Ollscoil na Gaillimhe
University of Galway

Scoil Ghnó agus Eacnamaíochta J.E. Cairnes
J.E. Cairnes School of Business and Economics

GROUP ASSIGNMENT COVER PAGE

Module Code	**<u>MS5107</u>**
Group Number	35
Student 1	Akshata Ghumatkar (24240627)
Student 2	Shubham Bhatnagar (24238834)
Student 3	Amit Kumar (24236930)
Student 4	Devraj Raghuwanshi (24232193)
Student 5	Ebad Abdul Rauf Shaikh (24237373)

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**
I am aware of the **University Academic Integrity Policy** https://www.universityofgalway.ie/media/registrar/policiesmay2023/QA220-Academic-Integrity-Policy-v2.0-Sept-2023.pdf and confirm the declaration below.	**Yes☒ Error! Bookmark not defined.**	☐
I have saved the files for submission (e.g., docx, .jar) **following strictly the format and naming required** (e.g., group_4_MS5107_A2.docx, group_4_MS5107_A2.xlsx).	**Yes☒**	☐

Declaration for this Assignment Submission*:*
We hereby declare that the work submitted is entirely our own work. It has not been taken from the work of others, except to the extent that such work has been cited and acknowledged within the text of our work. This work is not done in whole or in part by a machine or through Generative Artificial Intelligence, such as ChatGPT or else. We have not allowed, and will not allow, anyone to copy our work with the intention of passing it off as their own.

2425-MS5107 BUSINESS MODELLING AND ANALYTICS

QUESTION A) BUILD A MODEL THAT PREDICTS AVERAGE FARE ON A NEW ROUTE.

1: DESCRIBE THE MODEL BUILDING PROCESS AND WHY YOU BELIEVE THAT YOUR MODEL IS GOOD.

In the highly competitive industry like airline, introducing new routes open avenues for new sets of challenges and opportunities to exploit. A key factor for exploring new routes is accurate predictions of airfares. This report highlights the concept of making forecasting models for airfares using the dataset from 638 existing routes. Pivotal consideration is given to demographic, geographical, and market-related variables, while the response variable is the average airfare. Apart from these variables, the potential influence of competitors, specifically Southwest Airlines, is also examined. We followed the **SEMMA (sample, explore, modify, model, assess)** process that is followed for data mining and analysing our data that suited our assignment i.e. Data understanding, Data preparation, Modelling and Evaluation. Please find the detailed steps provided below:

Deleted: .

1. EXPLORING THE DATA

The dataset was scrutinized to decipher the valuable insights and critically evaluated the possible associated assumptions for each variable and data type it holds. This information includes different factors that help airlines figure out how much to charge for flights between U.S. airports. These factors range from whether the destinations are popular vacation spots to the income and population of both the departure and arrival cities, along with the current fares set by various airlines on that route. We also considered how many flights are needed to travel between two airports, the distance of the flight, whether a budget airline like Southwest flies that route, and the types of slots and gates available at the airports.

<u>Scatter Plots</u>:

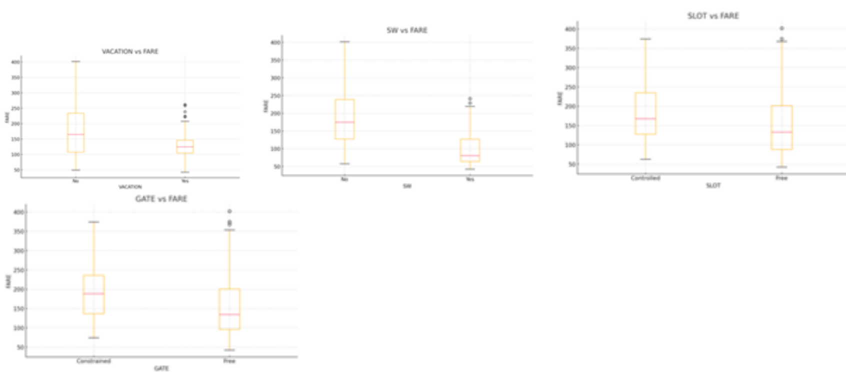
COUPON: A slight upward trend is visible; routes with more coupons (stops) tend to have higher fares.

DISTANCE: A positive correlation is evident, as fares increase with distance.

PAX: No clear linear trend, but fares may stabilize for routes with very high passenger numbers.

HI: Weak correlation with fares, suggesting other factors might have more influence.

Boxplots:



VACATION: Routes categorized as vacation destinations generally have slightly higher median fares.

SW: Routes served by Southwest Airlines have significantly lower fares, supporting the hypothesis in the assignment.

SLOT and GATE: Minimal differences in fares based on congestion or slot control.

2. MISSING VALUES

We checked the dataset for missing values, and we found that there were no missing entries. This shows the completeness of the data and makes the pre-processing easier.

3. TREATING OUTLIERS

For pruning the data values for continuous variables, we would be using statistical techniques i.e. outlier is any value that is outside the (mean - 3*standard deviation), (mean + 3* standard deviation) range, to ensure the normal distribution of data which enhances the later stage of model training. Please find the column wise details below:

1. Coupon

11 unusual values found for the coupon appear to be accurate as it is an average number of non-stops, and one stops flights for that route. Because of this, we decided to leave it as it is.

2. New

In NEW values, there are 34 outliers which have zero values. We kept these values because they represent there are no new carriers added to that route.

3. HI

In HI values, six values have been identified as outliers, they are still kept in the data they show high and low number of firms competing on that route which cannot be treated as outliers based on context of presence of competition

4. S_City Income

9 values are being identified as outliers in the data. We still kept them in the data they represent a level of income which can be very high and low for some cities.

5. PAX

22 values have been acknowledged as anomalies in the data. We have not removed them as PAX value shows density of passenger on that route which is valid.

Coupon Outlier	New Outlier	HI Outlier	S INCOME Outlier	E INCOME Outlier	S POP Outlier	E POP Outlier	DISTANCE Outlier	PAX Outlier	FARE Outlier
ok	ok	ok	ok	ok	ok	ok	ok	outlier	ok
ok	ok	ok	ok	ok	ok	ok	ok	outlier	ok
ok	ok	ok	ok	ok	ok	ok	ok	outlier	ok
ok	ok	ok	ok	ok	ok	ok	ok	outlier	ok
ok	ok	ok	ok	ok	ok	ok	ok	outlier	ok
ok	ok	ok	ok	ok	ok	ok	ok	outlier	ok
ok	ok	ok	ok	ok	ok	ok	ok	outlier	ok

Figure 1: Identifying Outliers

4. ENCODING

Before training the model, we need to convert the categorical variables into numerical values. We achieved this by using a straightforward encoding method. In our data, we found four important categorical variables: Vacation, SW, Slot, and Gate. Vacation tells us if there is a vacation route (encoded yes =1 and no = 0). SW indicates whether Southwest Airlines services that route (yes =1 and no = 0). For Slot, we check if either airport at the end of the route has slot controls (encoded free =1 and controlled = 0). Lastly, Gate looks at whether there are gate constraints at either endpoint airport (encoded free = 1 and Constrained = 0).

5. PARTITIONING THE DATA:

In our data analysis, we've focused on breaking down the data by looking at four important variables: Vacation, SW, Slot, and Gate. To make things simpler, we plan to cut down the number of categories. Starting with Vacation, we see that there are 1/4th instances where people say “yes” and 3/4th instances where they said ‘no’. Since no appears much more often, we can leave out the yes category in our review. The same goes for SW, where we have a few ‘yes’ and more ‘no’. Again, the higher count of no is what matters more for our data.

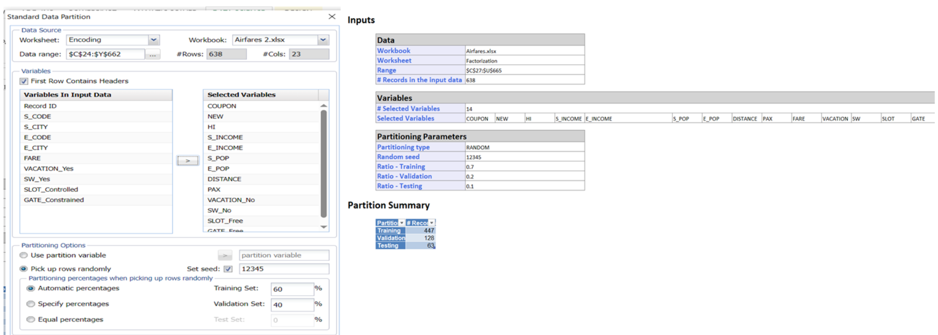


Figure 2: Standard Partitioning

Next, when we examine Slot, which refers to airport Slot control, we find less instances of ‘controlled’ slots and more of ‘free’ slots. Because free slots are much more common, we will concentrate on this category. For gate, the ‘constrained’ have low frequency compared to ‘free’ making it a primary focus. Lastly, we divided our dataset into two segments: 60% will be used for training and 40% for testing.

5. BUILDING THE MODEL

Six models were considered for our question and evaluated:

5.1 LINEAR REGRESSION

Linear regression was chosen as the baseline model due to its simplicity and interpretability. It assumes a linear relationship between the predictors and the response variable. While it provided reasonable accuracy, its inability to capture non-linear relationships limited its overall performance.

5.2 REGRESSION TREES

Regression trees were employed to address the limitations of linear models. Pruned trees were used to control overfitting and improve generalization. Despite their ability to capture non-linear interactions, the performance of regression trees was slightly inferior to other advanced methods.

5.3 NEURAL NETWORKS

Neural networks were considered for their ability to model complex, non-linear relationships. However, they exhibited high variance and overfitting, particularly given the dataset's size and structure. This made them less reliable compared to other methods like boosting.

5.4 BAGGING

Bagging combines multiple weak learners to reduce variance and improve robustness. While it was effective in mitigating noise, its overall performance was suboptimal for this dataset.

5.5 BOOSTING LINEAR REGRESSION

Boosting linear regression emerged as the most effective method. By iteratively focusing on the difficult-to-predict instances, it achieved the highest predictive accuracy across all evaluated metrics. This model was therefore selected for further analysis.

We use the partition data to build a linear regression model, as illustrated in the figure. All the variables are chosen as predictors except for 'fare,' which serves as the output variable. We also set the rescale method to normalization.

To check how well our model works, we looked at some key indicators shown in the figure.

- First up is SSE which stands for Sum of Squared Errors which measures total prediction error. Followed by RMSE, which stands for the square root of the average of the squared differences between what we predicted and the actual values. This number shows the size of prediction errors. The bigger values mean bigger errors. It's worth noting that RMSE can be affected significantly by outliers.
- Next, we have MAD or Mean Absolute Deviation. This one finds the average of the absolute differences between predicted and actual values. It tells us how big the errors are on average, no matter which way they go. MAD is more forgiving when it comes to outliers compared to RMSE.
- Lastly, $\mathbf{R^2}$, or the Coefficient of Determination, shows how much of the variation in the dependent variable our model explains. With a value of 0.7941, our model accounts for about 79.41% of that variation.

Predictor Screening

Predictor	Criteria	Included
Intercept	21.14237451	TRUE
COUPON	11.91596772	TRUE
NEW	21.11843264	TRUE
HI	21.11871208	TRUE
S_INCOME	21.08098056	TRUE
E_INCOME	17.84984071	TRUE
S_POP	16.71865859	TRUE
E_POP	14.65017807	TRUE
DISTANCE	18.49173421	TRUE
PAX	20.6753023	TRUE
VACATION	20.06633938	TRUE
SW	19.73322684	TRUE
SLOT	18.62475327	TRUE
GATE	20.92061145	TRUE

Figure 3 Predictors

The predictor screening in the figure helps us see if the variables we use to predict the true value are dependent on each other. The results indicate that there is no dependency among these predictor variables.

5.6 EVALUATION OF MODELS

Linear Regression

Metric	Value
SSE	147874.5748
MSE	1155.270116
RMSE	33.98926472
MAD	27.01270214
R2	0.794101363

Bagging Neural Networks

Metric	Value
SSE	788811.8
MSE	6162.592
RMSE	78.50218
MAD	69.18629
R2	-0.09833

Boosting Linear Regression

Metric	Value
SSE	147487.6
MSE	1152.247
RMSE	33.94476
MAD	26.98984
R2	0.79464

Decision Tree

Full tree

Metric	Value
SSE	178277.0889
MSE	1392.789757
RMSE	37.32009856
MAD	26.60942913
R2	0.751769297

Best Pruned

Metric	Value
SSE	161247.1
MSE	1259.743
RMSE	35.49286
MAD	26.33442
R2	0.775482

Minimum error

Metric	Value
SSE	161247.1
MSE	1259.743
RMSE	35.49286
MAD	26.33442
R2	0.775482

Neural Network

NetID	# Hidden Layers	# Neurons (Layer 1)	# Neurons (Layer 2)	Training SSE	Training RMSE	Training MSE	Validation SSE	Validation RMSE	Validation MSE
Net 18	2	2	3	2703303.6	77.77	6047.66	731458.78	75.59	5714.52
Net 42	2	6	3	2698665.9	77.7	6037.28	733606.6	75.71	5731.3
Net 54	2	8	3	2677367.3	77.39	5989.64	734714.32	75.76	5739.96
Net 24	2	3	3	2702947.5	77.76	6046.86	743865.91	76.23	5811.45

Figure 44 Performance Metrics from Different Models

5.7 Performance Metrics:

LINEAR REGRESSION:

SSE: 147,874.57
MSE: 1,155.27
RMSE: 33.99
MAD: 27.01
R^2: 0.794

Linear regression performed reasonably well, with an R^2 value indicating that approximately 79.4% of the variability in airfare could be explained by the predictors. However, the assumption of linearity limits its ability to capture complex relationships between variables.

REGRESSION TREES:

Regression trees were explored for their ability to model non-linear relationships and interactions. Two configurations were tested:

Pruned Tree:

SSE: 161,247.1
MSE: 1,259.74
RMSE: 35.49
R^2: 0.775

Regression trees were effective at handling non-linear relationships, but they underperformed compared to Linear Regression and Boosting Linear Regression. While pruning improved generalizability, the trees still exhibited a slight tendency toward overfitting, especially with complex datasets.

THE NEURAL NETWORKS:

Performance Metrics:

Validation SSE: 734,714.32
Validation RMSE: 75.76.

The neural networks demonstrated high variance and overfitting due to the limited size and complexity of the dataset. Additionally, the interpretability of neural networks is limited, making them less favorable for practical insights into key predictors of airfare.

BAGGING NEURAL NETWORKS

Bagging, or bootstrap aggregating, was employed to reduce variance by combining predictions from multiple models. However, its performance lagged significantly:

Performance Metrics:

SSE: 788,811.8
RMSE: 78.50
R^2: Negative, indicating poor model fit.

Bagging struggled to improve predictions, likely due to the dataset's limited size and the dominance of complex relationships that require targeted focus, which bagging does not inherently provide.

BOOSTING LINEAR REGRESSION

Boosting Linear Regression iteratively improves model performance by focusing on correcting prediction errors in successive iterations. This method achieved the best results among all models:

Performance Metrics:
SSE: 147,487.6
MSE: 1,152.25
RMSE: 33.94
MAD: 26.98
R^2: 0.795

2: SELECTION OF BEST MODEL TO PREDICT THE AVERAGE FARE ON A ROUTE

Boosting Linear Regression was chosen for it had the smallest RMSE (33.94) and the highest R^2 (0.795), indicating that it explained more variance in airfare and provided the most accurate predictions. The MAD (26.98) was also the lowest, showing that the average deviation of predictions from actual values was minimal.

Predicting fare value on the chosen model using the value given below:

Variable	Value
COUPON	1.202
NEW	3
VACATION	No
SW	No
HI	4442.141
S_INCOME	28760
E_INCOME	27664
E_POP	3195503
S_POP	4557004
SLOT	Free
GATE	Free
PAX	12782
DISTANCE	1976

Scoring

Record ID	Prediction: FARE
Record 1	235.6602822

Figure 5 5 Predicting the fare value

3: PREDICT THE REDUCTION IN AVERAGE FARE ON THE ABOVE ROUTE IF SOUTHWEST AIRLINES DECIDES TO COVER THIS ROUTE.

For this analysis, the Boosting Linear Regression model, identified as the best-performing model in Part A, was used.

COUPON	NEW	HI	S_INCOME	E_INCOME	S_POP	E_POP	DISTANCE	PAX	VACATION	SW	SLOT	GATE
1.202	3	4442.141	28760	27664	4557004	3195503	1976	12782	0	1	1	1

Scoring

Record ID	Prediction: FARE
Record 1	196.0698782

Figure 66 Fare Prediction

Fare With Southwest Airlines: When the variable **SW** is changed from "No" to "Yes," indicating the presence of Southwest Airlines, the predicted fare is reduced to **$196.07.**

The analysis reveals that Southwest Airlines has a significant impact on airfare pricing. The fare decrease represents a **$39.59(16.79%)** reduction compared to the baseline fare ($235.66).

QUESTION B) WHICH OF THE FACTORS (PREDICTOR VARIABLES) WILL NOT BE AVAILABLE FOR PREDICTING THE AVERAGE FARE FROM A **NEW AIRPORT**?

1: SELECTION OF VARIABLES:

The following are the results for considering each variable or disregarding it in the estimation of the average fare on new routes, based on availability before flights start operations:

Variables That Can Be Considered:

- VACATION (whether the road leads to a holiday spot): From tourism statistics, this is a recognizable characteristic of the destination city and can be determined well in advance of the beginning of operations.
- S_INCOME and E_INCOME (Starting and ending city's average income): These are readily available demographic information from economic or census reports.
- S_POP and E_POP (Beginning and ending population of the city): Like Income, demographic data is easily accessible through the national records.
- DISTANCE (Distance between two airports): The distance can also be inferred through the records and plays a significant role in the decision-making for introducing a new airport.
- SLOT and GATE (Airport congestion factors): These can be predicted well in advance and are defined by what infrastructure is in place.

Variables That Cannot Be Considered:

- PAX (Number of travelers on the route): No previous estimation of passenger demand is possible as flights need to begin first.
- SW (If the route is served by Southwest Airlines):
  Until it is announced, no one knows whether Southwest Airlines will decide to serve.
- HI (Herfindahl Index - market concentration):
  Market concentration depends on the competing airline's market shares, which are not available for a new route.
- NEW (Number of recently added carriers to the route):
  Until airlines confirm operational plans, their involvement is purely speculative.
- COUPON (Average number of coupons or stops on the route):
  This depends on the operational design of flights, which is not finalized before the route launches.
- FARE (Average fare): This is the target variable. Therefore, it cannot be included as a predictor.

2: DEVELOPMENT OF MODEL

As we found out that the performance of Boosting Linear Regression was better, we would be moving ahead with that model. We will be using the available predictors (**S_INCOME, E_INCOME, S_POP, E_POP, DISTANCE, VACATION, SW, SLOT, GATE**) to retrain the

Boosting Linear Regression model. While the model is applied to a reduced set of predictors, its iterative approach to minimizing error ensures robust performance even with fewer variables.

Inputs

Data	
Workbook	Book11
Worksheet	Factorization
Range	$C$27:$U$665
# Records in the input data	638

Variables									
# Selected Variables	9								
Selected Variables	S_INCOME	E_INCOME	S_POP	E_POP	DISTANCE	FARE	VACATION	SLOT	GATE

Partitioning Parameters	
Partitioning type	RANDOM
Random seed	12345
Ratio - Training	0.7
Ratio - Validation	0.2
Ratio - Testing	0.1

Partition Summary

Partition	# Records
Training	447
Validation	128
Testing	63

Figure 77 Selection of Parameters

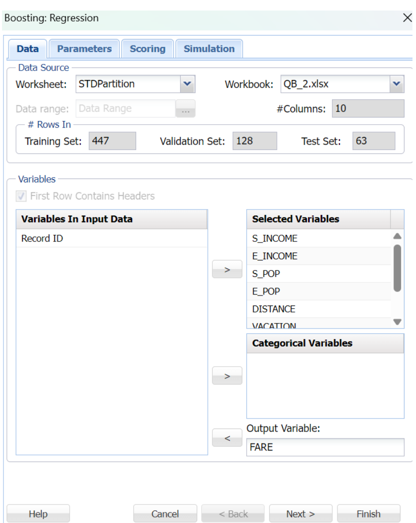


Figure 88 Selection of parameters for Model development

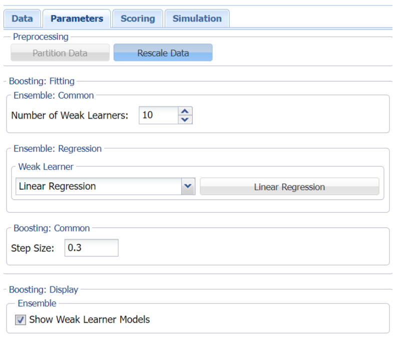


Figure 99 Setting up parameters

3: PREDICT THE AVERAGE FARE USING ONLY THE AVAILABLE (IN YOUR OPINION) DATA FROM THE RECORD

After building the model, we ran with the selected parameters, and we got the value of 233.433 which is very close to the model prediction in done with all the variables provided.

S_INCOME	E_INCOME	S_POP	E_POP	DISTANCE	VACATION	SW	SLOT	GATE
28760	27664	4557004	3195503	1976	0	0	1	1

Inputs

Data	
Workbook	QB_2.xlsx
Worksheet	Prediction 1
Range	$A$1:$M$2
# Records in the input data	1

Variables								
# Variables	8							
Model Variables	S_INCOME	E_INCOME	S_POP	E_POP	DISTANCE	VACATION	SLOT	GATE
Variables in New Data	S_INCOME	E_INCOME	S_POP	E_POP	DISTANCE	VACATION	SLOT	GATE

Scoring

Record ID	Prediction: FARE
Record 1	233.4335347

Figure 10 10Predictions using the model

4: COMPARE THE PERFORMANCE OF THIS MODEL WITH THE PERFORMANCE OF THE MODEL FROM ITEM A.

Error Metrics (SSE, MSE, RMSE, MAD):

There is a noticeable increase in errors across all metrics in the reduced model. For example, SSE increased by approximately 57%, and RMSE grew by about 25%, indicating a decline in predictive accuracy.

R^2 (Goodness of Fit):
The R^2 value dropped from 0.7946 to 0.6782, showing that the reduced model explains approximately 11.6% less variance in the average fare than the initial model.

The reduced model lacks crucial predictors such as COUPON, HI, and PAX, which carry significant information about pricing trends and market competition. Their absence likely results in weaker predictions.

Metric	Initial Model	Reduced Model	Difference
SSE	147487.5599	231101.5872	83614.0273
MSE	1152.246562	1805.48115	653.234588
RMSE	33.9447575	42.49095374	8.54619624
MAD	26.98983787	35.17179158	8.18195371
R^2	0.794640238	0.678217152	-0.11642309

Figure 11 Performance Comparison

MODEL FIT

The reduced model can provide rough predictions when no post-launch data (e.g., passenger volume or market competition) is available. It uses only economic, demographic, and operational data that are obtainable before flights commence. But, once the flights start operating, it is essential to reevaluate the model by incorporating the missing variables (e.g., **PAX** and **HI**) for more accurate predictions.