# MSBA 635 Individual Consulting Assignment 2 - By – Samruddhi Dhoble.

**CASE 1**

As an analytics consultant, you are tasked with 2 projects:

1. Flight delays are not only costly for airlines but also increase frustration among travelers. Therefore, reduction in flight delays results in decreasing airline operation costs and provides quality and timely services to the passengers. Provide FlightInc with a strategy to help prevent flight delays.

**Report:**

1. In interest of reduction in flight delays and provides quality and timely services to the passengers, business problem is to provide prediction and strategy to prevent flight delays.

In this report initially exploratory data analysis was done to find any missing, outliers or inconsistent data. Further different models were used over the training data to find which gives best prediction with less average square error. Finally, LASSO model was use to predict accuracy of the model.

1. Data Contents:
2. To predict the delay of flight there are 3593 observation and 10 predictors.
3. Predictor variables are independent variables that can be related to changes in response variable. For given data set there are 10 predictors and 1 response variable. These are used to determine the flight delays.
4. Exploratory Data Analysis findings:

i.) As per the summary table below there are **no missing values**. There are 4 carriers (EV,B6,EV,WN)

with **inconsistent** data for predictor Cleaning\_o. There are **outliers observed** for Number\_of\_flights, Support\_Crew\_Available etc.

The numeric values are in negative which depict that there might be no consideration delay for cleaning the flight. Another observation is for predictor weather. As per the observations values are delay due to weather condition ranked 0-10, with 0 being mild and 10 being extreme, however entire dataset has resulted values for delays are either 5 or 6. As weather is uncontrollable and unsteady condition how is the data consistent with value 5 or 6.

**ii.) Summary Statistics:**

Text, table

Description automatically generated

As per the statistics mean of Flight Arr\_Delay in minutes is 69.79 where minimum delay time is 0 mins and maximum delay time is 180 minutes.

**iii.) Histograms and density plots.**

* Following are Histograms and density plots for fueling\_o and Number of flights.

1. It is observed that fueling\_o is normally distributed hence skewedness is 0. Mean=Median.
2. Number of flights is slightly left skewed as Median is greater that mean by 113.09.

Table

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

* Histogram and box plot:

1. Blow chart shows that there are outliers for number of flights. As shown in summary statistics, the minimum count of number of flights is 29475 and maximum value is 53461. However, the box plot shows that there are many values which are out of the maximum and minimum range

Chart, histogram

Description automatically generated

1. Below chart displays the bar and box plot with no outliers which has minimum value of Delay due to weather condition ranked is 5 and maximum value is 6.

Graphical user interface, application

Description automatically generated

* **Build box plots for some variables by category:**

Carriers is categorical data with 14 carriers. Category with respect to flight delays caused due to support \_crew\_available, there are outliers for the category AA, B6, DL, EV, MQ, UA and WN.

**Similarly**, for arr\_delays, the data has different numbers of outliers for AA, B6, DL, EV, MQ, UA and WN.

**Chart, box and whisker chart

Description automatically generated**

**Fig: 2**

**Chart, box and whisker chart

Description automatically generated**

* **Scatter plots:**

1. The scatter plot displays that as the distance between the airports in miles increase the delay time for flights is increasing except for the few observation. For example the when distance is 475 then arr\_delay is highest

Chart, scatter chart

Description automatically generated

1. Delay due to weather condition more (6 them Arr)\_Delay is more.

Chart

Description automatically generated

1. When number of crews increases the arr\_Delay time decreases. Chart, scatter chart

   Description automatically generated
2. As Baggage\_loading\_time increases the Arr\_Delay also increases.

Chart

Description automatically generated

1. As per the plot when time in minutes for security checking in range of 25 to 55, the Arr\_delay is more.

Chart, scatter chart

Description automatically generated

**Modeling and results:**

1. **Generalization Approach**

Here training set is used to Estimates regression coefficients. Further, to decide best when to terminate the predictor variable selection process, validation data is used. Finally, to test our model how it fits well on new data, testing data is used

LASSO or subset selection techniques used as opposed to a liner regression model fit with all predictor variables because with many predictor variables—especially when the number of observations is close to the number of predictor variables analysis will have high variance and overfitting occur. Which means the model will ideally work well for history data but not for new data. For more interpretable model LASSO could be used as modern approach which will select best subset of all predictors to build the model.

**Model Fitting Linear regression:**

**Table

Description automatically generated**

Model has considered all predictors. With Average square error as 151.43390

**Table

Description automatically generated**

**Model Fitting lasso:**

Out of 10 predictors, LASSO choose 7 predictors and stopped the selection.

**Table

Description automatically generated**

**Table

Description automatically generated**

**Model Fitting forward stepwise:**

**Table

Description automatically generated**

**Table

Description automatically generated**

**Model Fitting backward stepwise:**

**Table

Description automatically generated**

**Table

Description automatically generated**

**Interpret results :**

I will choose LASSO for the prediction because as per model fitting the average square error is lest when the model used is LASSO as compared to others. (ASE shown in the table.). With lesser ASE the model will work better on the unseen data.

|  |  |
| --- | --- |
| Method | Testing set ASE |
| Linear Regression  (all predictor variables) | 151.43390 |
| LASSO | 150.96213 |
| Forward stepwise selection | 151.66450 |
| Backward stepwise selection | 151.66450 |

**Coefficient table estimation from LASSO as best model:**

Table

Description automatically generated

Table

Description automatically generated

**Recommendation**:

Considering data provided with 10 predictors we could say that there is relationship between the response variable and predictors. As per statistical analysis the R- square value is 81% which means that our model fits 81% well for testing data.

Analysis predicts that on change of predictors such as airport\_distance, weather, etc can influence the delay in the flights. But are these predictors enough to summarize the prediction?

It is better to know more about

* When was this data collected (year)? Is this data relevant to the current situation?
* Are these 11 characteristics enough to consider for testing and training data. For example, delay in flight timings could be due to technical issues in flight. How was the data collected? As these customer’s complaints? If yes, then are all of these complains true?

Hence, our model works well for given training and testing data for the given year, however, to know better accuracy of model we should have more data collected for different time (year)and more characteristics.

**CASE 2**

In spite of non-interest incomes on the rise, over half the money made by banks still comes from net interest earnings. A bank’s success heavily relies upon 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. Provide Mammoth Bank with a strategy to predict customers who default or not.

Mammoth Bank Consulting Problem

1. **Introduction (1/2 paragraph)** 
   1. **Purpose**

Build the model which will predict the bank customers who are default or not i.e., to predict the inability of the borrowers to pay back the loan in time.

* 1. In this report, initial exploratory data analysis has been performed and then data was partitioned in training and testing data. Found the coefficient for the variables. Train or fit the LASSO Logistic Regression Model, get prediction using testing set data. Finally, evaluate model performance by finding area under the curve.

1. **Exploratory data analysis (1 paragraph)**
   1. **Data Contents** 
      * 1. There is total 1000 observations and 15 predictors.
        2. Predictor variables are independent variables that can be related to changes in response variable. For given data set predictors are used to determine inability of the borrowers to pay back the loan in time.
   2. **Explain Exploratory Data Analysis findings (4-5 sentences)**

**Summary Statistics:**

Table

Description automatically generated

* + 1. Summary Statistics:

As per data exploration there are no missing values. There are outliers for predictors checking\_amount, Term, credit\_score, Home loan, Education loan, Amount, saving amount, age and number\_of\_credit\_acc. And credit score has inconsistencies too. For example, the maximum credit score possible is 850, but there are 82 observations in the dataset with more than 850 as the credit score.

**The summary from the skim function in R.**

A picture containing table

Description automatically generated

* + - 1. There are 2 categorical and 13 numeric data.
      2. Discuss 1-2 variables in depth—describe histogram, mean vs median etc.

**Variable Term:**

Text

Description automatically generated with medium confidence

For predictor TERM median is greater than mean, hence the histogram is slightly left skewed. Below is the histogram:

Chart, histogram

Description automatically generated

**Variable Emp\_Duration**:

Text

Description automatically generated with low confidence

Since mean is greater than median, the bar graph is right skewed.

Chart, bar chart, histogram

Description automatically generated

1. Box plot for Age over Default. Below plot has an outlier.

boxplot(credit\_data$Age~credit\_data$Default)

A picture containing chart

Description automatically generated

* + 1. There are total 700 0’s and 300 1’s in Default.

1. **1-2 paragraphs**
2. **Modeling and results**
   1. Generalization Approach (3-4 sentences)
      1. Using 80% of the data as training data with 5 fold cross validation, 12 seed and 20% of the data as testing data. Following in the selection summary:

Table

Description automatically generated

Here 800 observations are used as training set to Estimates regression coefficients. Further, to decide best when to terminate the predictor variable selection process, validation data is used. Finally, to test our model how it fits well on new data, testing data is used.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

Split the training data into k roughly equal size parts folds (or more), Choose the 𝜆 with the best average performance on the k validation sets. Refit all training data with the model that has the best 𝜆𝜆 . Evaluate final performance on the test set using the model refit on all training data with the model that has the best

* 1. **Model (1 sentence)**
     1. **Fit a LASSO logistic regression model to the data.**

Graphical user interface, application

Description automatically generated

* 1. Interpret results (4 sentences or more)
     1. Provide the ROC curve for the testing set for each model, and explain it fill out the following table.

Chart, histogram

Description automatically generated

|  |  |
| --- | --- |
| Method | Testing set AUC |
| LASSO Logistic regression | 0.9804762 |

Table

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Point in Fig 4 | False Positive | True Positive | Cutoff |
| A | 10.71% | 100.00% | 0.2093 |
| B | 2.14% | 88.33% | 0.6110 |

Using a cutoff value of 0.2093, Mammoth Bank can correctly identify 100% of the customers who will default on their loan. However, the model will also incorrectly identify over 10% of the customers who would likely pay back their loan as “will default,” resulting in them being disapproved for a loan and the bank losing that revenue.

By increasing the cutoff value to 0.6110, customers who should be approved for a loan but misidentified reduces to 2%, a reduction of 8.5%. HOWEVER, 12% of defaulters will slip through.

**Recommendation:**

Area under the curve is 98%. Which is close enough to prove that the model fits to work for the new data.

Mammoth Bank leadership can decide to use one of the points above, based on their desire to identify customers who will default, at the risk of missing an opportunity to approve other customers. If they are more concerned with losing potential customers, we will adjust the cutoff value to 0.6110. If the Mammoth Bank leadership is more concerned with not approving loans to customers who will likely default, we will adjust the cutoff value to 0.2093.

**Areas for further analysis**

One question that should be answered is, “Is it more important to miss an opportunity for a loan (False Positive) or lose money on a loan that is defaulted on?”