# **ASSIGNMENT COVER PAGE**

A picture containing graphical user interface

Description automatically generated

**J.E. CAIRNES SCHOOL OF BUSINESS & ECONOMICS**

**GROUP ASSIGNMENT**

**Module Name and Code: Business Modelling and Analytics – MS5107**

|  |  |  |
| --- | --- | --- |
| Group Number | Member Name | Student ID |
| 16 | Revanth Gengaiah | 22210054 |
| Vigneshwaran Raja | 22224111 |
| Lokeshwaran Ramaiah | 22225782 |
| Hemanth Kumar Ravilla | 22223108 |
| Dilip Venkatesan Sankar | 22225743 |

In submitting this assignment, we are aware that it is our responsibility to adhere to the submission guidelines. Please tick (double click->… or Y/N) for the following:

|  |  |  |
| --- | --- | --- |
|  | **Yes** | **No** |
| We are aware of what the NUI Galway plagiarism policy entails. |  |  |
| We have named the assignment file (MS Word docx , doc, xlsx) to contain our group number, module code, and assignment number (e.g., group\_4\_MS5107\_A2.docx or group\_4\_MS5107\_A2.xlsx). |  |  |

**Declaration for this Assignment Submission*:***

*In submitting this work we confirm that it is entirely our own. We acknowledge that we may be invited to interview if there is any concern in relation to the integrity, and we are aware that any breach will be subject to the University's Procedures for dealing with plagiarism (*[*http://www.nuigalway.ie/plagiarism*](http://www.nuigalway.ie/plagiarism) *).*

Table of Contents

**Section A**

1. Summary report for model building process and model selection

1.1 Overview

1.2 Business understanding

1.3 Data understanding

1.4 Data Preparation.

* Sampling
* Detecting NULL and duplicates
* Identifying outliers
* Variable handling – Creating Dummies
* Data partitioning and performance estimation strategies

1.5 Model Building

* Linear Regression
* Regression Trees

1. Fully Grown
2. Best Pruned
3. Minimum Error

* Ensemble Methods

1. Bagging
2. Boosting
3. Random Forest
4. Prediction of Airfare with custom values with SW=No
5. Prediction of Airfare with custom values with SW=Yes

**Section B**

1. Assumptions - Predicting Airfare based on predictor variables
2. Model building with predictor variables
3. Airfare prediction based on available variables
4. Performance and model comparison

## **Section A. Using your knowledge of Business Modelling and Analytics, build a model that predicts average fare on a new route:**

1**. Prepare a summary report to describe the model-building process and why you believe that your model is good.**

**Overview:**

In order to anticipate the optimal airfare for the route, we will research and analyze a dataset of air routes for this assignment.

To ensure the dataset is appropriate and full for our study, we must take a few steps before diving into the issue,

1. Data mining

* We will use Cross-industry standard process for data mining
* As part of this procedure, we must examine and pre-process the data to look for duplicates, null values, and outliers.
* We will focus on data rescaling, partitioning, and creating dummies for any categorical data
* We will be extracting useful information to find patterns, meaningful correlations

1. Model building

Once the data has been prepared and pre-processed, we will utilize the following approaches to forecast the cost of flying and select the most effective method based on the best outcome,

* Linear regression
* Regression trees
* Ensemble method

**Business understanding:**

A significant airline gathered data on 638 American air routes to price flights on these routes and the spreadsheet contains recent data related to the air routes. The distance travelled, the city's demographics, the destination city name, and if this location is a tourist destination are some aspects of these new routes that are known but we also have some unknown factors like the number of passengers who will travel, if Southwest airlines are planning to fly on these new routes. Southwest's approach differs significantly from the method used by the larger, more established airlines.

**Data understanding:**

The spreadsheet contains 639 rows including header and 18 columns which are explained in detail below,

|  |  |
| --- | --- |
| Data Column | Meaning |
| **S\_CODE** | Code of Starting Airport (Source) |
| **S\_CITY** | Starting City name |
| **E\_CODE** | Code of Ending Airport (Destination) |
| **E\_CITY** | Ending City name |
| **COUPON** | Average Coupon provided for that flight |
|  | Ø  One coupon indicated a non-stop flight |
|  | Ø  Two coupons indicated a one-stop flight |
| **NEW** | Total number of new flights |
| **VACATION** | Indicating the below |
|  | Ø  Yes – The fights take a vacation route |
|  | Ø  No – The flight does not take a vacation route |

|  |  |
| --- | --- |
| Data Column | Meaning |
| **SW** | Indicating the below |
|  | Ø  Yes – Southwest airline serves the route |
|  | Ø  No – Southwest airline does not serve the route |
| **HI** | Herfindal Index – the measure of market concentration |
| **S\_INCOME** | Average income of a person in the starting (source) City |
| **E\_INCOME** | Average income of a person in the ending (destination) city |
| **S\_POP** | Population of the starting (source) City |
| **E\_POP** | Population of ending (destination) city |
| **SLOT** | Indicates the below |
|  | Ø  Controlled – The destination airport is slot controlled |
|  | Ø  Free – Destination airport is no slot controlled |
| **GATE** | Indicated the below |
|  | Ø  Free - Destination disport does not have gate constraint |
|  | Ø  Constrained - Destination airport has gate constraint |
| **DISTANCE** | Distance between two endpoint airports in miles |
| **PAX** | Total Number of passengers present on the airline flying route during the period of data collection |
| **FARE** | The average fare on the route |

**Data preparation:**

In this process, we will use selection criteria to clean, construct, integrate, transform and re-scale the dataset.

Below are the steps involved in data preparation,

* **Pre-processing**

1. **Sampling:**

There is no need for sampling based on the provided dataset because there are a finite number of rows. When the data has more than 1000 rows, sampling of the data is an option.

1. **Detecting NULL and Duplicate values:**

Excel's algorithms cannot operate on null values, and our dataset contains no NULL values. However, we did find four duplicates, which were ignored.

***Fig 2.1 Null Values Fig 2.2 Duplicates values***

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

1. **Identifying outliers:**

In statistics, an outlier is a data point that differs significantly from other observations. There are a couple of steps to be performed to identify the outliers,

* Mean of all numerical data
* Standard Deviation of all numerical data

Once we have the mean and standard deviation, we are using the below formula to check if any of the values are outliers are not,

**[mean - 3\*stdev; mean + 3\*stdev]**

In excel, we will use the below formula based on our dataset,

=IF(OR((E2<E$641-3\*E$642),(E2>E$641+3\*E$642)),"outlier","okay")

Where E2 is a column name (**E.g., COUPON**) and E641 and E642 are the mean and standard deviation respectively where $ represent fixed column

In our dataset, we have a couple of outliers, and those we were not removed as it was not very significantly different and was not too far away from the present range.

1. **Variable handling - Creating dummies**

To make the values compatible with Excel, we have converted the categorical variables to numerical data,

* VACATION – Yes/No where Yes=1 and No=0
* SW – Yes/No where Yes=1 and No=0
* SLOT – Free/Controlled where Free=1 and Controlled-0
* GATE – Free/Constrained where Free = 1 and constrained=0

And to avoid muti-collinearity and dummy variable trap, we have removed a few columns,

* VACATION\_No
* SW\_No
* SLOT\_Controlled
* GATE\_Constrained

***Fig 4.1 Creating dummies***

Graphical user interface, application

Description automatically generated

The details can be found in the **STDPartition** sheet in Excel.

* **Data Partitioning and Performance estimation strategies**

In this process, we have segregated the data into,

1. Training – Contains training partition of 380 records
2. Validation – Contains validation partition of 254 records

**Fig 4.2 Data Partitioning**

Graphical user interface, application

Description automatically generated

|  |  |  |
| --- | --- | --- |
| **Partition Summary** | | |
|  |  |  |
|  | **Partition** | **# Records** |
|  | **Training** | 380 |
|  | **Validation** | 254 |

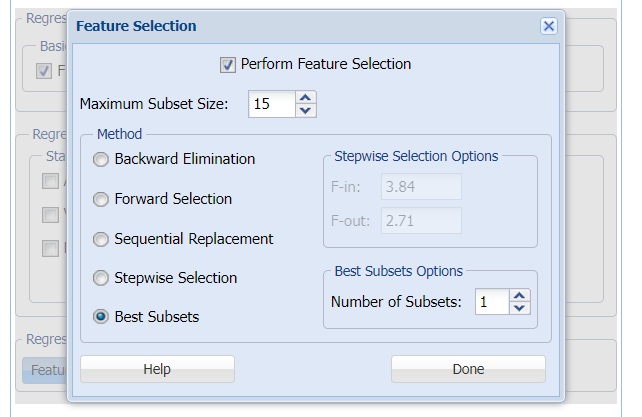
**Model Building**

**Linear Regression**

A predictive analysis method called linear regression is used to fit a linear function to a given dataset. This method makes use of independent factors to forecast the value of a dependent variable. Simple linear regression occurs when there is just one independent variable; multiple linear regression occurs when there are numerous independent variables. By fitting a linear equation to the observed data, linear regression models establish a link between dependent and independent variables,

**Y = b0 + b1X1 + b2X2 +…. + bnXn + Ꜫ**

***Fig 1.5.1 Linear Regression Fig 1.5.2 Best Subsets***

Graphical user interface, application, Word

Description automatically generated

We will select every value for the Selected variables with the exception of Record ID, which we will put in the Output variable in order to estimate the airfare.

After the Linear regression is processed we will see a couple of sheets created,

***Fig 1.5.3 Datasheet of STDPartition***

Table

Description automatically generatedThe workbook name and the total number of records in the training data and validation data are both listed on the LinReg Output sheet.

However, as the model is projected based on its value, this section will concentrate more on the coefficients table.

The model is judged on the below metric values in the LinReg\_ValidationScore sheet,

1. Sum of squared error
2. Mean squared error
3. Root mean square error
4. Mean absolute deviation

***Fig 1.5.4 Metric values***

Table

Description automatically generated We'll concentrate on the RMSE and MAD in order to predict the value. The model with lower values is considered to be superior because RMSE and MAD values reflect prediction accuracy. And we have taken RMSE=35.56 and MAD=28.09 from the Linear\_regression\_Validation sheet.

Based on the above findings, a few variables will be removed in order to see if the model's performance improves. And to do this, we'll run the linear regression model again with different feature selection options, then use the Best Subset approach to make the model better.

***Fig 1.5.5 Subset selection option***

Table

Description automatically generated

Based on the details we will consider the following details to choose the subset,

1. **RSS** (residual sum of squares) - the lower the better subset is.
2. **Mallow’s Cp** –Adequate subsets are those with Cp roughly equal to the number of the subset variables and/or Cp is at a minimum. So, in our case we can eliminate most of the subsets from 1 to 10 since the Cp value is greater than 14.
3. **R2** / **Adj**. **R2** - goodness-of-fit (how well the model fits to the data). The higher the better subset is.
4. **Probability** – hypothesis test if a subset is acceptable. If Probability<0.05, rule out that subset.

So, will use Subset 11, 12, and 13 as there are satisfying the conditions. The linear regression will be performed again for the three subsets by eliminating respective values,

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| We choose Subset 12 as our model | | | |  |  |  |  |
| subset 12 |  |  | subset 13 |  |  | subset 11 |  |
| **Metric** | **Value** |  | **Metric** | **Value** |  | **Metric** | **Value** |
| SSE | 320350.7 |  | SSE | 321756.3 |  | SSE | 322754.8 |
| MSE | 1261.223 |  | MSE | 1266.757 |  | MSE | 1270.688 |
| RMSE | 35.51371 |  | RMSE | 35.59153 |  | RMSE | 35.64671 |
| MAD | 28.04926 |  | MAD | 28.12332 |  | MAD | 28.22579 |
| R2 | 0.781502 |  | R2 | 0.780543 |  | R2 | 0.779862 |

We can see that RMSE = 35.51 and MAD = 28.09 and R2 = 0.28 for subset12 with twelve variables is the best model compared to the other subsets.

Note: We have considered all the values except Record ID, S\_CODE, E\_CODE, S\_CITY, and E\_CITY as they do not provide any significant change in building the model for prediction.

**Regression Trees**

The next model we are using to predict airfare is the Regression tree. Regression tree analysis is when the predicted outcome can be considered real. Two of the strengths of this method are on the one hand the simple graphical representation by trees, and on the other hand, the compact format of the natural language rules we are using this since it is continuous data, and it is a prediction model.

We will be using the below trees for our prediction under Tree for scoring,

1. Fully grown
2. Best pruned
3. Minimum error

***Fig 1.5.6 Regression selection option***

Graphical user interface, application, Word

Description automatically generated Graphical user interface, application, Word

Description automatically generatedGraphical user interface, application, Word

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

On running the regression, we were able to obtain the RMSE And MAD based on which we will decide the significant predictor,

***Fig 1.5.7 RMSE and MAD values for regression trees***

Table

Description automatically generated

Based on the above even though for fully gown RMSE = 36.31 AND MAD = 26.68, Best pruned RMSE = 36.60 and MAD = 27.25 but we can see the RMSE is the lowest for the Min error tree RMSE = 36.21 and MAD = 26.76. Even though Best pruned has Nodes = 57 and height = 7 which is less compared to the Min error tree with nodes = 65 and tree height = 7, we are taking the minimum error tree as the best model taking RMSE and MAD value into consideration.

***Fig 1.5.8 Minimum error tree chart***

Chart

Description automatically generated

And the Prediction score for regression tree with SW = NO and SW =YES are given below respectively,

***Fig 1.5.9 Regression tree scoring value***

 Table, timeline

Description automatically generated

 Timeline

Description automatically generated

**Ensemble methods**

By mixing numerous models rather than relying just on one, ensemble approaches seek to increase the accuracy of findings in models. The integrated models considerably improve the results' accuracy.

Using the same learning algorithm, it creates multiple weak models and combines them together to form a strong model.

They are classified into three types,

1. Bagging
2. Boosting
3. Random forests

The above methods can be run using different algorithms like Decision trees, neural networks, K-Nearest Neighbours, and Linear regression. We will be using the Decision trees algorithm for all the methods to predict the best airfare. After the data is pre-processed (removing duplicates, identifying outliers, creating dummies, rescaling, and creating partitions) we will start performing the operations,

***Fig 1.5.10 Ensemble method***

Graphical user interface, application

Description automatically generated

***Fig 1.5.11 Ensemble method using Decision tree***

Graphical user interface, text, application, email

Description automatically generated

We will be calculating the RMSE and MAD values for all three methods and based on the values we will choose the suitable method,

***Fig 1.5.12 Method selection***

Graphical user interface, application

Description automatically generated

* **Bagging:**

Bootstrap aggregating is commonly used in classification and regression, and also. It creates several datasets from the original dataset using random sampling. Through decision trees, it improves the models' accuracy, greatly reducing variance.

We will perform the Ensemble bagging on the STDPartition sheet with rescaling as normalization,

Table

Description automatically generated

In the bagging process, we get RMSE = 33.7668 and MAD = 25.031. Bagging is helpful because it creates one strong learner from several weak base learners, which is more stable than individual learners. Additionally, it gets rid of any variance, which lessens overfitting in models. The computational cost of bagging is one of its drawbacks.

* **Boosting:**

Boosting is an ensemble strategy that improves future predictions by learning from previous predictor errors. The method greatly increases model predictability by combining numerous weak base learners into one strong learner i.e., it produced a strong model from a weak model by focusing on misclassified records. The prediction technique produces the below values,

Table

Description automatically generated

In the boosting process, we get RMSE = 26.68 and MAD = 20.32. Boosting works by placing weak learners in sequential order so that they can learn from the subsequent learner to improve their predictive models. But the only drawback of this method relies heavily on the computation speed and the performance of the model.

* **Random forest:**

It is a variation of the bagging process which is the most accurate algorithm currently available, and it can process very big datasets. Similar to the Bagging approach, this trains numerous separate decision trees simultaneously.

Based on the dataset, we can see the below values obtained after the validation in the random trees prediction model,

Table

Description automatically generated

In the Random Forest process, we get RMSE = 33.71 and MAD = 24.94 which seems to be comparatively more than the Bagging and Boosting prediction methods.

Based on all the models, we can see that the Ensemble Boosting prediction method using a decision tree produces the best results compared to the others which are represented in the table below,

***Fig 1.5.13 Overall values for all models***

|  |  |  |  |
| --- | --- | --- | --- |
| No | Modeling technique | RMSE | MAD |
| 1 | Linear Regression | 35.51 | 28.04 |
| 2 | Regression Trees (Min error) | 36.00 | 26.62 |
| 3 | Bagging method | 33.76 | 25.03 |
| 4 | Boosting method | 26.68 | 20.31 |
| 5 | Random forest | 33.71 | 24.94 |

### **2. Using the model, predict the average fare on a route with the following characteristics: COUPON=1.202, NEW=3, VACATION=No, SW=No, HI=4442.141, S\_INCOME = $28760, E\_INCOME=$27664, S\_POP=4557004, E\_POP=3195503, SLOT=Free, GATE=Free, PAX=12782, DISTANCE=1976 miles.**

**Prediction of Airfare with custom values with SW=No**

The Ensemble boosting approach is the best model for this investigation, according to the aforementioned analysis. In the event that Southwest Airlines does not offer service on the route, we will forecast the airfare using custom numbers (SW=No),

Timeline

Description automatically generated

Based on the scoring method we have predicted the airfare as 277.89

### **3. Predict the reduction in average fare on the above route if Southwest Airlines decides to cover this route.**

**Prediction of Airfare with custom values with SW=Yes**

We need to predict the airfare if Southwest airline decides to cover the route which means similar to question 2, we will use the ensemble boosting method for the prediction where SW=Yes and below are the values used,

COUPON=1.202, NEW=3, VACATION=No, **SW=Yes**, HI=4442.141, S\_INCOME = $28760, E\_INCOME=$27664, S\_POP=4557004, E\_POP=3195503, SLOT=Free, GATE=Free, PAX=12782, DISTANCE=1976 miles and below is the obtained result,

Timeline

Description automatically generated

Based on the scoring method we have predicted the airfare as 255.03

## **Section B. In reality, which of the factors (predictor variables) will not be available for predicting the average fare from a new airport?**

### **1. Briefly comment on your assumptions.**

**Predicting Airfare based on predictor variables**

As the flight is not yet operational, we have a total of 18 variables based on the provided dataset, and we will be discarding a few of them to observe whether there is any increase or decrease in the airfare.

From the initial step, we have not considered **Record ID, S\_CODE, E\_CODE, S\_CITY, and E\_CITY** as they do not provide any significant change in building the model for prediction.

We have eliminated the below,

1. **COUPON** – It is not a valid variable because we cannot predict how many coupons will be given out for each flight because the flight is still not in operation.
2. **NEW** – Due to the airport's ongoing operation, we are unable to predict the number of carriers on each route.
3. **HI** – Before the airport is open for business, we cannot calculate the Herfindel index, which measures market concentration.
4. **PAX** – Until the airport is operational, it is impossible to estimate the number of passengers there.

***Fig 1.1 Partition values for new model prediction***

Graphical user interface, application

Description automatically generated with medium confidence

We have considered the below variables as all these are known and few of them are fixed values,

VACATION, SW, S\_INCOME, E\_INCOME, S\_POP, E\_POP, DISTANCE, FARE, SLOT, FREE, GATE.

Based on these values will create another partition sheet called STDPartition1\_ASSUMPTION using which we will predict the airfare from the best-chosen model which is the Ensemble Boosting from Section A.

### **2. Based on the settings and findings of the model from item A, build another model using the available variables only**

**Model building**

We have chosen the Ensemble boosting prediction method as the suitable method based on Section A analysis as it had low RMSE, and MAD compared to the other prediction method.

Based on the variable elimination, the partition is created, and we have the below columns,

***Fig 2.1 Partition summary with training and validation values***



Upon running the Ensemble booting method using the decision tree, we get RMSE= 34.5 and MAD = 24.05,

Graphical user interface, application

Description automatically generated ***Fig 2.2 Choosing variable for Ensemble Boosting method with RMSE and MAD value***

Table

Description automatically generated

### **3. Use this model to predict the average fare using only the available (in your opinion) data from the record in item A.2.**

**Airfare prediction based on available variables**

We are using the Ensemble boosting method to predict the air fare using scoring technology if southwest airlines decide to operate in the routes where SW=NO,



Before predicting the score, we matched all the variables by name,

Graphical user interface, application

Description automatically generated

And we can the below score,

Timeline

Description automatically generated

### **4. Compare the performance of this model with the performance of the model from item A.** **Is this model good enough, or is it worthwhile re-evaluating the model once flights commence on the new route?**

**Model Comparison and Conclusion**

From the 18 variables present in the given dataset,

* For Model 1 we have excluded 4 variables which are S\_CODE, S\_CITY, E\_CODE, E\_CITY
* For Model 2 we have excluded 8 variables S\_CODE, S\_CITY, E\_CODE, E\_CITY, COUPON, NEW, HI and PAX

All these variables were excluded as they didn’t add any significant different to the prediction and they were unknown factors.

Table

Description automatically generatedModel 1:

With 14 variables, model 1 produced an RMSE = 26.28, MAD = 20.31, and R2 = 0.87 which gives 87 percent accuracy in the prediction

Model 2:

Table

Description automatically generated With 10 variables, model 2 produced an RSME = 34.52, MAD = 24.05, and R2 = 0.79 which given 79 percent accuracy in the prediction

Comparing both the models, we can conclude that Model 1 performed better than Model 2 which has lower RMSE and MAD scores which means Model 1 has lower errors compared to Model 2.

Additionally, Model 1's R2 values are higher than Model 2's, indicating that Model 1's prediction accuracy

In order to avoid influencing the model's ability to make a better forecast, we have evaluated both models with all variables and have run them through their full ranges of possibilities. Therefore, in the event that the airport is operational, and the airline decides to start operations, Model 1 with 14 variables would be used because it had better RMSE, MAD, and R2 values as well as better prediction accuracy than Model 2 with 10 variables.