Analysis report

DATA MINING METHODOLOGIES AND APPLICATIONS WITH AIRLINES INDUSTRY DATASET

Quyen Thuc Nguyen

2 December 2022

Table of Contents

I. Introduction 4

II. Methodologies 4

1. Multilinear Regression Analysis 4

2. Two-Step Clustering 4

3. Artificial Neural Network: Multilayer Perceptron 5

III. Results and Discussions 6

1. Multilinear Regression Analysis 6

2. Two-Step Clustering Analysis 11

3. Artificial Neural Network – Multilayer Perceptron 15

IV. Practical implications for end-users 21

1. Revenue forecasting 21

2. Customer satisfaction improvement 21

V. Conclusion 22

List of Tables

Table 1 Correlations between continuous variables 6
Table 2 Descriptive statistics of revenue, average number of seats, and cost of ticket 6
Table 3 Variables entered/Removed 7
Table 4 Model Summary 7
Table 5 ANOVA results 7
Table 6 Coefficients of multilinear regression model 8
Table 7 Residuals Statistics 9
Table 8 MLP case processing summary 15
Table 9 MLP network information 15
Table 10 MLP model summary 17
Table 11 Classification of MLP 17
Table 12 Parameter estimates 18
Table 13 Area under the curve of MLP model 19
Table 14 Independent variable importance (IVI) 20
Table 15 Customer satisfaction frequency 21

Table of Figures

Figure 1 Normal P-P plot of regression standardised residual 9
Figure 2 Histogram of regression standardised residual 9
Figure 3 3D scatterplot of multilinear regression model 10
Figure 4 Two-Step clustering model summary 11
Figure 5 Cluster sizes 11
Figure 6 Cluster details 12
Figure 7 Predictor importance of Two-Step clustering 12
Figure 8 Cluster comparison 14
Figure 9 MLP architecture with two hidden layers 16
Figure 10 ROC curve for MLP model 19
Figure 11 Normalised importance of predictor variables 20

I. Introduction

As many other types of services sectors, aviation industry has adopted big data analytics to improve their operations and service quality as well as financial and risks management. In addition, applications of data mining plays a crucial role in identifying underlying patterns of historical data, diagnose reasons behind existing problems and predict future outcomes, and thus help businesses to gain actionable insights and make data-driven decisions.

With given dataset of airlines industry, this report will unravel factors that influence revenues and customer satisfaction by using data mining techniques which are multilinear regression, Two-Step clustering and artificial neural network (ANN).

II. Methodologies

This study will explore the airline industry dataset in purpose to determine factors influencing revenue and customer satisfaction. The data mining techniques used are Multilinear Regression, Two-Step Clustering, and Artificial Neural Network. The technical procedures are performed using SPSS.

1. Multilinear Regression Analysis

Tickets, CO2 emissions, average seats, and mileage are being considered for multilinear regression. These variables are statistically independent and continuous, which meets the assumptions for linear regression.

Second, regression analysis may determine how independent factors affect a dependent variable, unlike correlation coefficients. Multiple linear regression is effective for investigating airline revenue components in this report.

Limitations of the technique

Data must be independent and regularly distributed — which is not always the case. Revenues, ticket prices, CO2 emissions, and mileages in this dataset have severe outliers. Heteroskedasticity, caused by outliers, can drastically affect expected outcomes.

2. Two-Step Clustering

Two-Step clustering is implemented in SPSS with the categorical variable being customer satisfaction, and scale variables including mileage covered, cost of tickets, revenue, and CO_2 emissions. In addition, a new variable of clustering membership is also created to complement the later application of a multilayer perceptron neural network in predictive analysis.

Limitations of the technique

Due to various merging criteria, different clustering approaches might yield varied outcomes in data mining. Except for simple linkage, clustering outcomes are easily altered by variable sorting. Cluster merging is predicated on the likeness of one observation to the cluster, therefore when instances are nullified, the analysis becomes unstable.

3. Artificial Neural Network: Multilayer Perceptron

This report forecasts airline customer satisfaction using SPSS MLP. MLP is a feedforward network without back-loops. The MLP network predicts customer happiness using category and scale factors.

Neural network analysis is done in SPSS. The dependent variable is customer satisfaction on a scale of 1 to 5, with 1 being severely unhappy, 2 being unsatisfied, 3 being neutral, 4 being satisfied, and 5 being extremely satisfied. The independent variables include each observation's Two-Step cluster membership, average number of seats, ticket price, revenue, CO_2 emissions, miles covered, and taxi-in time. Two hidden layers make to the model's architecture.

Limitations of the technique

MLPs have restrictions. First, MPL networks need many patterns and iterations to learn. Second, dataset properties affect learning convergence. MLP hidden layer neurons and layers are hard to count. Finally, MLP is a versatile predictive analytic model, but its synaptic weights are hard to understand, therefore it fails to explain connection determination.

III. Results and Discussions

1. Multilinear Regression Analysis

Pearson's correlation coefficients demonstrate that total revenue has substantial positive connections with ticket cost and average seat count, with r-values of 0.864 and 0.366, respectively (see Table 1). The cost of tickets and average number of seats are utilised as independent variables in a multilinear regression model with revenue as the dependent variable since their p-values are less than 0.001.

Table 1 Correlations between continuous variables

Correlations

		Cost of ticket (£)	Sum of Total_CO2_emissions	Average Number of seat	Total revenue	Taxi-in Time (Minutes)	Sum of Mileage covered
Cost of ticket (£)	Pearson Correlation	--					
	N	347					
Sum of Total_CO2_emissions	Pearson Correlation	.000	--				
	Sig. (2-tailed)	.999					
	N	347	347				
Average Number of seat	Pearson Correlation	-.006	.049	--			
	Sig. (2-tailed)	.914	.364				
	N	347	347	347			
Total revenue	Pearson Correlation	.864**	.011	.366**	--		
	Sig. (2-tailed)	<.001	.832	<.001			
	N	347	347	347	347		
Taxi-in Time (Minutes)	Pearson Correlation	-.017	.020	-.007	-.026	--	
	Sig. (2-tailed)	.748	.713	.898	.630		
	N	347	347	347	347	347	
Sum of Mileage covered	Pearson Correlation	-.049	-.013	-.004	-.045	.073	--
	Sig. (2-tailed)	.366	.805	.940	.400	.176	
	N	347	347	347	347	347	347

**. Correlation is significant at the 0.01 level (2-tailed).

This analysis uses 347 observations per variable (see Table 2). Table 2 also includes means and standard deviations, which reflect the dispersion of data around the means. Revenue averages 556,198.07 and varies by 600,854.58. The mean and standard deviation for average seat count are 1,44.27 and 572.598. Finally, ticket prices average 385.94 and vary 360.14.

Table 2 Descriptive statistics of revenue, average number of seats, and cost of ticket

Descriptive Statistics

	Mean	Std. Deviation	N
Total revenue	556,198.07	600,854.580	347
Average Number of seat	1,444.27	572.598	347
Cost of ticket (£)	385.94	360.140	347

Table 3 shows that there are 3 independent variables entered and no variables are removed, and dependent variable is total revenue.

Table 3 Variables entered/Removed

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Cost of ticket (£), Average Number of seat[b]	.	Enter

a. Dependent Variable: Total revenue

b. All requested variables entered.

Table 4 shows that the regression is well-fitted. R, the multiple correlation coefficient, is 0.940, indicating that ticket price and average seat count greatly affect revenue. R2 shows that cost of tickets and average number of seats explain 88.4% of revenue volatility, which is considerable. Adjusted R2 is adjusted R2 if the model has unimportant predictors (Corporate Finance Institute, 2022). All independent variables are significant in this model because adjusted R2 = R2. The estimate's standard error, 204,958.461, evaluates the model's accuracy and residual variability around the regression line.

Table 4 Model Summary

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.940[a]	.884	.884	204,958.461

a. Predictors: (Constant), Cost of ticket (£), Average Number of seat

b. Dependent Variable: Total revenue

ANOVA (Analysis of variance) results shows the F statistics F(2,344) = 1,1314.802 with the significance level of less than 0.001 and much smaller than 0.05, indicating that the multilinear regression model developed can significantly forecast revenue (see Table 5).

Table 5 ANOVA results

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.105E+14	2	5.523E+13	1314.802	<.001[b]
	Residual	1.445E+13	344	42007970628		
	Total	1.249E+14	346			

a. Dependent Variable: Total revenue

b. Predictors: (Constant), Cost of ticket (£), Average Number of seat

Table 6 interprets the gradients and intercept of regression model, and that considered predictors and constant are all significant as the p-values are all <0.001 and much smaller than 0.05.

Table 6 Coefficients of multilinear regression model

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-563593.843	32198.745		-17.504	<.001
	Average Number of seat	389.090	19.244	.371	20.219	<.001
	Cost of ticket (£)	1445.422	30.596	.866	47.242	<.001

a. Dependent Variable: Total revenue

The developed multilinear regression model is as follows:

$$\hat{y} = 1{,}445.422 * (x_1) + 389.09 * (x_2) - 563{,}593.843$$

With ŷ as the estimated value of revenue, x_1 as the value of cost of tickets, and x_2 as the value of average number of seats.

Cost of ticket's gradient of 1,445.422 means that for every extra £1, the revenue increases by £1,445.422 on an average.

The average number of seats' gradient of 389.09 means that for every extra seat, revenue increases by £389.09 on average.

The constant of -563,593.843 indicates that when the cost of tickets and the average number of seats equal 0, there will be an average loss of -£563,593.843.

The residual mean of the regression model equals zero indicating a normal distribution in the residuals (see Table 7). Histogram of the residual's normal distribution is shown in Fig. 1.

Table 7 Residuals Statistics

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-345,008.78	2,416,959.25	556,198.07	565,031.925	347
Residual	-1,487,892.875	715,040.813	.000	204,365.237	347
Std. Predicted Value	-1.595	3.293	.000	1.000	347
Std. Residual	-7.259	3.489	.000	.997	347

a. Dependent Variable: Total revenue

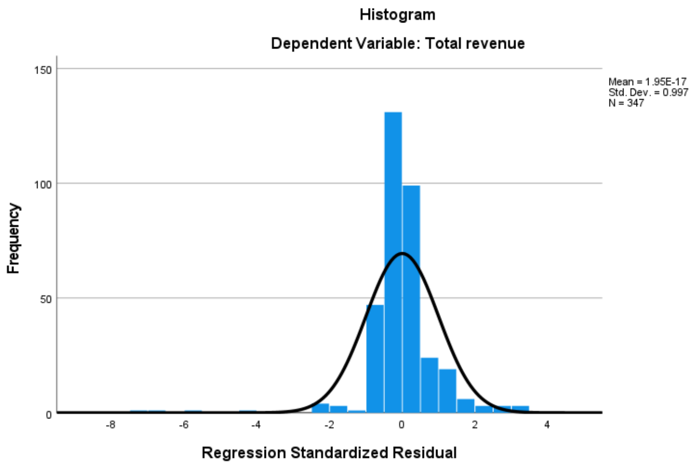


Figure 2 Histogram of regression standardised residual

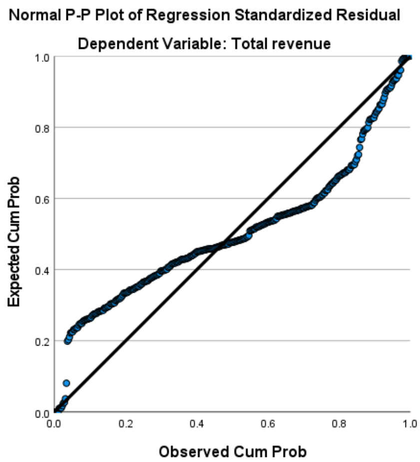


Figure 1 Normal P-P plot of regression standardised residual

The normal probability plot depicts the cumulative frequency of the model's distribution of standardised residuals against residuals with a normal probability graph scale (see Fig. 2). Outliers and heteroskedasticity are shown by residual data points that curve. Heteroskedasticity indicates residual variance is not constant, which may impact prediction accuracy. However the phenomenon is only observed rather than resolved.

A 3D scatterplot is produced to visualise the relationship between revenue, cost of tickets and average number of seats on a dimension (see Fig. 3). This allows to observe how revenue changes across different levels of the other two variables.

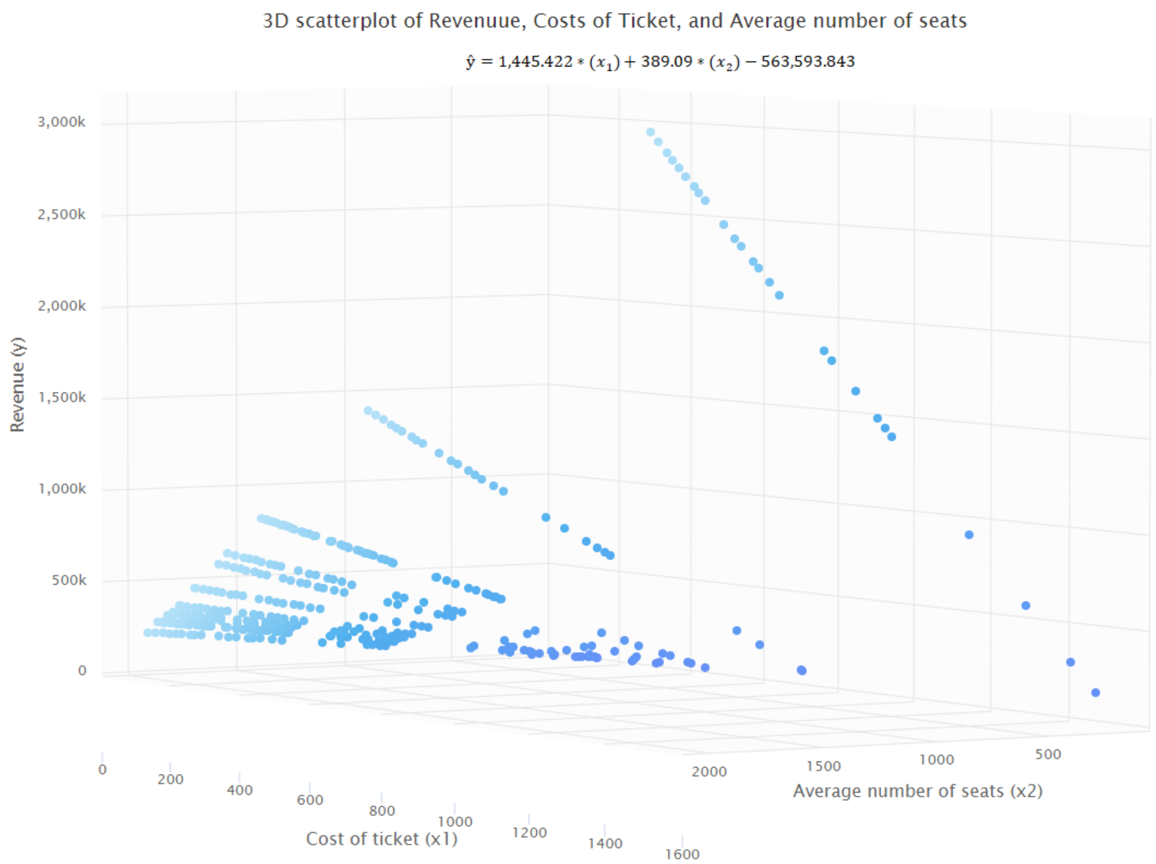


Figure 3 3D scatterplot of multilinear regression model

2. Two-Step Clustering Analysis

The summary of Two-Step clustering model (Fig. 4) shows 7 clusters were produced based on 5 input features. Furthermore, the model's quality is evaluated as 'Good', which implies that clustering results are reliable.

Model Summary

Algorithm	TwoStep
Inputs	5
Clusters	7

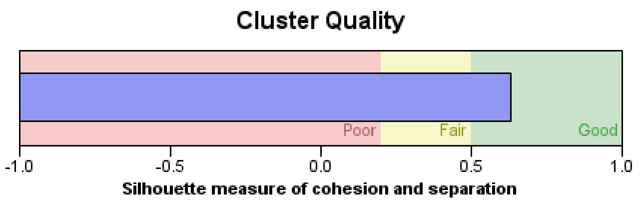
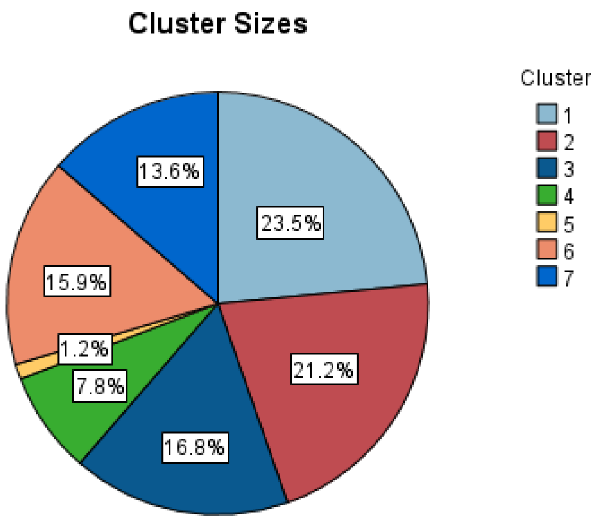


Figure 4 Two-Step clustering model summary

The frequency of each cluster is depicts in the figure of cluster sizes (Fig. 5). Among 7 clusters, cluster 1 is the largest with 81 members, making up 23.5% of the sample size. Following is cluster 2 with 21.3%. The smallest cluster is cluster 5 with only 4 members, making up 1.2% of the sample size.



Size of Smallest Cluster	4 (1.2%)
Size of Largest Cluster	81 (23.5%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	20.25

Figure 5 Cluster sizes

Clusters

Input (Predictor) Importance
1.0 0.8 0.6 0.4 0.2 0.0

Cluster	1	2	3	6	7	4	5
Label							
Description							
Size	23.5% (81)	21.2% (73)	16.8% (58)	15.9% (55)	13.6% (47)	7.8% (27)	1.2% (4)
Inputs	Sum of Mileage covered 54,889,300.07	Sum of Mileage covered 37,494,846.88	Sum of Mileage covered 23,482,314.14	Sum of Mileage covered 39,192,032.47	Sum of Mileage covered 16,091,597.09	Sum of Mileage covered 32,933,033.56	Sum of Mileage covered 8,106,525,000.00
	Customer satisfaction 2 (100.0%)	Customer satisfaction 1 (100.0%)	Customer satisfaction 3 (100.0%)	Customer satisfaction 4 (100.0%)	Customer satisfaction 5 (100.0%)	Customer satisfaction 2 (29.6%)	Customer satisfaction 1 (50.0%)
	Cost of ticket (£) 314.69	Cost of ticket (£) 268.63	Cost of ticket (£) 309.83	Cost of ticket (£) 302.18	Cost of ticket (£) 293.40	Cost of ticket (£) 1,450.74	Cost of ticket (£) 222.50
	Total revenue 501,635.31	Total revenue 364,432.33	Total revenue 456,238.45	Total revenue 473,535.45	Total revenue 323,218.30	Total revenue 2,096,386.30	Total revenue 301,225.00
	Sum of Total_CO2_ emissions 4,191,744.66	Sum of Total_CO2_ emissions 5,158,443.89	Sum of Total_CO2_ emissions 8,167,771.44	Sum of Total_CO2_ emissions 5,387,954.34	Sum of Total_CO2_ emissions 6,070,743.60	Sum of Total_CO2_ emissions 28,311,944.43	Sum of Total_CO2_ emissions 2,482,065.96

Figure 6 Cluster details

The order of the clusters is from left to right, sorted by cluster sizes (Fig. 6). The means of each variable in each cluster are shown, indicating the clusters are well distinguished.

Across variables used to classify airline flights, the variable of mileage covered is the most important with a ratio of 1.0, meanwhile CO_2 emissions factor is the least important predictor with a ratio close to zero (Fig. 7).

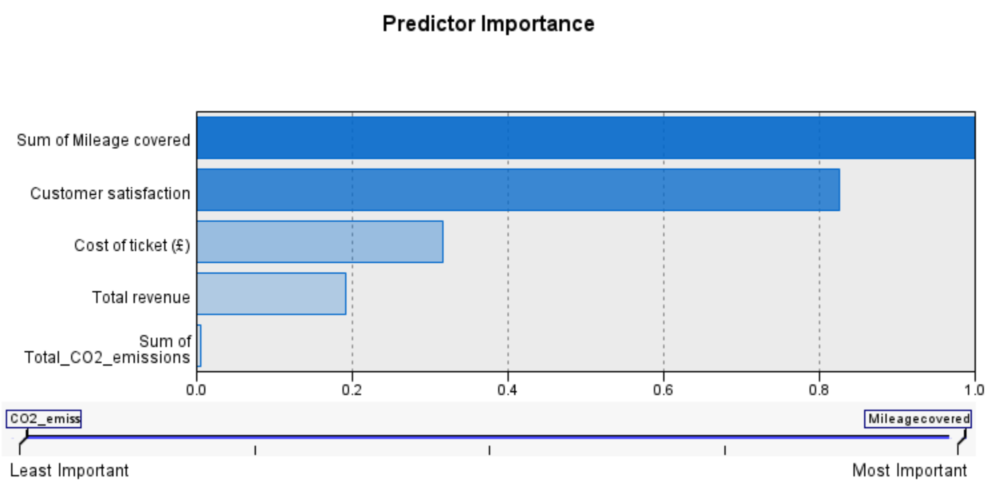


Figure 7 Predictor importance of Two-Step clustering

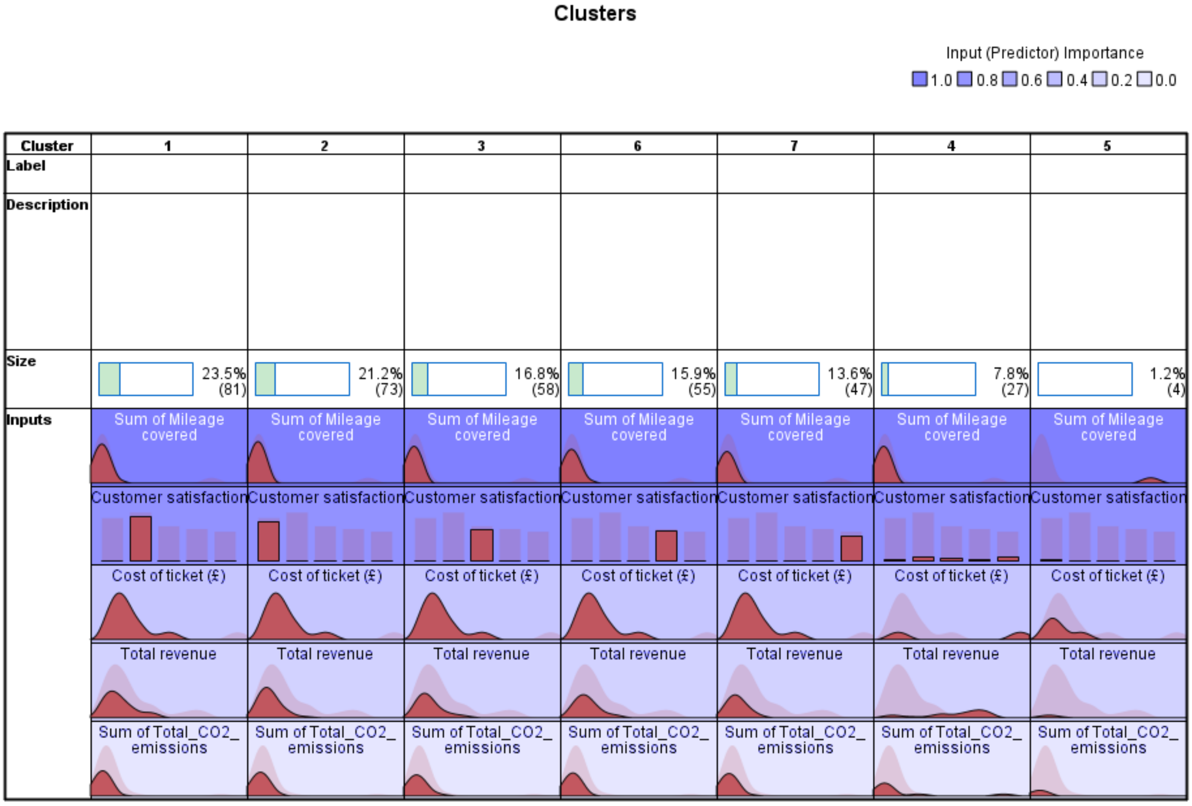


Figure 8 Clusters cell distributions summary

Looking at the cell distributions (Fig. 8), characteristics of each cluster can be defined as follows:

- Airplane flights in cluster 1, 2, 3, 6, and 7 all have a low to moderate amount of mileages covered, cost of tickets, revenues, and low total CO_2 emissions. The factor separating them is the customer satisfaction. Cluster 1, 2, 3, 6, and 7 are flights with customer satisfaction of 2, 1, 3, 4, and 5, respectively.
- Airline flights in cluster 4 has a low to moderate mileage covered, however these flights' numbers of cost of ticket, revenue, customer satisfaction and CO_2 emissions range from low to high, only with a low frequency.
- Airline flights in cluster 5 share the same traits as cluster 4, except for high numbers of mileage covered.

Interpreting from cluster comparison (Fig. 9), cluster 7 are air flights with the highest customer satisfaction, which also often account for the highest number of mileages covered, cost of tickets, revenue, and total CO_2 emission. On the contrary, cluster 2 are flights with the lowest customer satisfaction and they also have low mileages covered, cost of tickets, revenue, and CO_2 emissions.

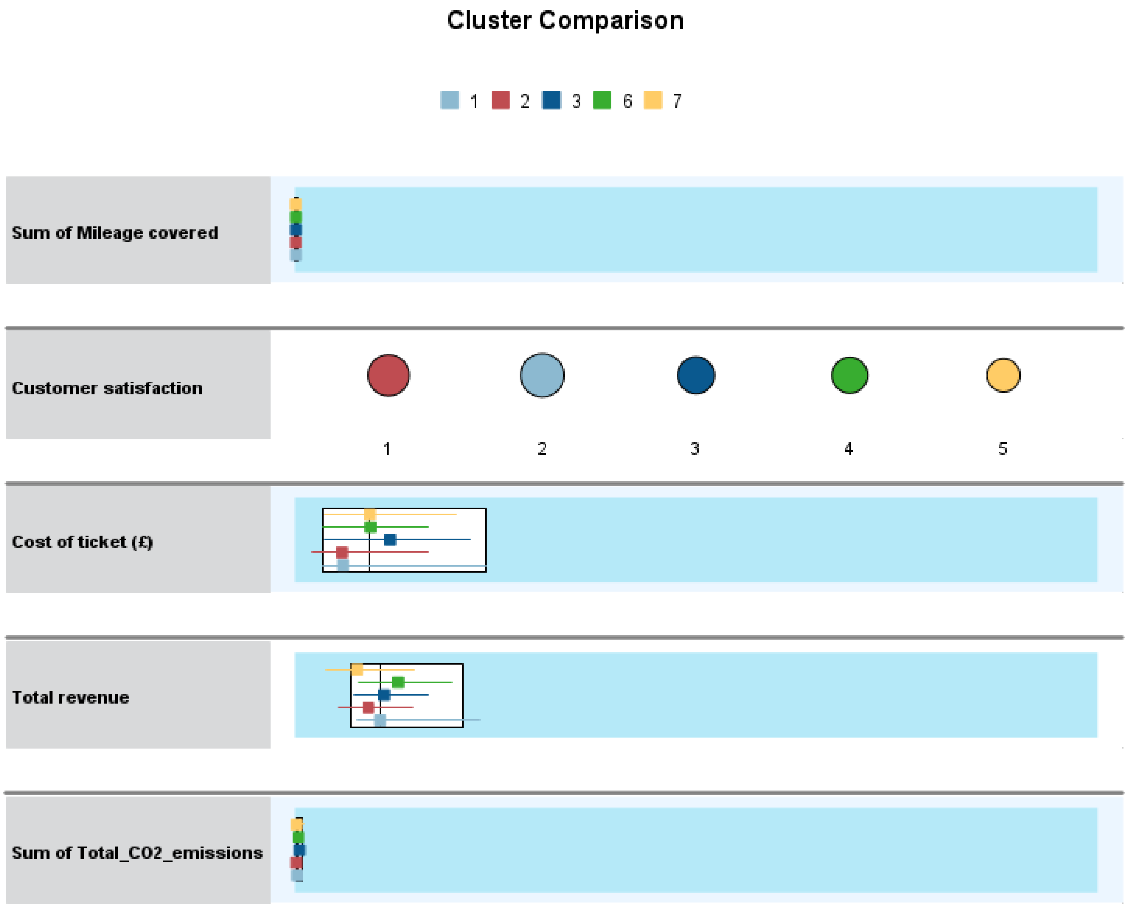


Figure 8 Cluster comparison

3. Artificial Neural Network – Multilayer Perceptron

The case processing summary (Table 8) shows that in total of 347 samples, there are 222 training samples (64.3%), 123 testing samples (35.7%), and 2 samples are excluded.

Table 8 MLP case processing summary

Case Processing Summary

		N	Percent
Sample	Training	222	64.3%
	Testing	123	35.7%
Valid		345	100.0%
Excluded		2	
Total		347	

Table 9 and Fig. 9 summarise the MLP model. The input layer has one component, clustering membership, and six covariates: average number of seats, taxi-in time, ticket price, total CO2 emissions, revenue, and mileage. Two buried levels have 9 and 7 nodes, respectively. Hyperbolic tangent activates hidden layers. Customer happiness is the sole variable in the output layer. Output layer Softmax activation.

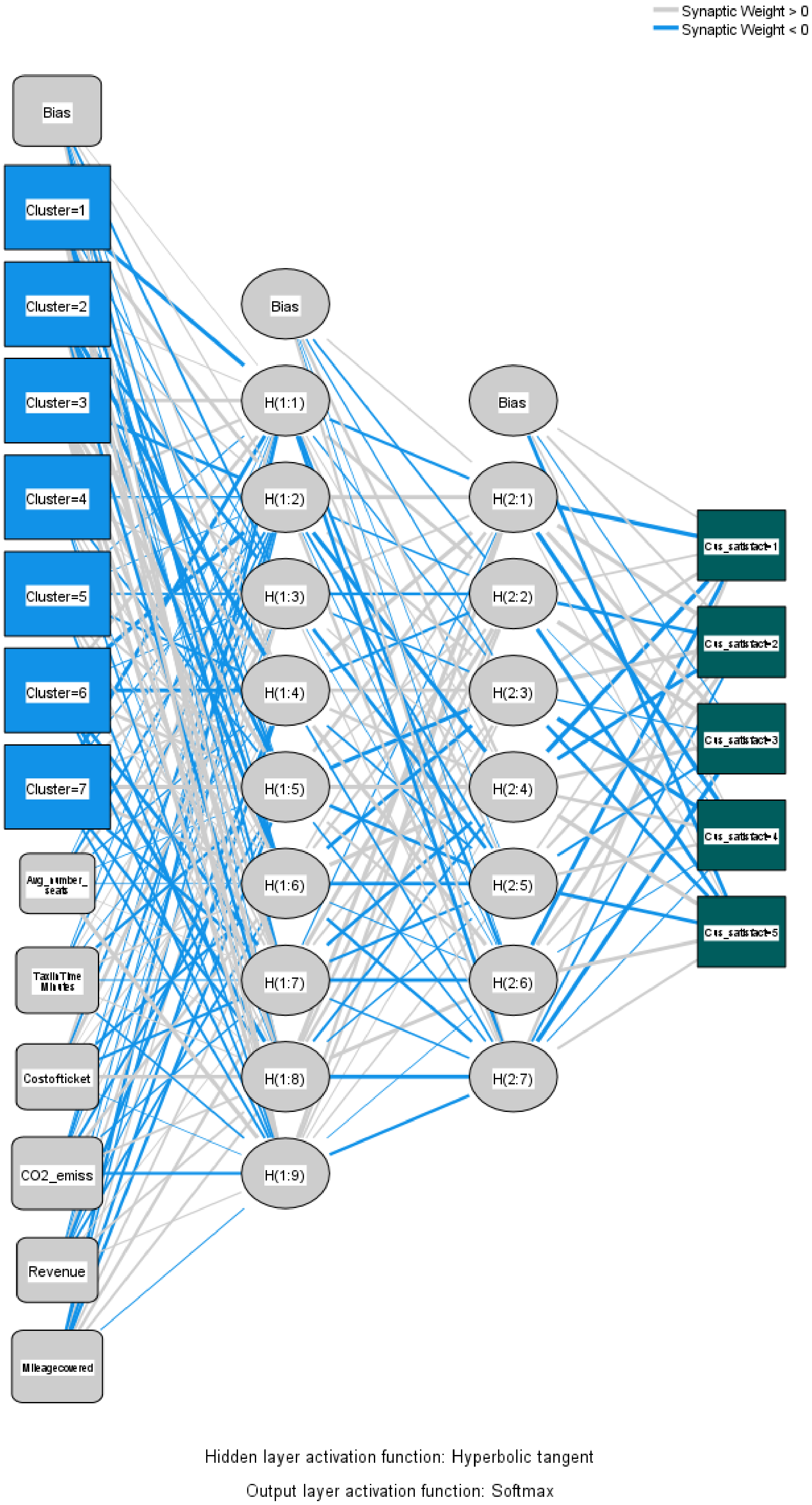
Table 9 MLP network information

Network Information

Input Layer	Factors	1	TwoStep Cluster Number
	Covariates	1	Average Number of seat
		2	Taxi-in Time (Minutes)
		3	Cost of ticket (£)
		4	Sum of Total_CO2_emissions
		5	Total revenue
		6	Sum of Mileage covered
	Number of Units[a]		13
	Rescaling Method for Covariates		Standardized
Hidden Layer(s)	Number of Hidden Layers		2
	Number of Units in Hidden Layer 1[a]		9
	Number of Units in Hidden Layer 2[a]		7
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	Customer satisfaction
	Number of Units		5
	Activation Function		Softmax
	Error Function		Cross-entropy

a. Excluding the bias unit

Figure 9 MLP architecture with two hidden layers



The model summary (Table 10) reveals that the developed MLP model has a substantially high rate of accuracy as there are only 2.7% incorrect predictions in training process and 8.9% incorrect predictions in testing process.

Table 10 MLP model summary

Model Summary

Training	Cross Entropy Error	26.299
	Percent Incorrect Predictions	2.7%
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	0:00:00.03
Testing	Cross Entropy Error	32.843
	Percent Incorrect Predictions	8.9%

Dependent Variable: Customer satisfaction

a. Error computations are based on the testing sample.

The classification table displays the rate of accuracy of customer satisfaction predictions by partition and overall. Within the training and testing samples across 5 levels of customer satisfaction, the highest percentage of correction is 100% and the lowest is 85.2%.

Table 11 Classification of MLP

Classification

		Predicted					
Sample	Observed	1	2	3	4	5	Percent Correct
Training	1	51	0	1	1	0	96.2%
	2	1	60	0	0	0	98.4%
	3	0	0	43	0	0	100.0%
	4	0	0	0	38	0	100.0%
	5	0	1	1	1	24	88.9%
	Overall Percent	23.4%	27.5%	20.3%	18.0%	10.8%	97.3%
Testing	1	23	0	1	2	0	88.5%
	2	0	25	2	1	0	89.3%
	3	0	0	20	1	0	95.2%
	4	0	0	0	21	0	100.0%
	5	0	1	1	2	23	85.2%
	Overall Percent	18.7%	21.1%	19.5%	22.0%	18.7%	91.1%

Dependent Variable: Customer satisfaction

In Table 12, parameter estimates for input, hidden, and output layers are presented for the MLP network.

Table 12 Parameter estimates

Parameter Estimates

		Predicted																				
		Hidden Layer 1									Hidden Layer 2							Output Layer				
Predictor		H(1:1)	H(1:2)	H(1:3)	H(1:4)	H(1:5)	H(1:6)	H(1:7)	H(1:8)	H(1:9)	H(2:1)	H(2:2)	H(2:3)	H(2:4)	H(2:5)	H(2:6)	H(2:7)	[Cus_satisfact =1]	[Cus_satisfact =2]	[Cus_satisfact =3]	[Cus_satisfact =4]	[Cus_satisfact =5]
Input Layer	(Bias)	.037	.151	-.312	.189	-.054	-.152	-.340	-.005	.345												
	[Cluster=1]	-.986	.809	-.135	.457	-.473	-.767	-.102	.994	.922												
	[Cluster=2]	.082	.172	-.600	-1.489	-1.012	-.984	1.106	-.497	.189												
	[Cluster=3]	.645	-.626	.242	.819	1.328	1.184	.894	1.051	1.001												
	[Cluster=4]	.383	-.186	.191	.005	-.256	.120	.191	-.019	-.218												
	[Cluster=5]	-.094	-.052	-.262	-.375	.313	.135	.319	.179	-.269												
	[Cluster=6]	.657	-1.269	-.082	-.922	.572	.040	-.682	-.957	-.341												
	[Cluster=7]	-.868	-.131	.414	.199	.949	.724	-.550	-.475	-.304												
	Avg_number_seats	-.007	-.317	.378	-.145	-.026	-.029	.019	.022	1.034												
	TaxiinTimeMinutes	-.040	-.212	.072	-.071	-.420	.232	-.053	.125	-.226												
	Costofticket	.311	-.281	.716	.011	.116	-.382	-.360	.498	-.008												
	CO2_emiss	-.071	-.377	-.087	.091	-.372	.487	-.172	.288	-.397												
	Revenue	-.104	-.031	.267	-.618	-.106	-.293	.057	.296	.133												
	Mileagecovered	-.541	-.191	.158	-.457	-.481	.262	.493	.320	-.073												
Hidden Layer 1	(Bias)										.161	-.234	-.135	.468	-.015	-.037	.376					
	H(1:1)										-.383	.788	-.148	.083	.265	-.523	-.696					
	H(1:2)										.918	-.220	.407	-1.024	-.309	-.036	.514					
	H(1:3)										.277	-.434	-.228	.149	-.660	-.015	.133					
	H(1:4)										1.163	-.350	.643	.451	-.259	.975	-.243					
	H(1:5)										.654	.228	-.920	.859	-.619	.390	-.221					
	H(1:6)										.221	.365	-.743	1.120	-.801	-.085	-.471					
	H(1:7)										-.129	1.067	.288	-.603	-.378	-.663	-.207					
	H(1:8)										1.472	.539	.524	-.326	-.180	.779	-.660					
	H(1:9)										.531	.446	.229	.086	.181	-.023	-.478					
Hidden Layer 2	(Bias)																	.181	.319	-.094	.181	-.704
	H(2:1)																	-1.681	1.145	1.216	-1.955	.125
	H(2:2)																	.254	-.632	1.732	-.004	-.919
	H(2:3)																	.615	1.366	-.148	-1.245	-.723
	H(2:4)																	-1.730	-.847	.861	.642	1.484
	H(2:5)																	.239	.404	-.349	.402	-.682
	H(2:6)																	-.943	.939	.253	-.074	.888
	H(2:7)																	.527	-.303	-1.401	-.092	.353

Figure 3 illustrates the sensitivity and specificity diagram of estimated results of customer satisfaction, where 1 is extremely dissatisfied, 2 is dissatisfied, 3 is neutral, 4 is satisfied, and 5 is extremely satisfied. The straight line in 45 degrees from bottom left corner to the upper right corner of the chart defines the circumstance of random guessing. The further the curves deviate away from the line, the more accurate the estimates.

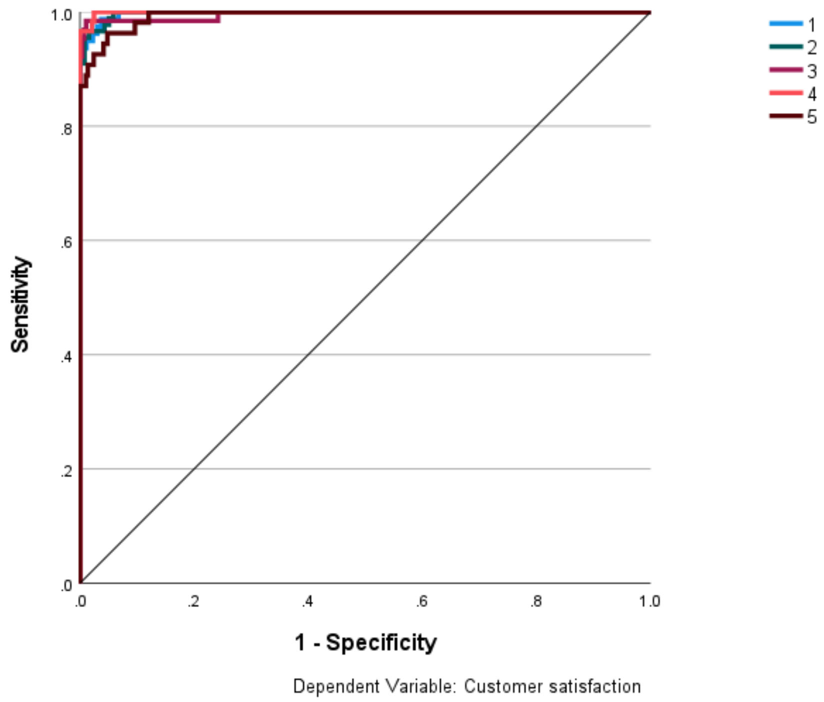


Figure 10 ROC curve for MLP model

The percentage of area under the curve of each customer satisfaction level is extremely high at 99.7% (Table 13).

Table 13 Area under the curve of MLP model

Area Under the Curve

		Area
Customer satisfaction	1	.998
	2	.998
	3	.998
	4	.999
	5	.997

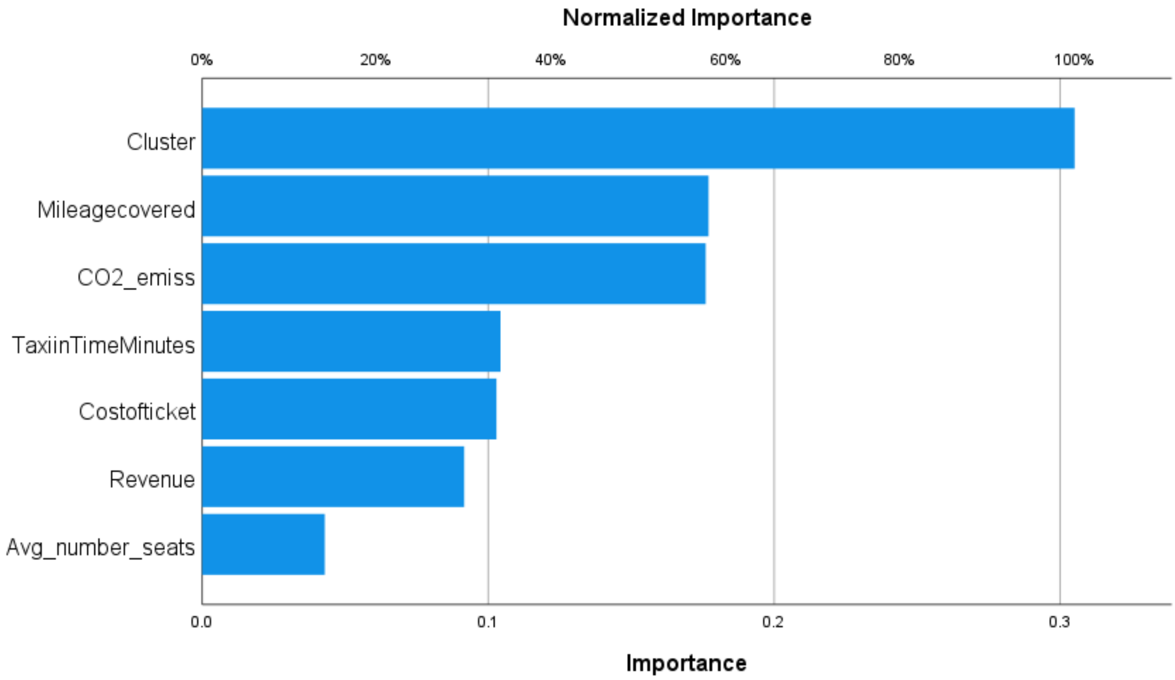
The independent variable importance (Table 14) and normalised importance (Fig. 11) shows that the Two-Step cluster membership is the most influencing predictor with percentage of normalised importance of 100%, followed by mileages covered (58%) and CO_2 emissions (57.7%).

Table 14 Independent variable importance (IVI)

Independent Variable Importance

	Importance	Normalized Importance
TwoStep Cluster Number	.305	100.0%
Average Number of seat	.043	14.0%
Taxi-in Time (Minutes)	.104	34.2%
Cost of ticket (£)	.103	33.7%
Sum of Total_CO2_emissions	.176	57.7%
Total revenue	.092	30.0%
Sum of Mileage covered	.177	58.0%

Figure 11 Normalised importance of predictor variables



IV. Practical implications for end-users

1. Revenue forecasting

With the developed multilinear regression model, airlines business can manipulate influencing factors on revenue, i.e., cost of tickets and the average number of seats, to make informed decisions.

Revenue forecasts is crucial in business plan, as it could help a firm to strategise their growth rates, allocate budgets and manage their cash flows both short-term and long-term. In other words, forecasting could assist business foresee the future's challenges as well as opportunities, and thus take control over financial and risks management to purposefully navigate their firm.

2. Customer satisfaction improvement

Analysis results show that customers that have extremely satisfied experience with flights that cover a high amount of mileages, have expensive cost of ticket, high revenue, and high amount of total CO_2 emissions, in which the mileages covered is the most influencing factor of customer satisfaction.

Furthermore, the data set reveals that 22.8% customers feel extremely dissatisfied and 25.6% of them feel dissatisfied (Table 15). With analysis results from Two-Step Clustering and MLP model, airlines could adjust their business operation focus on short to moderate flights, moderate price of ticket and minimise CO_2 emissions of each flight.

Table 15 Customer satisfaction frequency

Customer satisfaction

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	79	22.8	22.9	22.9
	2	89	25.6	25.8	48.7
	3	64	18.4	18.6	67.2
	4	59	17.0	17.1	84.3
	5	54	15.6	15.7	100.0
	Total	345	99.4	100.0	
Missing	System	2	.6		
Total		347	100.0		

On the other hand, variables considered are not quite related to customers' own opinions regarding other qualities, for example, pre-flight, in-flight, and post-flight services (Namukasa, 2013), but rather than airlines' statistics themselves, which is a limitation in developing the most suitable model to predict customer satisfaction.

V. Conclusion

Analysis results show that within the airlines industry, factors influencing revenue are cost of ticket and the average number of seats, which can explain approximately 88.4% variation of revenue.

On the other hand, the findings of the analysis establishes a linkage between high customer satisfaction with low to moderate mileages covered, cheap ticket price, and low revenue. 27 observations in cluster 4 indicates that high revenue often associates with high cost of tickets, short flights, low CO_2 emissions and customer satisfaction varied from extremely dissatisfied to extremely satisfied.

In general, with the current dataset, business can utilise findings and results from multilinear regression , clustering, and multilayer perception neural network to forecast revenue to make informed decisions and manage influencing factors to improve customer satisfaction.