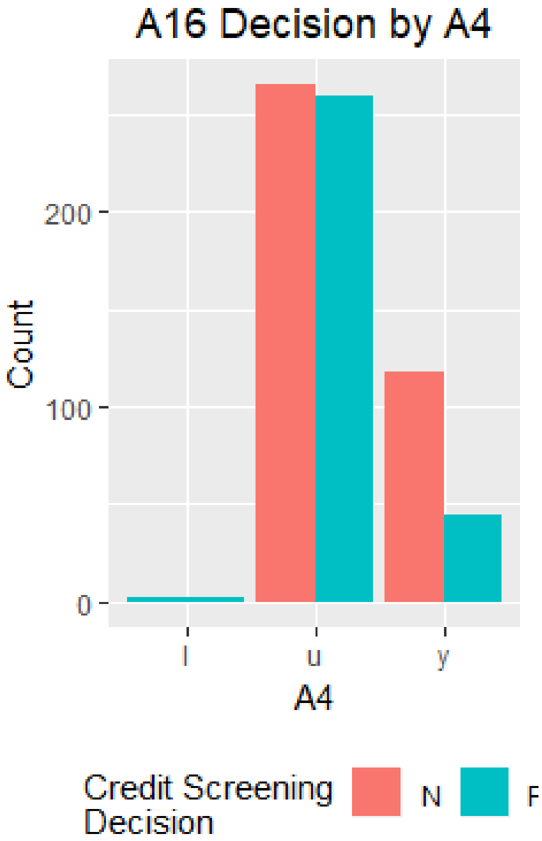
Some of the more interesting things I notice looking at the relationships are:

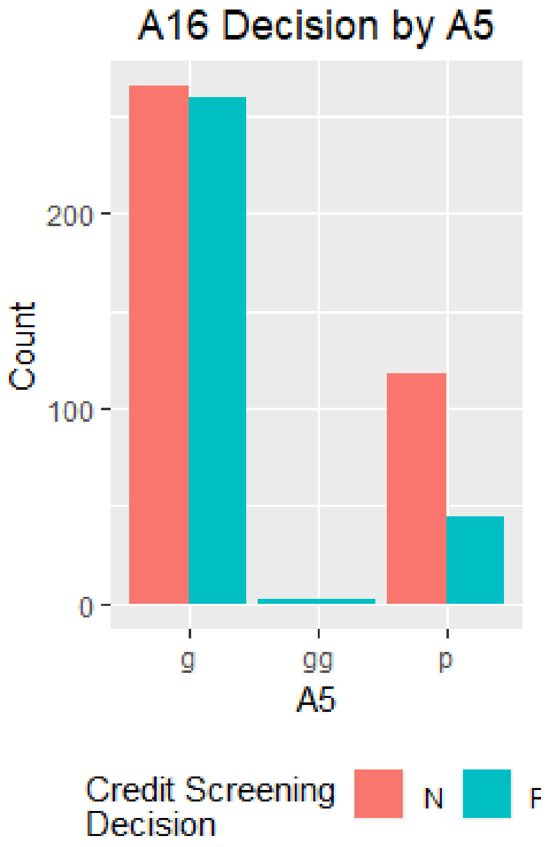
- A higher *A3* value indicates higher likihood toward a '+', or Positive in the credit screening
- Observations with an 'l' for *A4* receive all Positive screening results
- Observations with gg for *A5* receive all Positive screening results
- For *A6*, 'cc', 'q', 'r', 'w', and 'x' result in mostly 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.

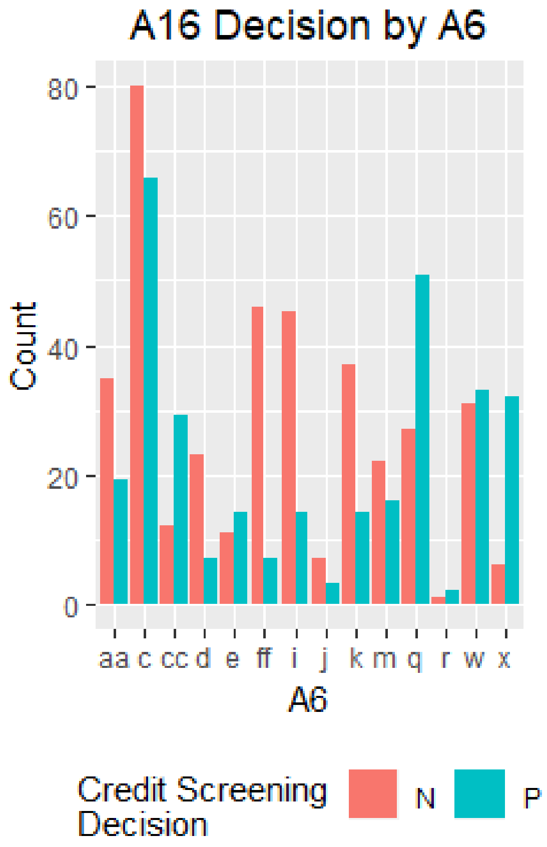

A16 Decision by A1

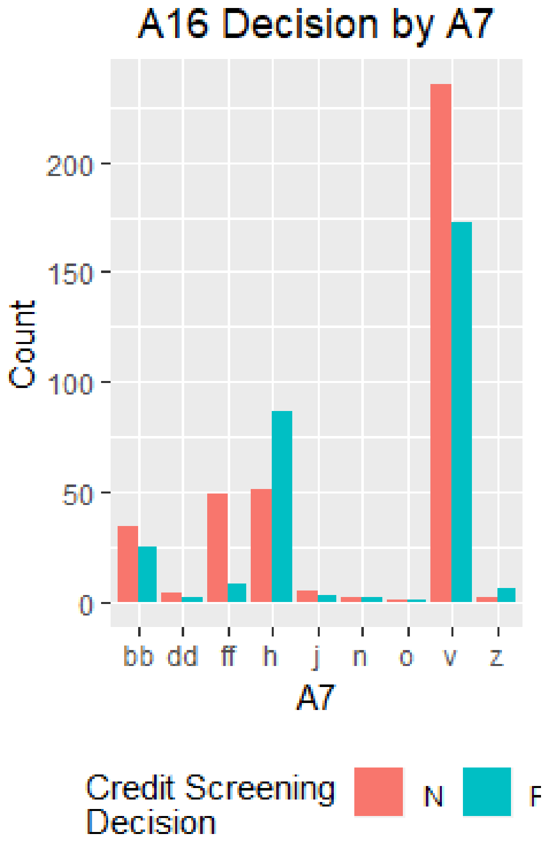

A16 Decision by A2

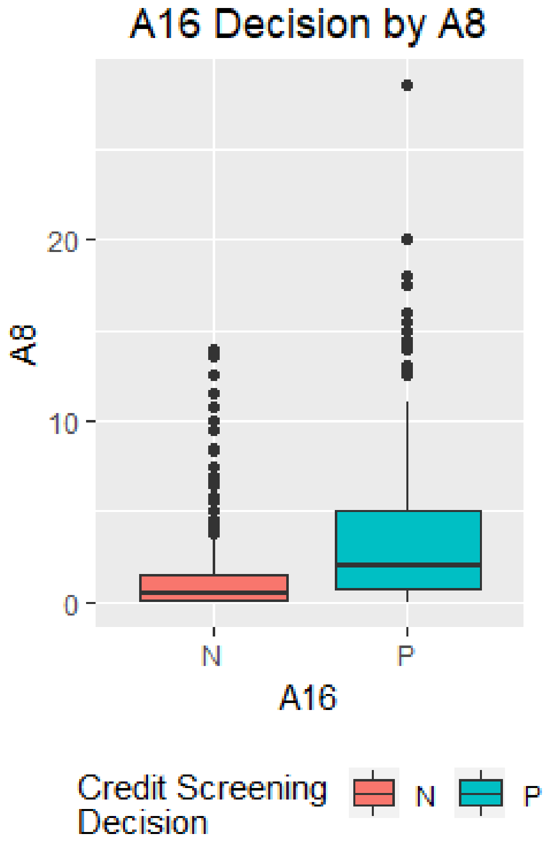

A16 Decision by A3

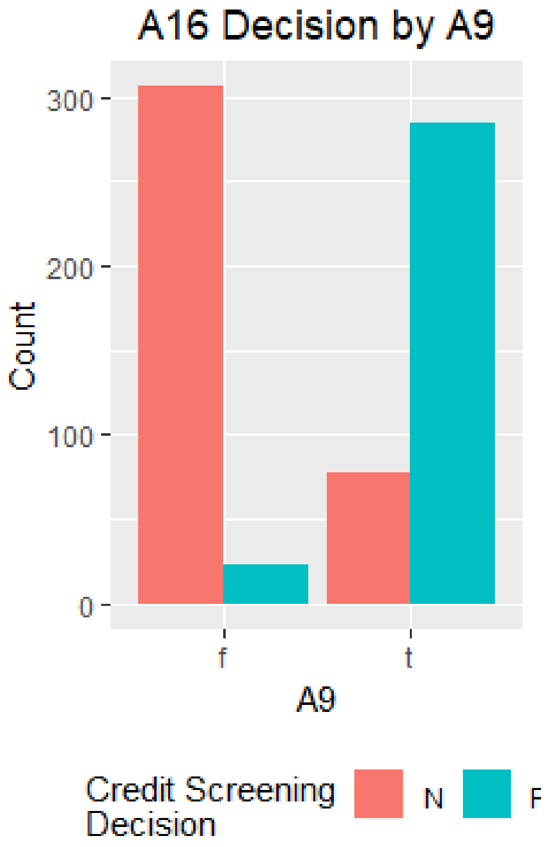

A16 Decision by A4

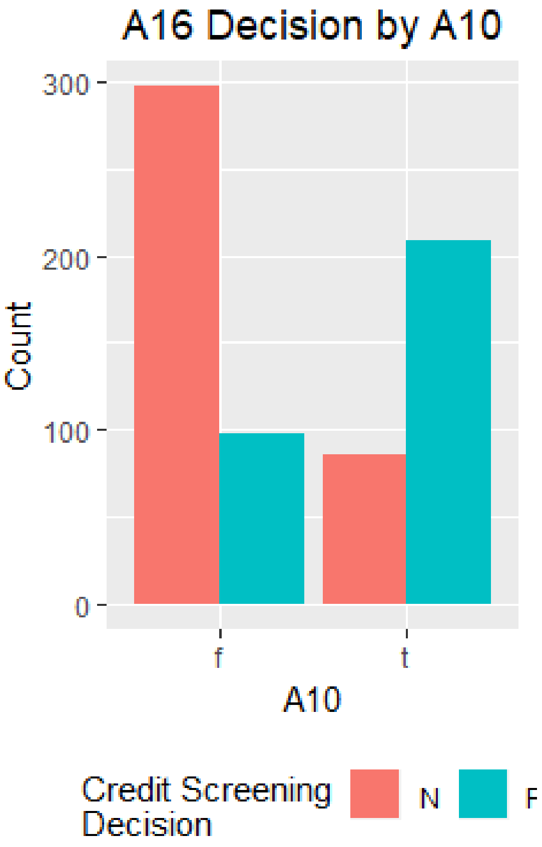

A16 Decision by A5
Count
200
100
0
g
gg
p
A5
Credit Screening Decision
N
F

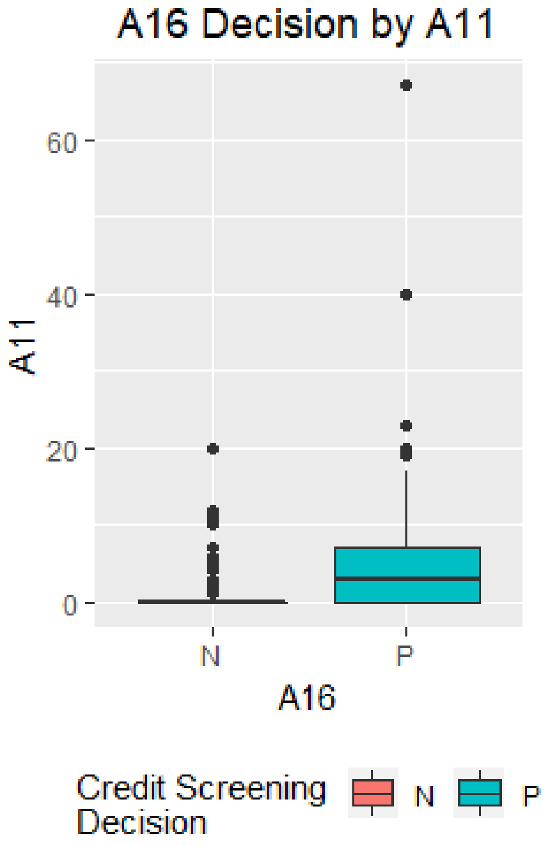

A16 Decision by A6
Count
80
60
40
20
0
aa
c
cc
d
e
ff
i
j
k
m
q
r
w
x
A6
Credit Screening Decision
N
P

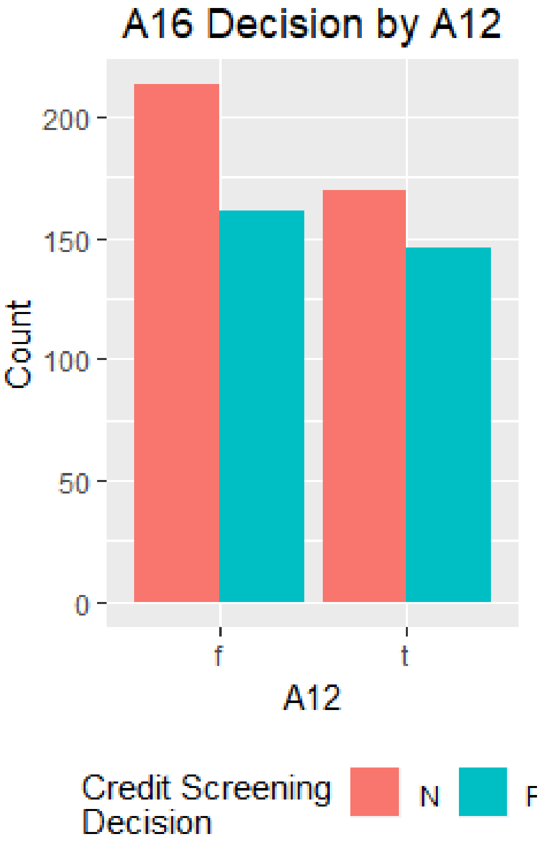

A16 Decision by A7
Count
200
150
100
50
0
bb
dd
ff
h
j
n
o
v
z
A7
Credit Screening Decision
N
F

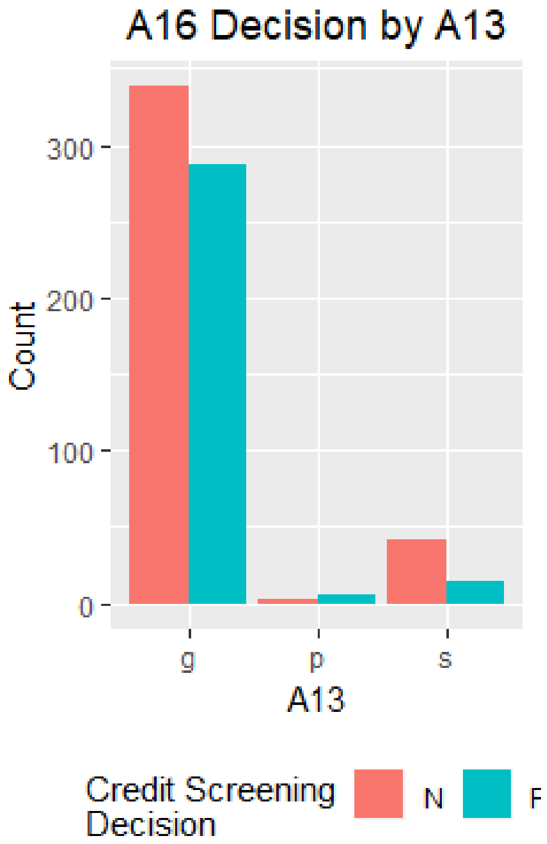

A16 Decision by A8
A8
20
10
0
N
P
A16
Credit Screening Decision
N
P

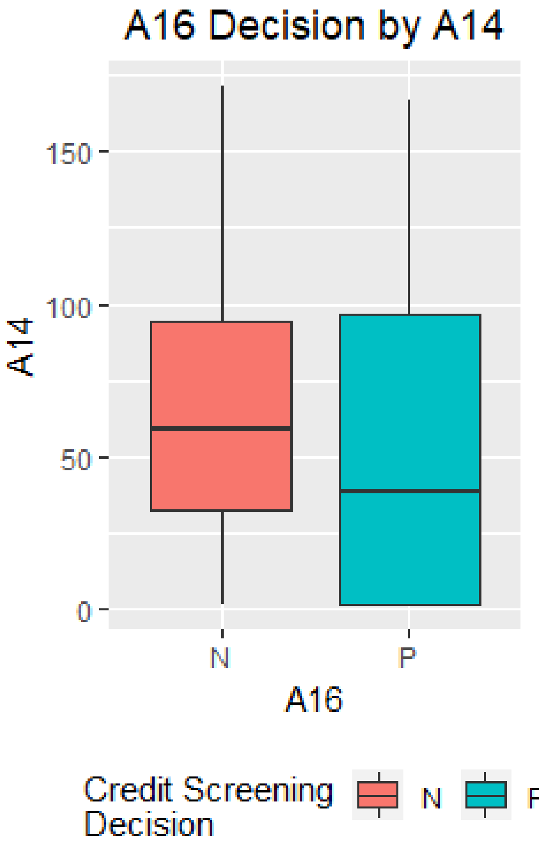

A16 Decision by A9
Count
300
200
100
0
f
t
A9
Credit Screening Decision
N
F

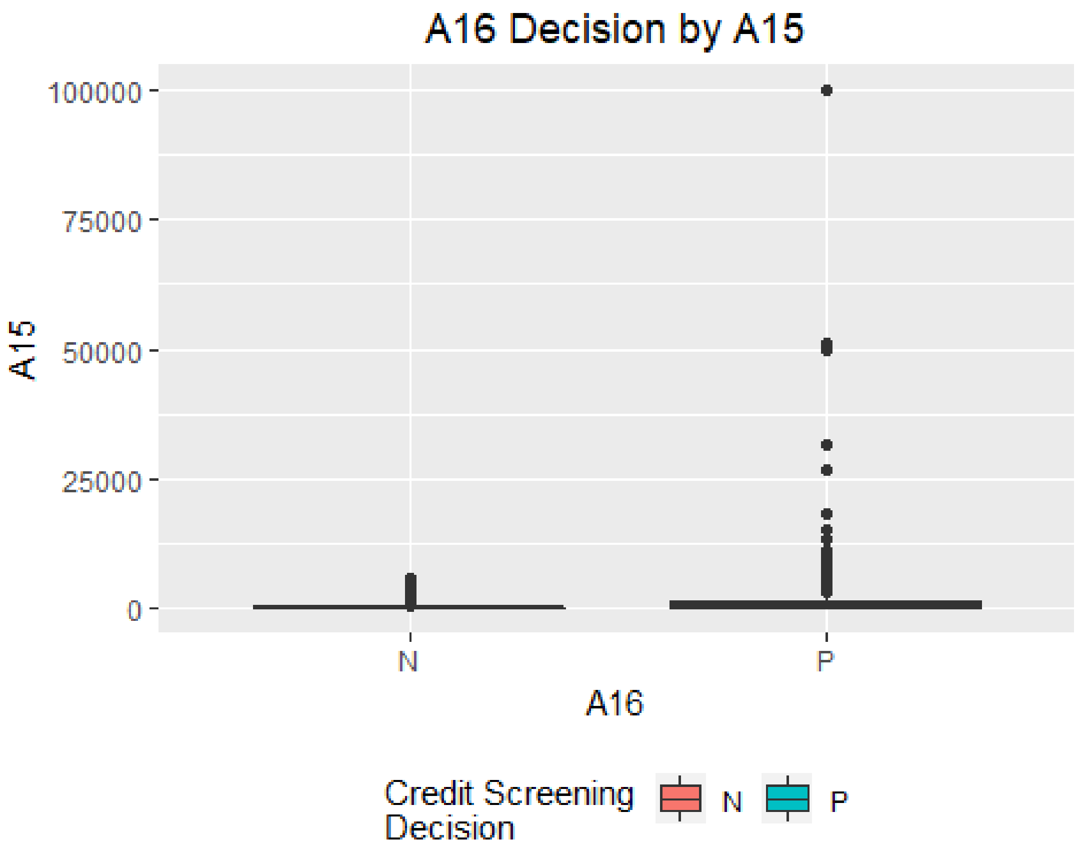

A16 Decision by A10
Count
300
200
100
0
f
t
A10
Credit Screening Decision
N
F


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
F

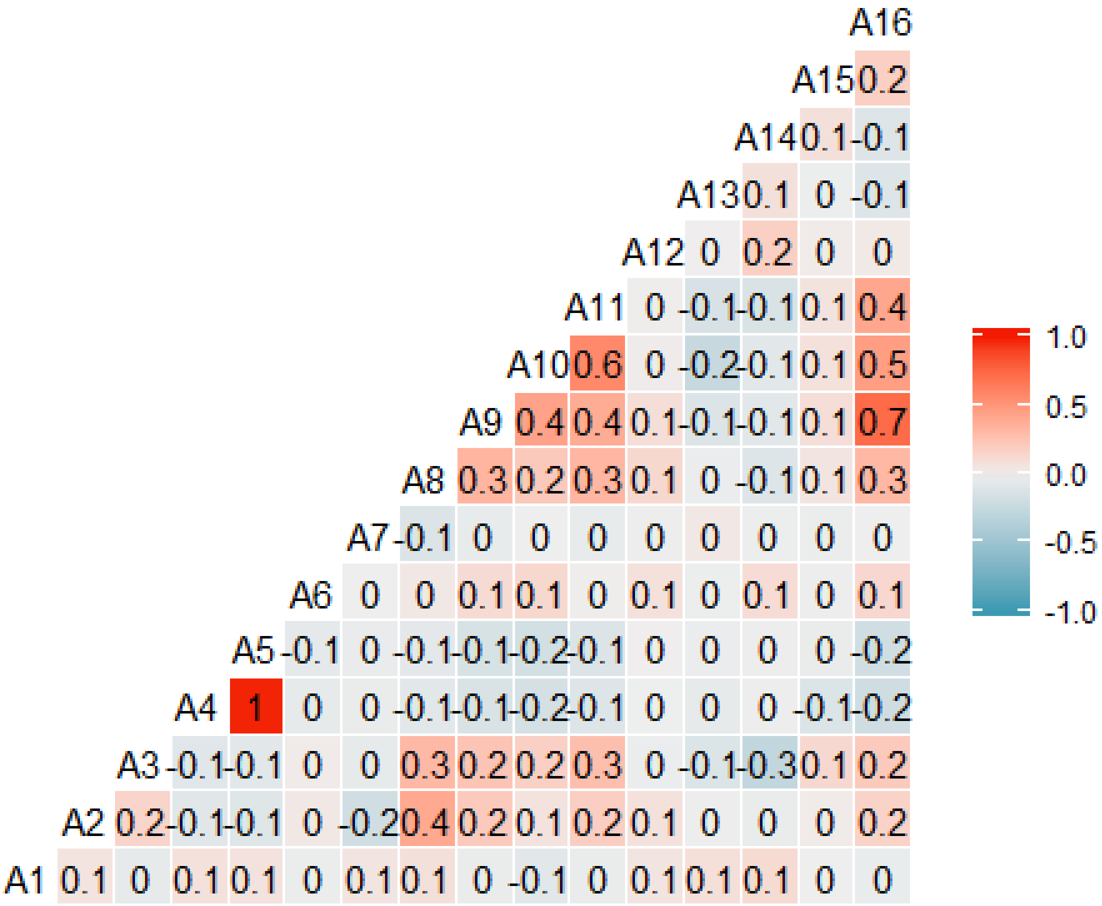

A16 Decision by A13
Count
300
200
100
0
g
p
s
A13
Credit Screening Decision
N
P


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


A16 Decision by A15
A15
100000
75000
50000
25000
0
N
P
A16
Credit Screening Decision
N
P

Now that I've examined the relationships between values of the predictors and the value of the dependent variable, I am going to look at the correlations between the independent and dependent variable.

From the correlation plot below, we see that most of the variables have a very weak impact on the dependent variable. A9, A10, and A11 have the greatest correlation.



	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16
A15															0.2
A14														0.1	-0.1
A13													0.1	0	-0.1
A12												0	0.2	0	0
A11											0	-0.1	-0.1	0.1	0.4
A10										0.6	0	-0.2	-0.1	0.1	0.5
A9									0.4	0.4	0.1	-0.1	-0.1	0.1	0.7
A8								0.3	0.2	0.3	0.1	0	-0.1	0.1	0.3
A7							-0.1	0	0	0	0	0	0	0	0
A6						0	0	0.1	0.1	0	0.1	0	0.1	0	0.1
A5					-0.1	0	-0.1	-0.1	-0.2	-0.1	0	0	0	0	-0.2
A4				1	0	0	-0.1	-0.1	-0.2	-0.1	0	0	0	-0.1	-0.2
A3			-0.1	-0.1	0	0	0.3	0.2	0.2	0.3	0	-0.1	-0.3	0.1	0.2
A2		0.2	-0.1	-0.1	0	-0.2	0.4	0.2	0.1	0.2	0.1	0	0	0	0.2
A1	0.1	0	0.1	0.1	0	0.1	0.1	0	-0.1	0	0.1	0.1	0.1	0	0

To summarize, this data set presented an interesting challenge that I had not previously considered in that the variable names were "masked" for confidentiality. In working with the dataset, this removed the biases that I may otherwise have had, as mentioned above.

Above I've analyzed some key factors of the response, *A16*, that I think would be beneficial in beginning variable selection and model building. I would begin by building different types of models - glm, lda, qda, knn, mclustda, etc. using a combination and interactions of the most influential variables from the correlation plot shown above.