# MSBA 635 Group Consulting Assignment 2

Team 5:

# Business Problem

TAC, LLC a wonderful organization brought to you by five impressively diverse Data Analytics professionals, working intangibly as a third-party consulting group to Flight Inc., a company based out of central Denver, which also specializes in flight training and aircraft rental services and has leading airline services across the US, also spanning flights to and from Europe, Asia, and Africa. The foundation of any airline is in its cargo and making sure it reaches a destination safely and on time. Flight Inc. is looking to us for help on determining causes for flight delays to help minimize and prevent. As flight delay is one of the major issues in aviation systems all over the world. This report explores the internal workings of flight departure delays for Flight Inc., analyzing factors that include but not limited to late flights, cancellations, issues with the air carrier, extreme weather, national aviation system (NAS), late-arriving aircraft, security, scheduling, baggage, fueling, etc. and through all factors, we divide results statistically into two known factor categories, propagation, and non-propagation to determine the best causes in delay and learning which determinates we should have our client put focus on. Doing so analyzes the risk and helps with finding them, to reduce the cost that delays are causing. To alleviate the harm of flight delay, a considerable amount of work is required. It is our task to examine all factors and work to provide sound evidence that Flight Inc. can utilize to prevent delays.

# Summary for our Work.

As data analysts, it is important for us to understand the various features of the dataset, their distribution patterns, trends, anomalies, and behaviors against the determinant or predictive variable. For this, we ran analysis in SAS to conduct initial data exploration tasks which helped us for better visualization and graphical representation. Once we explored the dataset features, we wanted to understand how well we can predict the potential flight delays with our model and to address the business problems of Flight Inc.

In order to predict the best variables that are impacting the flight delays, we wanted to build a regression model. Regression model gives us the flexibility of splitting the data and making assumptions about the predictor variables and their correlations with the response variable. Primarily, we ran the LASSO model of regression to exclude the variables which might give misleading results because of outliers or any data anomalies. Along with this model, we also used the sub selection of the variables using forward and backward stepwise selection.

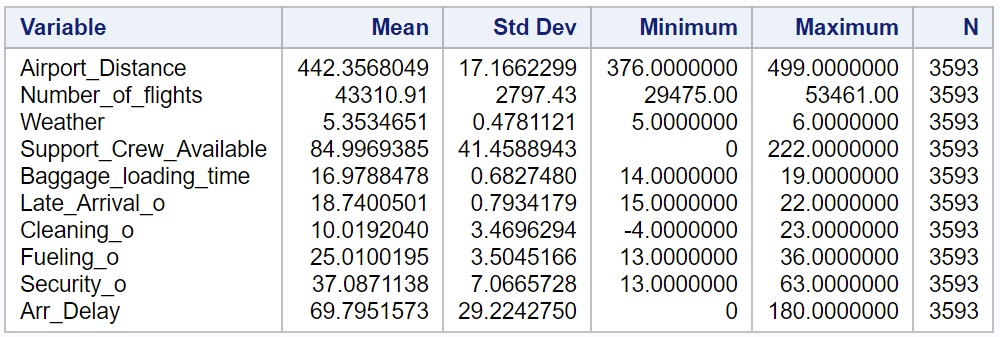
The reason we used new methods other than linear regression model is that we wanted to increase the predictability, interpretation and to best fit the data into the model we developed reducing the oversimplification which is technically called ‘bias’ and overspread of the data called ‘variance’.

# Understanding Our Data

The total number of observations in this dataset is 3593 and the feature variables are 11. Out of 11, one variable is the response variable and the rest of 10 variables are predictor variables.

It is always crucial to understand the predictor variables and their behaviors with the response variable. In our flight's dataset, we have 10 such variables and they individually have their features as in giving the data about the total number of flights, distance between the airports in miles, weather conditions which gives the mild and extreme levels, number of support crew, and the time in minutes for loading, fueling, cleaning, late arrivals, time taken for security check-ins.

# Summary Statistics for Numerical Variables



# Exploratory Data Analysis

Missing Data Patterns across Variables:



There are no missing values in the data. This is checked using the “Describe missing data” tab in “Data” task.

**To understand more descriptive nature of the variables, the following plotting and graphs are used to understand the trends in data.**

|  |  |
| --- | --- |
|  | The histogram plotted on the left shows us the distribution of percentage of aircraft fueling. The data is normally distributed. The average time taken for fueling the aircraft is 25 minutes. |

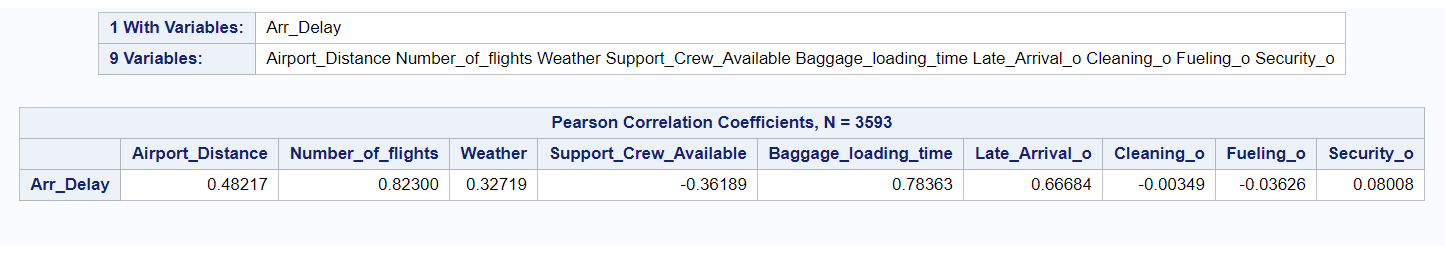
|  |  |
| --- | --- |
|  | The average fueling time of the aircraft is 25 minutes. There are outliers but we are not considering them. |

|  |  |
| --- | --- |
|  | As we can see in the boxplot with fueling and carrier as the categorical variable, there is a significant unevenness between the data points. This indicates that the HA carriers are taking uneven fueling timing from 21 minutes to 31 minutes. Whereas the AS carriers are taking more time than the average fueling time that is 26 minutes to 31 minutes. |

|  |  |  |
| --- | --- | --- |
|  | | This is a histogram plotted for the late arrivals of the flights. As we can see in the plot, 50% of the time, the flights are 18 minutes late on an average to the airport which could potentially impact the flight arrivals. The data spread is slightly right skewed and there are very less times that the flights arrived beyond 20 minutes. |
|  | This boxplot for late\_arrival is skewed upwards slightly. The outliers are not much impactful, and we are not considering here. The data spread is slightly right skewed and there are very less times that the flights arrived beyond 20 minutes. | |

|  |  |
| --- | --- |
|  | This is an interesting boxplot plotted against the carrier variable which is categorized. We have plotted against late\_arrival and found that HA carriers have significantly more delays when compared to other carriers. The HA carriers average time is delayed by 18-20.5 minutes every time. Whereas the other carriers are consistent in their arrivals. |

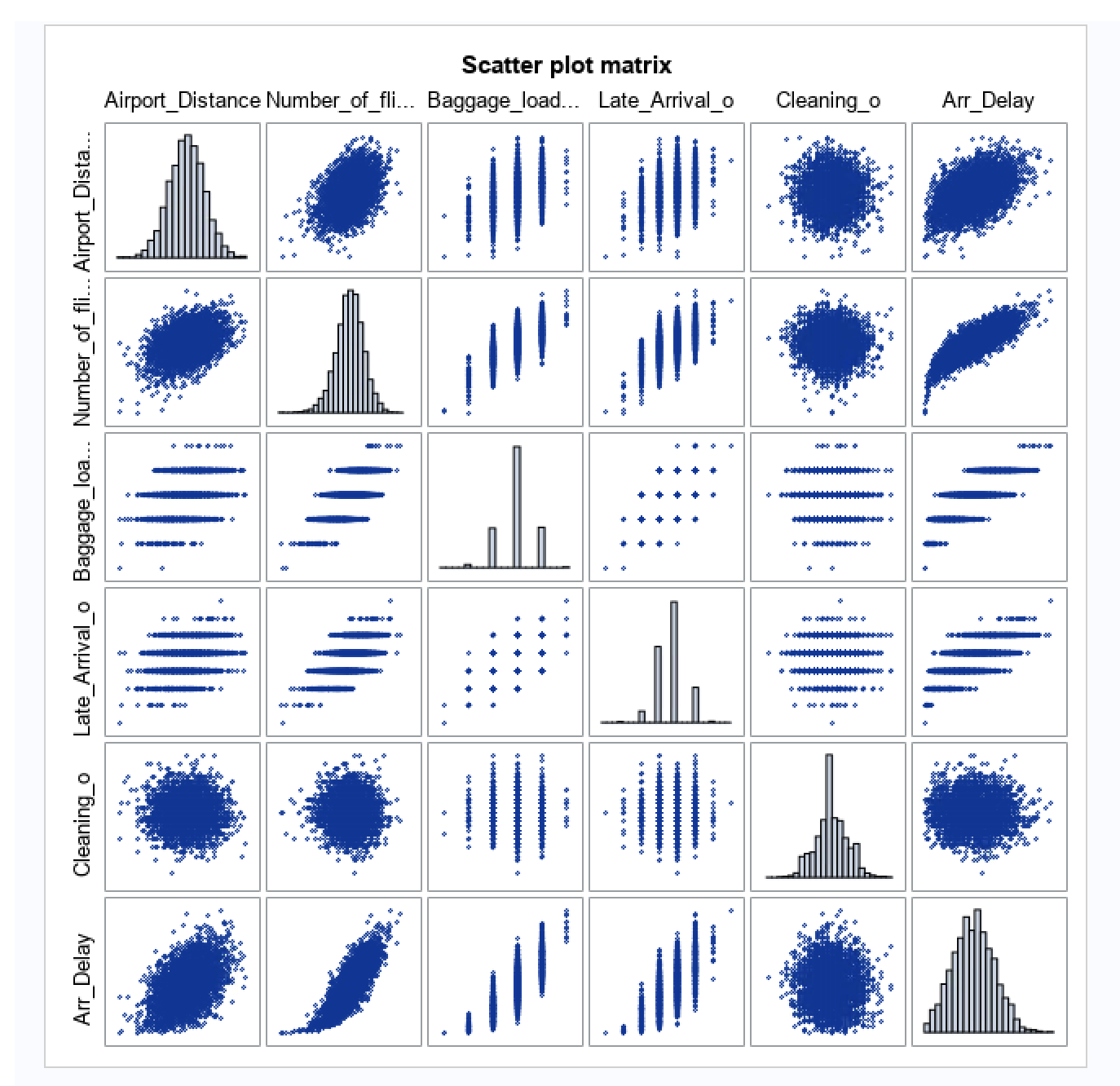
# Scatter Plots: Observed Correlation between the Response and Predictor variables



The correlation between the response variable and the predictor variables can be viewed in the table above and from this we learned that the number of flights, baggage loading time and the late arrivals to the airport are very well correlated to the arrival delays. We observed that the correlations are positive.

To explain in simple terms, as the number of flights is more in the airport and they are stopped for some reason, that will delay the arrivals more. Likewise, if there are more flights arriving late to the airport, that will clearly worsen the arrival times of the flights. The baggage loading time is also positively impacting the arrival delays.

The below scatterplot helped us to visualize the correlation between the predictor variables and the response variable.



# Modeling and Results

As our dataset is labelled with specific variables and the response variable is a continuously varying variable, we are using Supervised Linear Regression Model to understand the relationship between predictor variable and response variables. Our target variable is Arr\_Delay and it is the number of minutes the flight got delayed. Regression modelling gives a flexibility of choosing the best fit of the data by splitting the entire data set to training, validation and test data. Instead of regular approach, we wanted to improve our modelling in terms of interpretation and predictability and hence, we learned a new method of Subset selection and LASSO of the data to understand where we can fit the data in the model.

The reason why we shifted our gears to apply new modelling techniques is that we wanted to understand the most important variables that impact the predictor variables and get rid of the least important variables. This can be attained by the above-mentioned techniques. In Subset selection, we have two methods:

1. Forward Subset selection: This begins with model containing no predictor and adds one predictor at a time until no further improvement is possible.
2. Backward Subset selection: This begins with model containing all predictors and then deleting one predictor at a time until no further improvement is possible.

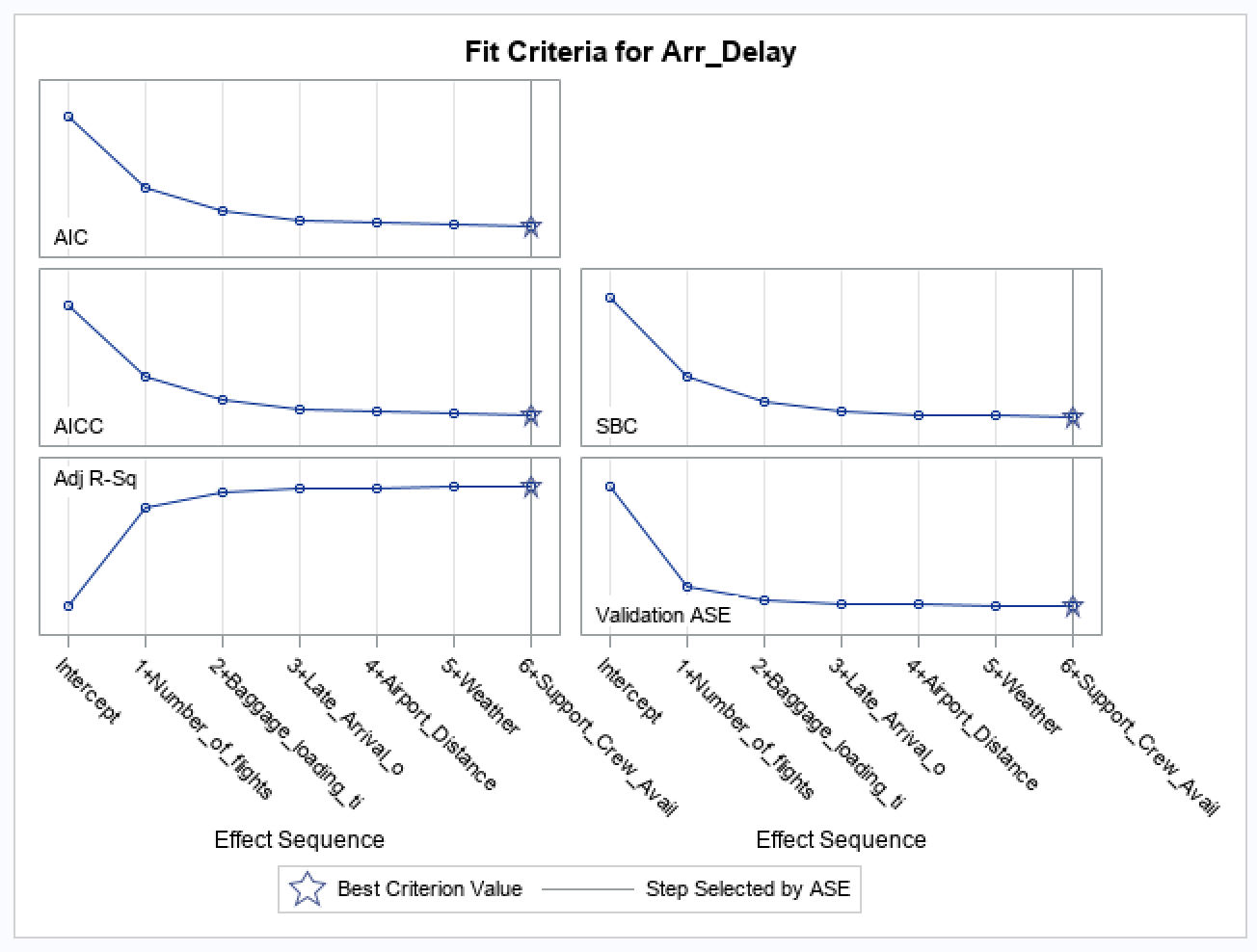
LASSO Regression is a type of regularized regression which improves the model that shrinks the variables with large coefficient and removes the less important features to avoid overfitting. As LASSO tries to fit the best model through three chunks on data (training, validation & test), this makes the final model easier and simpler to interpret.

All in all, all these models are used to remove the least important variables in the dataset. But the question is where we stop removing the variables. In order to know where to stop the exclusion, we are using validation set (is also a part of training set) which helps determines when to terminate the predictor variable selection. Considering validation set gives us the best model predictability using the test set errors on unseen data.

We used **60%** of our data for training data to fit our model. **20%** of our data was used for the validation set. Because we want our machine learning model to generalize to data it has not seen before we use another **20%** of the data for testing set to measure the generalization performance. A random seed of **11** is used to build our model.

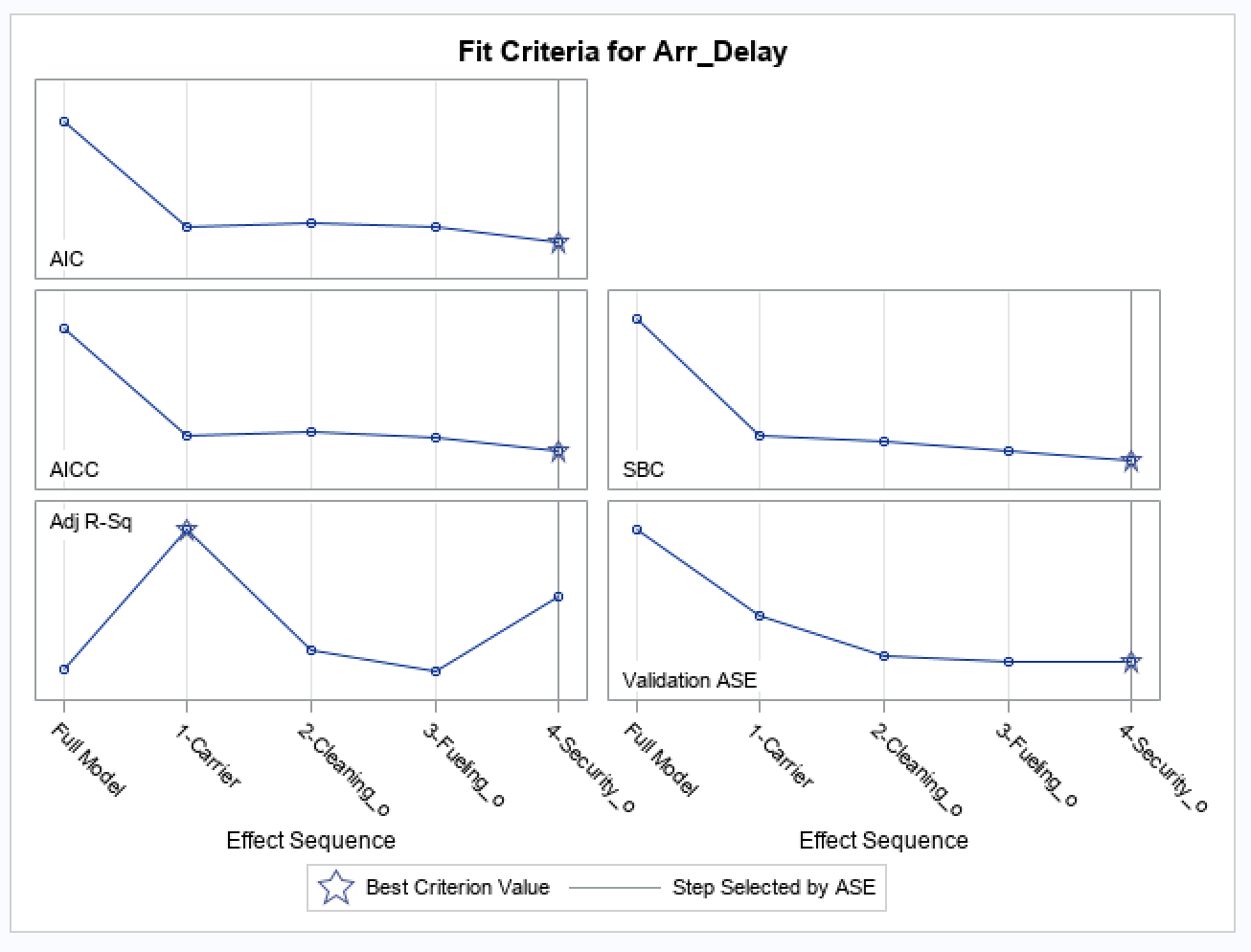
We fit a linear regression model, a LASSO model, a Forward Stepwise Selection, and a Backward Stepwise Selection model using SAS studio.

# Forward Subset Selection Method



In the forward subset selection, the elimination of the predictors can we observed at last as this model chooses the best correlated variables with the response variable. Hence, we can we in step 1, there is the highest correlated variable which is number\_of \_flights, step 2 highest is baggage\_loading\_time, so on and so forth.

# Backward Subset Elimination Selection Method



In the backward subset selection, the model eliminates the least correlated variable and as we can see from the fit criteria above, the least correlated variable is the carrier. This process continues with the rest of the least important or technically correlated variables.

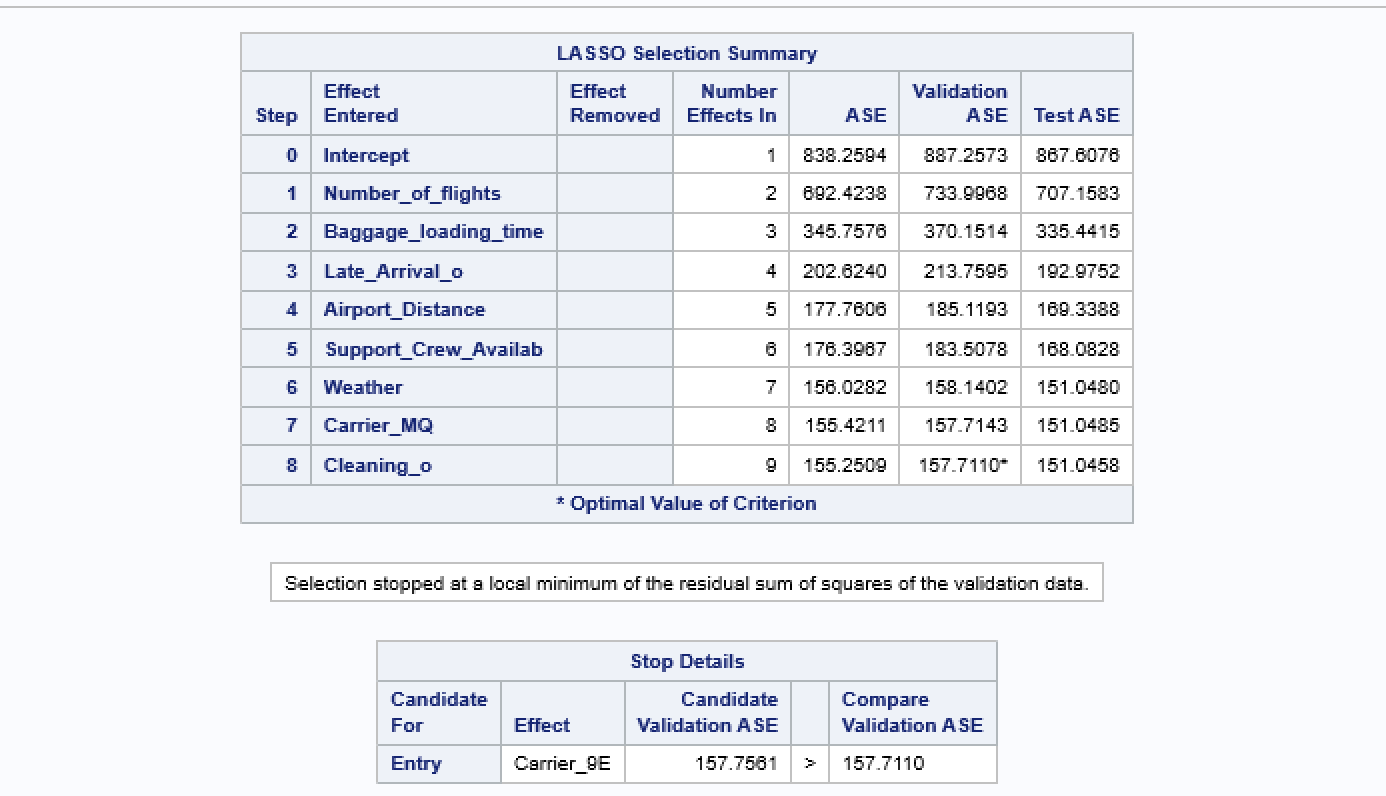
|  |  |
| --- | --- |
| Method | Testing set ASE |
| Linear Regression  (all predictor variables) | 152.35953 |
| LASSO | 151.04579 |
| Forward Stepwise Selection | 151.66450 |
| Backward Stepwise  Selection | 151.66450 |

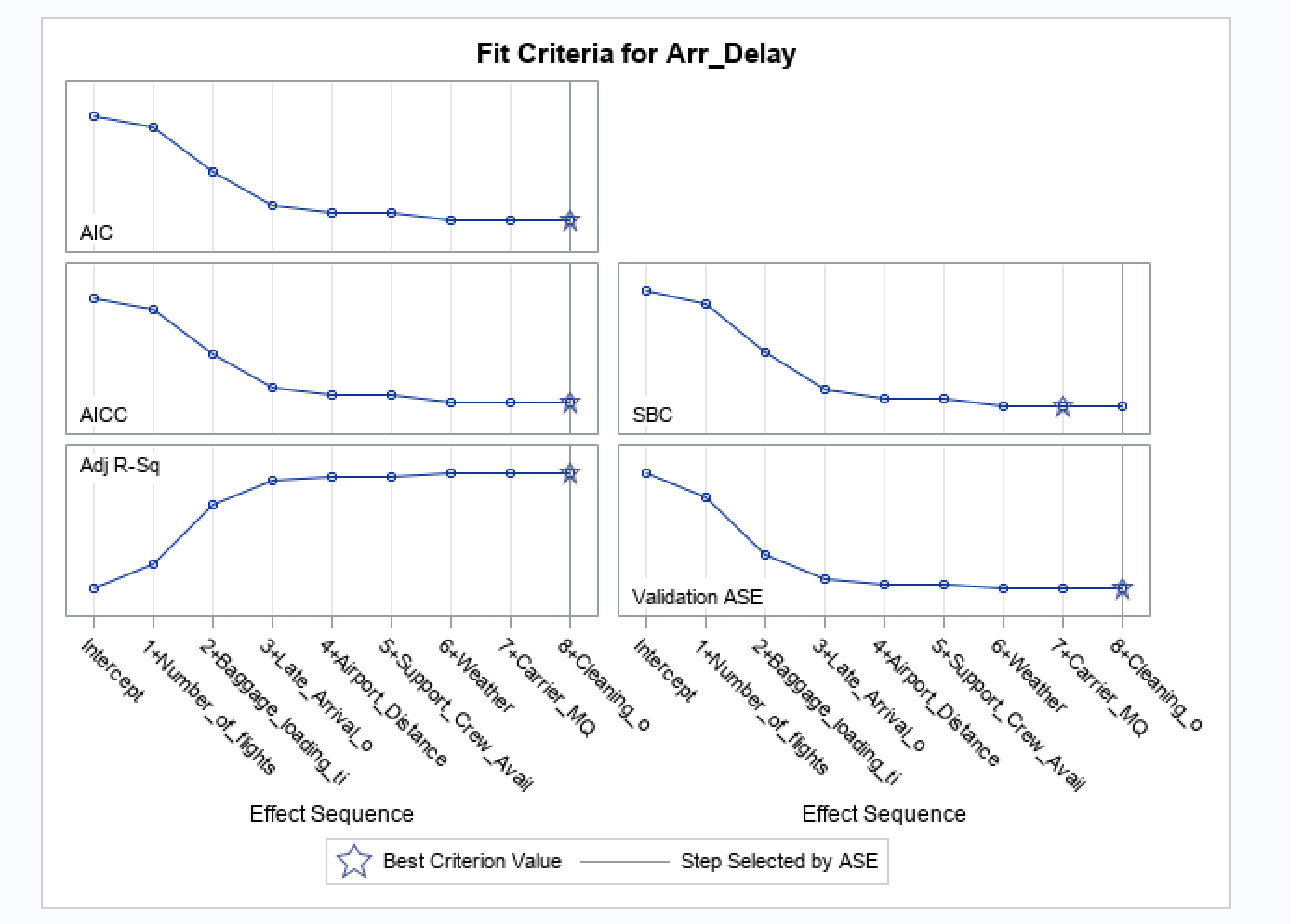
We chose the LASSO model because it helped us to select only the important features of the flight dataset. As the Forward and backward subset selection is a time-consuming process of adding the predictors forward and backward, LASSO has a more advanced system on shrinking the least important variables that do not impact the response variables.

The best model is LASSO with ASE(Test) of **151.04579.**

# LASSO

Below is the table of coefficient estimates from the best model.







# Recommendations to our client

As a group of analysts to solve the business problem of reducing the flight delays for Flight Inc, we had an opportunity to run various advanced regression models on their historic data. Apart from the classic Vanilla Linear Regression modelling, we used LASSO regression to see how our data behaves when we remove the least impacting variables on the response variable. To our surprise, the model really has least errors and the prediction capability has been improved a lot better than the Subset selection method.

As per the model dynamics, we strongly recommend being watchful about the late\_arrivals and the number\_of\_flights features of the airports. These two features are highly impacting the delay times in the flight arrivals to the airport. The average 25 minutes delays can be reduced in the future by noticing why the same flight in a different airport is landing or departing late. There can be other reasons for the delay like the baggage\_loading \_time where only 50% of the time the baggage is being loaded without delay. But the delay time can be reduced by decreasing the loading time by 30% into the aircraft which also significantly increases the flight arrival time.

In simple terms, if the operating management of the connecting airports fail to load and unload the baggage and delay in fueling the planes which could be an impactful feature, there may be a delay in departures and arrivals. In the future, we recommend the Flight Inc., to plan and investigate the reasons for the delays in late\_arrivals and baggage\_loading\_time and reduce these timings in minutes to avoid major arrival delays. This gives a profit to the organizations and a peace travel to the passengers.

# **Mammoth Bank:**

Our Business Problem

TAC, LLC widely known for their impressive insight and due diligence in statistical reporting and elaboration in facts, an organization ran by five unique individuals, has been asked to consult for Mammoth Bank.  Mammoth Bank, founded in Lexington Kentucky, and known as one of the largest most profitable banks in its industry since 1964.  Our group is being asked to come in as an arbitrary group to analyze provided data on customers to best predict those suited, for Mammoth Bank’s continued success in staying profitable in the banking industry.  As you may not know, over half of the money made by banks still comes from net interest earnings. The success of Mammoth Bank, is heavily reliant on how many loans it can give out while maintaining low default rates, where default means the inability of the borrowers to pay back the loan in time.  In our report, we work to discuss risk vs rewards among Mammoth’s customer base.  As you may already know, the banking industry handles cash, credit, and other financial transactions as well, providing a safe place to store money for customers' convenience.  Our focus will be on providing one or a few determinates to Mammoth Bank that contribute to a strategy that helps predict those customers who won’t default on a given loan.  As we work to understand these determinants consider the following, bank capital is a measure that appears on the liability side of the bank’s balance sheet.  The greater the capitalization, liquidity risks, poor credit quality, greater cost inefficiency, and banking industry size works to significantly increase the production of non-performing loans; yet, when the bank is more profitable it helps to reduce non-performing loans.  In the best fit of our predictive variables, we will incorporate several indicators and measures, so to best help, our client move forward with those customers best suited.

Summary of Our Work

In this report, initial exploratory data analysis has been performed to help better understand the data- to identify any outliers, correlation, or patterns. We put more emphasis on Age and Employment Duration as we began our data exploration.  Following that, a LASSO logistic regression model was implemented to predict the factors that may or may not contribute to a borrower defaulting on a loan.

Understanding Our Data

There are 1000 observations of historical loan accounts provided in the dataset, with 14 varying factors that is used to assess the loan candidate. Also provided in the dataset is whether or not the loan was defaulted on or not. There are 700 zeros in the response variable, indicating that 700 of the loans given out in this data were not defaulted on. There are 300 ones, indicating that 300 loans were defaulted on.  This is going to what we are trying to limit in future instances and will serve as our response variable.

The other variables include basic information that your bank would know for their loan clients like the term, type, and amount of the loans granted by the bank. It also includes information that is typically used to assess whether or not the customer will be a good borrower like employment status, duration, number of other credit accounts, and credit score. They also have basic customer information included that is usually given once you set up an account like age, marital status, saving and checking account balance, and gender of the borrower.

There are 7 ‘true’ numerical variables in the data provided. There are 8 categorical variables including the variable we want to be able to predict in future occurrences, whether a borrower will default or not. Some categorical variables like the ones concerning loan type have been coded using numbers to make it easier to use to build the predictive model.

Summary Statistics of Numerical Data

The data provided was complete with no missing values in the set. There are some observable outliers in the data provided. Most of the borrowers are between 20 and 40 years of age. There are only a few who do not fit in the age range. Most of the credit scores fell between 500 and 900 points, with a few who scores were higher who defaulted. There is also a borrower who score is below 400, an extreme outlier.  There is about a $2000 spread between the largest and small checking amount balances.

Below is the data provided from the Skimr report as created in R

-- Data Summary ------------------------

                           Values

Name                       loan\_data

Number of rows             1000

Number of columns          15

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Column type frequency:

  character                2

  numeric                  13

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Group variables            None

-- Variable type: character ------------------------------

# A tibble: 2 x 8

  skim\_variable  n\_missing complete\_rate   min   max empty

\* <chr>              <int>         <dbl> <int> <int> <int>

1 Marital\_status         0             1     6     7     0

2 Emp\_status             0             1     8    10     0

  n\_unique whitespace

\*    <int>      <int>

1        2          0

2        2          0

-- Variable type: numeric --------------------------------

# A tibble: 13 x 11

   skim\_variable    n\_missing complete\_rate     mean

 \* <chr>                <int>         <dbl>    <dbl>

 1 Default                  0             1    0.3

 2 Checking\_amount          0             1  362.

 3 Term                     0             1   17.8

 4 Credit\_score             0             1  760.

 5 Car\_loan                 0             1    0.353

 6 Personal\_loan            0             1    0.474

 7 Home\_loan                0             1    0.056

 8 Education\_loan           0             1    0.112

 9 Amount                   0             1 1219.

10 Saving\_amount            0             1 3179.

11 Emp\_duration             0             1   49.4

12 Age                      0             1   31.2

13 No\_of\_credit\_acc         0             1    2.55

        sd    p0   p25   p50   p75  p100 hist

 \*   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>

 1   0.458     0    0     0     1      1 ▇▁▁▁▃

 2 301.     -665  165.  352.  554.  1319 ▁▃▇▃▁

 3   3.24      9   16    18    20     27 ▁▆▇▆▁

 4  77.6     376  726.  770.  812   1029 ▁▁▆▇▁

 5   0.478     0    0     0     1      1 ▇▁▁▁▅

 6   0.500     0    0     0     1      1 ▇▁▁▁▇

 7   0.230     0    0     0     0      1 ▇▁▁▁▁

 8   0.316     0    0     0     0      1 ▇▁▁▁▁

 9 306.      244 1016  1226. 1420.  2362 ▁▅▇▂▁

10 340.     2082 2951  3203  3402.  4108 ▁▃▇▆▁

11  37.8       0   15    41    85    120 ▇▅▂▃▃

12   4.09     18   29    32    34     42 ▁▃▇▆▁

13   1.65      1    1     2     3      9 ▇▃▂▁▁

Default    Checking\_amount       Term        Credit\_score    Marital\_status

 Min.   :0.0   Min.   :-665.0   Min.   : 9.00   Min.   : 376.0   Length:1000

 1st Qu.:0.0   1st Qu.: 164.8   1st Qu.:16.00   1st Qu.: 725.8   Class :character

 Median :0.0   Median : 351.5   Median :18.00   Median : 770.5   Mode  :character

 Mean   :0.3   Mean   : 362.4   Mean   :17.82   Mean   : 760.5

 3rd Qu.:1.0   3rd Qu.: 553.5   3rd Qu.:20.00   3rd Qu.: 812.0

 Max.   :1.0   Max.   :1319.0   Max.   :27.00   Max.   :1029.0

    Car\_loan     Personal\_loan     Home\_loan     Education\_loan   Emp\_status

 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Length:1000

 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   Class :character

 Median :0.000   Median :0.000   Median :0.000   Median :0.000   Mode  :character

 Mean   :0.353   Mean   :0.474   Mean   :0.056   Mean   :0.112

 3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.000

 Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000

     Amount     Saving\_amount   Emp\_duration         Age        No\_of\_credit\_acc

 Min.   : 244   Min.   :2082   Min.   :  0.00   Min.   :18.00   Min.   :1.000

 1st Qu.:1016   1st Qu.:2951   1st Qu.: 15.00   1st Qu.:29.00   1st Qu.:1.000

 Median :1226   Median :3203   Median : 41.00   Median :32.00   Median :2.000

 Mean   :1219   Mean   :3179   Mean   : 49.39   Mean   :31.21   Mean   :2.546

 3rd Qu.:1420   3rd Qu.:3402   3rd Qu.: 85.00   3rd Qu.:34.00   3rd Qu.:3.000

 Max.   :2362   Max.   :4108   Max.   :120.00   Max.   :42.00   Max.   :9.000   

Default Checking\_amount  Term Credit\_score Marital\_status Car\_loan Personal\_loan Home\_loan

    <int>           <int> <int>        <int> <chr>             <int>         <int>     <int>

1       0             988    15          796 Single                1             0         0

2       0             458    15          813 Single                1             0         0

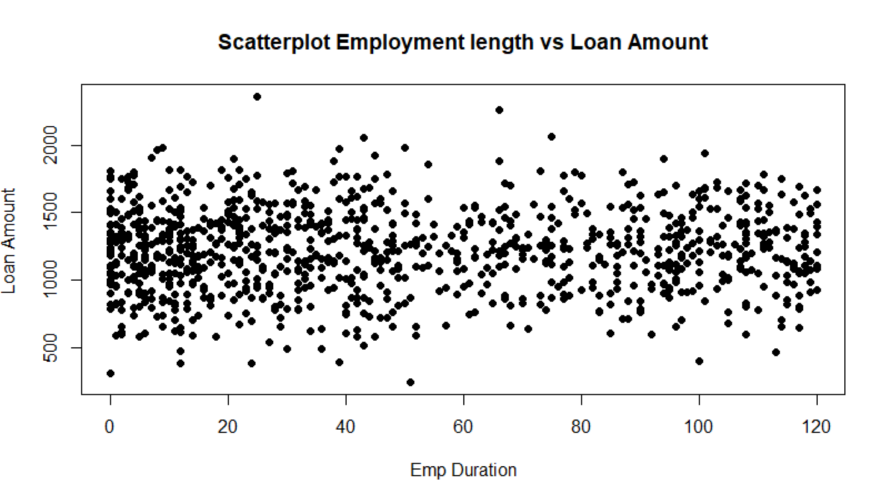
3       0             158    14          756 Single                0             1         0

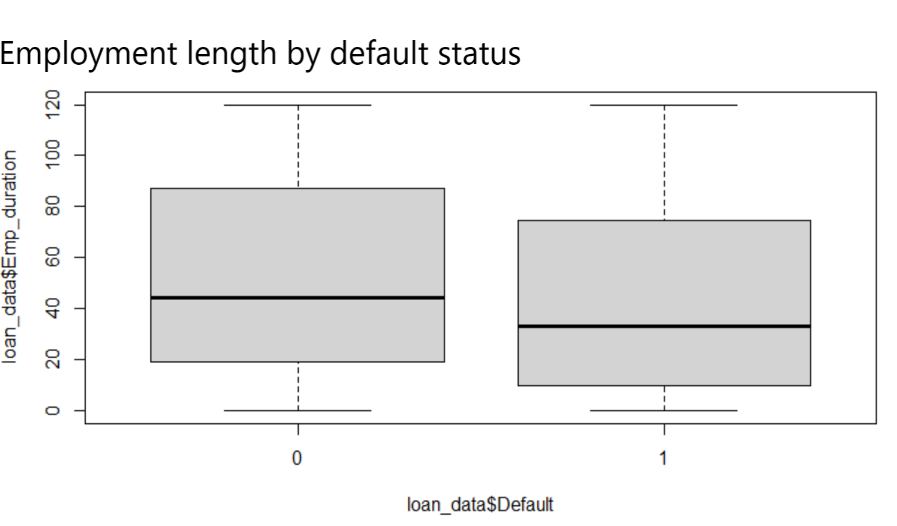
4       1             300    25          737 Single                0             0         0

5       1              63    24          662 Single                0             0         0

6       0            1071    20          828 Married               1             0         0

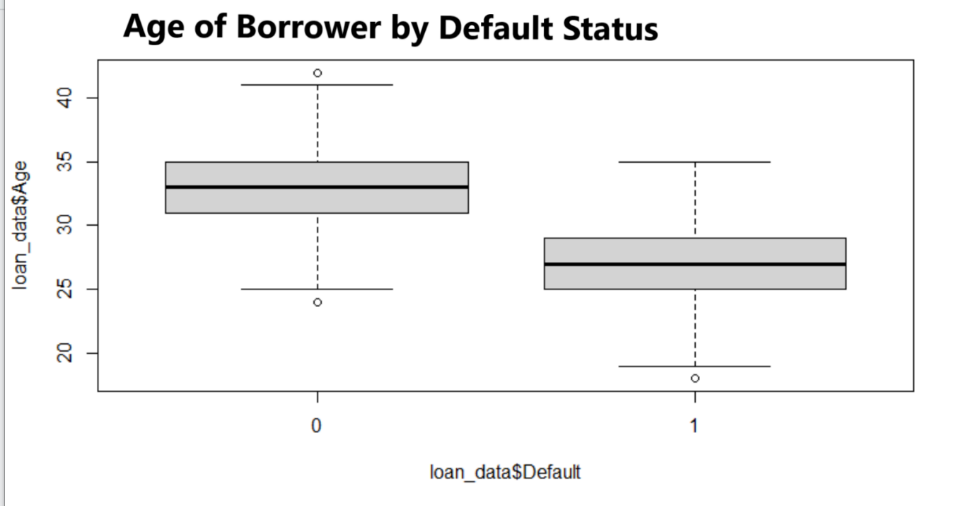
The most frequent length of employment from borrowers is 0 months. The median value is 41 months but the average is 49 months. The median is skewed slightly down from the mean due to the heavy amount, a quarter of all borrowers in this data who have less than 25 months experience on their current job. There is also someone in the set that has 120 months on their job. Borrowers who default tend to have less tenure in their role. The median employment duration of those who default is less than 40 months and the median duration is longer than 40 months, closer to 45 months for those who did not default. This is also evident in the graphs below and the summary statistics provided above. There is no general trend between how much is borrowed and the length of time that the borrower has worked, though we thought there might be.

 <- **Scatterplot of Employment Vs Loan Amount:** The most frequent length of employment from borrowers is 0 months. There is no general trend between how much is borrowed and the length of time that the borrower has worked, though we thought there might be.

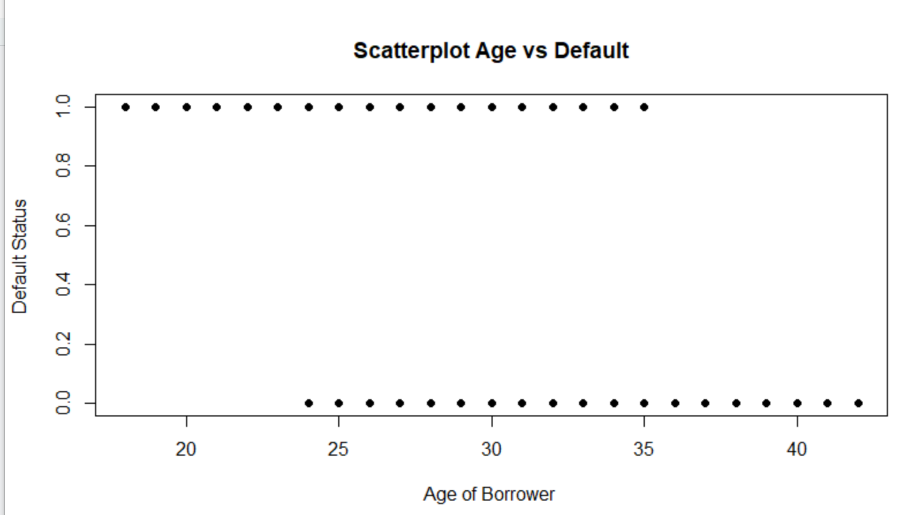


**Box Plot of Employment length by loans defaulted on (1) or not (0):** Borrowers who default tend to have less tenure in their role. The median employment duration of those who default is less than 40 months and the median duration is longer than 40 months, closer to 45 months for those who did not default. ->

The most frequent age to borrow from Mammoth Bank is 32 years old. The median age is also 32 years old. and the average age of borrowers is 31. The age of those who default on their loan tend to be lower, indicating younger loan candidates may present a higher risk. However, the median age of those who default on their loan is closer to 27. The median age of those who do not default on their loan is more closely aligned with the dataset’s average.  There are a few outliers present in the age data. Most of the borrowers are between 20 and 40 years of age. There is a borrower who defaulted that is 18 years old and who did not that is 42 years of age. This is also evident in the graphs below and the summary statistics provided above.



**Box Plot of Age of Borrowers by Default Status:**The median age is also 32 years old. and the average age of borrowers is 31. <-



**Scatterplot Age vs Default Status of the Loan:** The age of those who default on their loan tend to be lower, indicating younger loan candidates may present a higher risk. ->

There is more instances of young borrower, 25 and younger, who have not been at the job long who have defaulted than other age groups. The age of borrowers is about 35 years old for most borrowers who have not worked long in their role in this category of loans. The younger borrower who did not default usually have longer tenures at their current role than their counterparts in the defaulted loan group.

**Chart, bar chart

Description automatically generatedHistogram displaying Age and Employment Length by month for defaulted loans:** There is more instances of borrowers, who have not been at the job long, 25 months or less, who have defaulted than other age groups.

**Chart, bar chart

Description automatically generatedHistogram displaying Age and Employment Length by month for non-defaulted loans:** The age of borrowers is about 35 years old for most borrowers who have not worked long in their role in this category of loans. The younger borrower who did not default usually have longer tenures at their current role than their counterparts in the defaulted loan group.

Modeling and Results

We use training data to build the potential models. We then use cross validation to determine which potential model would be the best choice by looking at the average performance on the training data. Then, we use testing data to evaluate the performance of the best model. Our goal is the build a model that can make predictions, given the indicated information to make a decision about whether or not a loan should be given to reduce the likelihood of future borrowers from defaulting. The testing set allows us to see how accurately our model can predict prior to trying it out on a brand-new borrower.

For this analysis, we used a LASSO logistic regression model with k-fold cross validation. Our measure of the effectiveness of our model is 0.975. This is good since an acceptable measurement is around 0.7.

|  |  |
| --- | --- |
| Method | Testing set AUC |
| LASSO Logistic regression | 0.9754497 |

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

**Chart, scatter chart

Description automatically generated**

**Plot of Receiver Operations Characteristic Curve for Loan Default Prediction Model:** Our goal is to provide the best point on the curve to help the bank choose the best borrower in the future by focusing on where the true positive rate is high and the false positive rate is low. We want to be right about who we say will default and have low rates of incorrectly labeling of those wouldn’t default default.

**Chart

Description automatically generated LASSO Logistic Regression Variable Importance Plot:** The most important variables in determining whether future loan applicants present a risk of defaulting on their loan are if the loan is loan type and age of the borrower.

Recommendations for the client

Whether or not a loan was an educational loan or not was the strongest factor predicting default, and our model has a high measure of effectiveness. Therefore, we can say with confidence borrowers of education loans are more likely to default than other types of loans. Whether or not a loan was for a home was a strong predictor in who would not default on the loan. Therefore, we can say that the institution can continue loaning money for homes without too much fear of default.  The client should also look at the age of the applicant as a determining factor for whether they should lend or not. Younger borrowers present more risk than older borrowers.