Data Mining for Business SCH MGMT 655

Final Project Report

Predicting Airfare on Novel Routes

Group members -:

Subhadeep Prasad Bose

Sakhi Namireddy

Tanay Vaddiparthi

Executive Summary

This report deals with predictions related to the airfare in novel routes after the 1978 air deregulation act which resulted in significant alterations in the US airline industry, thus freeing fares and routes from regulatory control. This in turn led to the emergence of budget air carriers like Southwest Airlines and some others that led to heightened competition and a gradual increase in demand to travel by air and thus leading to airline companies hiring analysts to make predictions for new routes and airfare. Some of the major objectives of the project include creating predictive models on Analytics Solver and creating regression models using R and SAC and then finally comparing the results of all the outputs to get an optimum predictive model that could be used.

To make the regression models using the Analytics Solver we use methods like Linear Regression, k-Nearest Neighbors (kNN), CART. Following that, there would be regression models generated by R Programming and SAC. In the conclusion we have made a solitary comparison between all the three models to suggest which model is supposed to be deemed as the best approach for this dataset based on our emperical observations.

Subhadeep Prasad Bose

Sakhi Namireddy

Tanay Vaddiparthi

Regression Using Analytics Solver

1. Linear Regression Model

- The initial step for any regression model is Typecasting. The following table shows the Typecasting used in the running model.

○ S_Code	○ Categorical
○ S_City	○ Categorical
○ E_Code	○ Categorical
○ E_City	○ Categorical
○ Coupon	○ Numerical
○ New	○ Numerical
○ Vacation	○ Categorical
○ SW	○ Categorical
○ Hi	○ Numerical
○ S_Income	○ Numerical
○ E_Income	○ Numerical
○ S_Pop	○ Numerical
○ E_Pop	○ Numerical
○ Slot	○ Categorical
○ Gate	○ Categorical
○ Dsitance	○ Numerical
○ Pax	○ Numerical
○ Fare	○ Predictive Variable

  ○

- In the linear regression model, we start by deciding if we need to sample the existing data set under the study. In this case of the project, the database provided to us has 638 rows of input, so the decision was made based on the size of the dataset as to not sample the data into a smaller segment.
- Moving forward the transform data tool was used to perform imputation on the existing data to make sure there are no missing variables. As a result, it was found that there were 0 missing data points.
- Following the previous step, data partitioning was done to segregate the data into training, validation and testing partitions so as to enable a better performance of the predictive model. The partition was made as 50% of the data for training, 30% of the data for validation and 20% of the data for testing.
- Upon running the linear regression model, using feature selection was found that the RMSE for validation partition was 37.21 which was decent, and R2 at 76% which also indicated that this model is capable of predicting roughly 76% of the variation caused by the regression.

Validation: Prediction Summary

Metric	Value
SSE	264542
MSE	1385.037
RMSE	37.21608
MAD	28.75873
R2	0.76086

- The values for the test partition were observed to be ambiguous for scaled and re-scaled data from which a clear conclusion was not being able to be made. This might be a sheer stance of overfitting model where the data performs well in training partition and decently in validation partition but when it comes to test partition, it fails to create a conclusive result.
- A point to be noted was that the subset SW(No) was important in the analysis after feature selection which gave us the answer to the question posed in the problem statement that yes, including the presence or absence of Southwest Airlines impacted the predictive model.

2. K-Nearest Neighbor Model

- To perform the kNN predictive model, we went all the way back to the partition that was made in the initial section of the problem solving. In the beginning of this model, typecasting is necessary as well but since we have already performed that step in the prior section of the run-through, it is not necessary to re-do it.
- Now, in the first step using all the numerical variables as inputs, and using K = 10, and not re-scaling the data, we found out that the RMSE value coming out at 60.83 and the R2 at roughly 36%.

Validation: Prediction Summary

Metric	Value
SSE	706915.8
MSE	3701.13
RMSE	60.83691
MAD	41.54586
R2	0.360965

- In the next step, we re-scaled the data by standardization. Keeping the variables and the K as same as before, we got a slight variation in the results as compared to the ones obtained previously. The RMSE stood at 38.72 which is a stark diffference but the R2 went on to be roughly 74%.

Validation: Prediction Summary

Metric	Value
SSE	286429.9
MSE	1499.633
RMSE	38.72509
MAD	27.07744
R2	0.741074

- Following the previous step, this time the data was re-scaled by normalization and we again got some different results as compared to the previous one. In this case, the K to be used is 8 (which ensures there is no overfitting), the RMSE for the validation partition is found out to be 37.93 and the R2 stood at 75%. This observation suggested that in this case, while performing kNN prediction, using Normalization is the better approach to re-scale the data.

Validation: Prediction Summary

Metric	Value
SSE	274817.8
MSE	1438.837
RMSE	37.932
MAD	26.02017
R2	0.751571

- In conclusion to this section, we found out that given the data set, to run a kNN prediction model to get optimum output on the validation partition, Normalization should be used with K = 8.
- We also performed predictions based on the test partition of the model, but since the sample size is relatively smaller, there were only 128 entries in the test partition which is again a smaller number. Owing to which, the test partition was again inconclusive and inconsistent with the results depicted.
- The following are the observations from the test partition -:
- RMSE (Test) = 58.81296 (Regular), R2 = 0.401495
- RMSE (Test) = 36.94121 (Standardization), R2 = 0.763874
- RMSE (Test) = 37.932 (Normalization), R2 = 0.751571

3. CART Analysis

- In case of CART analysis, we attempted multiple iterations including limits and nodes
- Iterations were performed starting from limits 6 till limit 100 and it was observed that RMSE became constant at 42.34 from level 7 onwards.
- Increasing and decreasing of nodes did not effectively change the RMSE throughout the entire operational tenure.

- The best case scenario from the CART analysis is depicted below -:

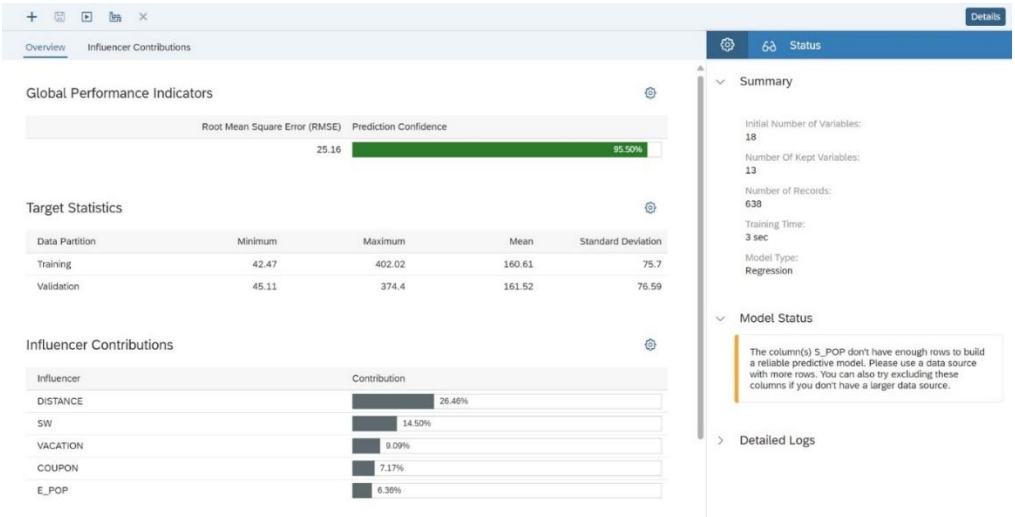
Validation: Prediction Summary

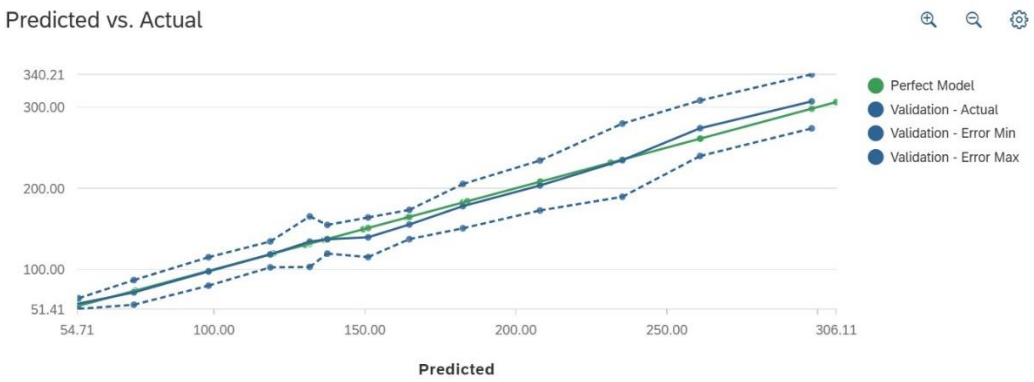
Metric	Value
SSE	342439.5
MSE	1792.877
RMSE	42.34238
MAD	29.38554
R2	0.690443

- We also tried re-scaling the data both ways by standardizing and normalizing but it also did not pose any significant change in the RMSE or R2.

Regression Using SAP Analytics Cloud

- Linear Regression model was created using SAC in which we found a number of defining outputs.
- The RMSE for the validation partition comes out to be at 25.16 which is better as compared to the RMSE which we got for the entirity of the Analytics Solver.
- The prediction confidence for the SAC stood at 95.50%
- The initial number of variables were 18, and the number of variables kept were 13. The variables which were excluded were Slot, New, E-Code and S-City. These variables were potentially rejected because they did not impact the analysis to any extent.
- The pictorial depictions are as follows -:





Influencer Contributions

Influencer	Contribution
DISTANCE	26.46%
SW	14.50%
VACATION	9.09%
COUPON	7.17%
E_POP	6.36%
E_INCOME	6.23%
HI	6.15%
PAX	6.06%
S_POP	5.50%
S_INCOME	4.92%
S_CODE	3.32%
GATE	2.38%
E_CITY	1.85%

- In this case too, the inclusion of Southwest Airlines poses a significant impact towards the analysis of the regression model.

Linear Regression Using R Programming

(Partitioning of the train-test model was done into 50-50%)

- The number of variables being in play for the R programming part was as follows -:

	Df	Sum of Sq	RSS	AIC
<none>			365330	2270.8
- S_INCOME	1	3533	368863	2271.9
- S_POP	1	9351	374681	2276.9
- E_POP	1	10999	376329	2278.3
- E_INCOME	1	11855	377186	2279.0
- SLOT	1	12790	378121	2279.8
- GATE	1	16673	382003	2283.1
- PAX	1	31749	397080	2295.4
- HI	1	34399	399729	2297.5
- VACATION	1	60698	426029	2317.9
- SW	1	71284	436615	2325.7
- DISTANCE	1	433714	799044	2518.5

- In this case, the performance of both RMSE and MAPE are above benchmark level as depicted in the Code file. This suggests that this is a good model.

mape	0.214400816842479
mape_bench	0.502396054037328
rmse	36.8350340738056
rmse_bench	78.2690586698939

- Here, the RMSE is 36.83 which is better than the RMSE of Analytics Solver but lags behind as compared to the SAC.

Conclusion Comparison

We have progressively ran models in all the three platforms including Analytics Solver, SAC and R Programming. In conclusion to which we found out that the RMSE in linear regression for analytic solver came up to 37.93. Wheras for the SAC the regression RMSE was 25.16 similarly in case of R programming the regression RMSE came out as 36.84. From this we can conlcude that for a given data set of this type, running a linear regression model using SAP Analytics Cloud is more beneficial as comapared to R and Excel Analytic Solver. Having said that, we do not imply that the models created in R and Excel are of lower quality or higher overfitting. It just comes to this part where for this data set, SAC comes out to be on the top.

Now, answering the question posed in the problem statement as to whether Southwest Airlines's presence or absence effects the fare or not, our observations suggest that in both R and Analytic Solver the coefficients obtained from the model are negative and imply that the presence of Southwest Airline in that route tends to decrease the airfare.