Airfares

a) Explore the numerical predictors and response (FARE) by creating a correlation table and examining some scatterplots between FARE and those predictors. What seems to be the best single predictor of FARE?

Distance seems to be the best single predictor of FARE from the Correlation Table. By looking at the scatter plot, with a change in distance we can see a positive change in FARE. The second best predictor is coupon which has second highest magnitude on the correlation table and scatter plot. The other numerical predictors have scatter plots distributed randomly which doesn't give sufficient evidence that it correlates with FARE. We can refer the table below from sheet correlation.

	COUPON	*NEW*	*HI*	*S_INCOME*	*E_INCOME*	*S_POP*	*E_POP*	*DISTANCE*	*PAX*	*FARE*
COUPON	1									
NEW	0.020223	1								
HI	-0.34725	0.054147	1							
S_INCOME	-0.0884	0.026597	-0.02738	1						
E_INCOME	0.046889	0.113377	0.082393	-0.13886	1					
S_POP	-0.10776	-0.01667	-0.1725	0.517187	-0.14406	1				
E_POP	0.09497	0.058568	-0.06246	-0.27228	0.458418	-0.28014	1			
DISTANCE	0.746805	0.080965	-0.31237	0.028153	0.176531	0.018437	0.11564	1		
PAX	-0.33697	0.010495	-0.16896	0.138197	0.259961	0.284611	0.314698	-0.10248	1	
FARE	0.496537	0.09173	0.025195	0.209135	0.326092	0.145097	0.285043	0.670016	-0.09071	1

b) Explore the categorical predictors (excluding the first four) by computing the percentage of flights in each category. Create a pivot table with the average fare in each category. Which categorical predictor seems best for predicting FARE?

	Column Labels		
	Controlled	Free	Grand Total
Count of SLOT	182	456	638
% of Flights	28.53%	71.47%	100.00%

Row Labels	Average of FARE
Controlled	186.0593956
Free	150.8256798
Grand Total	160.8766771

	Column Labels		
	No	Yes	Grand Total
Count of SW	444	194	638
% of Flights	69.59%	30.41%	100.00%

Row Labels	Average of FARE
No	188.1827928
Yes	98.38226804
Grand Total	160.8766771

	Column Labels		
	Constrained	Free	Grand Total
Count of GATE	124	514	638
% of Flights	19.44%	80.56%	100.00%

Row Labels	Average of FARE
Constrained	193.1290323
Free	153.0959533
Grand Total	160.8766771

	Column Labels		
	No	Yes	Grand Total
Count of VACATION	468	170	638
% of Flights	73.35%	26.65%	100.00%

Row Labels	Average of FARE
No	173.5525
Yes	125.9808824
Grand Total	160.8766771

From the above pivot tables we can say that South West operates 30.41% of total flights operated and the average fare for a South West operated flight is around 98$ which is least average fare when compared to all other categorical predictors. So, SW column best predicts the fare of the flight in the categorical predictors. This table can be referred from the pivot table.

c) Find a model for predicting the average fare on a new route:

i. Convert categorical variables (e.g., SW) into dummy variables. Then partition the data into training and validation sets. The model will be fit to the training data and evaluated on the validation set.

the

Inputs

Data	
Workbook	Airfares.xlsx
Worksheet	Encoding
Range	$C$24:$Y$662
# Records in the input data	638

Variables					
# Selected Variables	18				
Selected Variables	COUPON	NEW	HI	S_INCOME	E_IN

Partitioning Parameters	
Partitioning type	RANDOM
Random seed	12345
Ratio - Training	0.6
Ratio - Validation	0.4

Partition Summary

Partition	# Records
Training	383
Validation	255

This table can be referred from Encoding sheet.

ii. Why should the data be partitioned into training, and validation? What will the training set be used for? What will the validation set be used for?

Partitioning data into training and validation datasets allows the user to develop highly accurate models that helps not only analyzing the data which is there but also helps to collect the future data. Usually the data is partitioned into 60% training data and 40% validation data. Training is subset of the data which is used to learn the relationships between data and outcome value. Validation is also a subset of data which is run on the data to estimate the performance of the model after the training has been done on the data. This can be referred from the STD partition sheet.

iii. Use stepwise regression to reduce the number of predictors. You can ignore the first four predictors (S CODE, S CITY, E CODE, E CITY). Report the estimated model selected.

Regression Summary

Metric	Value
Residual DF	369
R2	0.785567445
Adjusted R2	0.778012911
Std. Error Estim	35.92391229
RSS	476204.638

Predictor Screening

Predictor	Criteria	Included
Intercept	1.53E-15	FALSE
COUPON	2.877009329	TRUE
NEW	15.83731177	TRUE
HI	32251.39024	TRUE
S_INCOME	99174.25422	TRUE
E_INCOME	219391.7811	TRUE
S_POP	106685127.2	TRUE
E_POP	69411958.24	TRUE
DISTANCE	11519.3443	TRUE
PAX	215642.951	TRUE
VACATION_No	5.23769E-15	FALSE
VACATION_Ye	8.149752609	TRUE
SW_No	1.01429635	TRUE
SW_Yes	8.779115831	TRUE
SLOT_Controlle	3.95078E-15	FALSE
SLOT_Free	8.020632568	TRUE
GATE_Constrai	2.93261E-15	FALSE
GATE_Free	6.943145248	TRUE

Tolerance for e	9.07264E-06

Coefficients

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	0	0	0	0	N/A	N/A
COUPON	10.24575172	-21.54811806	42.03962149	16.1684543	0.633687768	0.526677195
NEW	-1.75953839	-6.356993349	2.837916569	2.33798971	-0.752586028	0.452178542
HI	0.009122908	0.006570801	0.011675014	0.001297848	7.029257093	1.01014E-11
S_INCOME	0.001254741	-4.88504E-05	0.002558332	0.000662928	1.892724869	0.059176837
E_INCOME	0.001846147	0.00089259	0.002799704	0.000484922	3.807101403	0.000164596
S_POP	4.41605E-06	2.77757E-06	6.05453E-06	8.33233E-07	5.299897973	1.99832E-07
E_POP	4.31136E-06	2.37782E-06	6.2449E-06	9.83282E-07	4.384666615	1.51709E-05
DISTANCE	0.074125826	0.065057053	0.083194599	0.004611834	16.07296164	1.93279E-44
PAX	-0.000893952	-0.001281595	-0.00050631	0.000197132	-4.534799904	7.80584E-06
VACATION_No	0	0	0	0	N/A	N/A
VACATION_Ye	-33.74914644	-43.11228096	-24.38601192	4.761528358	-7.087881014	6.96628E-12
SW_No	-23.21350588	-92.85910078	46.43208901	35.4175703	-0.655423443	0.512603565
SW_Yes	-58.32119611	-125.2494242	8.607032006	34.03568062	-1.713531066	0.087454746
SLOT_Controlle	0	0	0	0	N/A	N/A
SLOT_Free	-17.45963666	-27.33574137	-7.583531945	5.022394215	-3.476357273	0.000568827
GATE_Constrai	0	0	0	0	N/A	N/A
GATE_Free	-17.39758297	-27.89963716	-6.895528783	5.340714557	-3.257538441	0.001228019

Feature Selection

Best Subsets

Subset ID	Intercept	COUPON	NEW	HI	S_INCOME	E_INCOME	S_POP	E_POP	DISTANCE	PAX	VACATION_No	VACATION_Yes	SW_No	SW_Yes	SLOT_Controlled	SLOT_Free	GATE_Constrained	GATE_Free
Subset 1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Subset 2	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Subset 3	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
Subset 4	0	1	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0
Subset 5	0	1	0	1	0	0	0	0	1	0	0	1	0	1	0	0	0	0
Subset 6	0	1	0	1	0	0	1	0	1	0	0	1	0	1	0	0	0	0
Subset 7	0	1	0	1	0	0	1	1	1	0	0	1	0	1	0	0	0	0
Subset 8	0	1	0	1	0	0	1	1	1	1	0	1	0	1	0	0	0	0
Subset 9	0	1	0	1	0	1	1	1	1	1	0	1	0	1	0	0	0	0
Subset 10	0	1	0	1	0	1	1	1	1	1	0	1	0	1	0	1	0	0
Subset 11	0	1	0	1	0	1	1	1	1	1	0	1	0	1	0	1	0	1
Subset 12	0	0	0	1	0	1	1	1	1	1	0	1	0	1	0	1	0	1
Subset 13	0	0	0	1	1	1	1	1	1	1	0	1	0	1	0	1	0	1

Best Subsets Details

Subset ID	#Coefficients	RSS	Mallow's Cp	R2	Adjusted R2	Probability
Subset 13	11	477771.0628	9.213786447	0.78486209	0.779078815	0.749784072
Subset 11	11	482081.4	12.55374061	0.782921	0.777085732	0.209454471
Subset 12	10	484604.3895	12.50877383	0.78178508	0.77651984	0.166724076
Subset 10	10	495258.116	20.76410101	0.77698776	0.771606767	0.005819331
Subset 9	9	507772.241	28.46100804	0.77135271	0.766461861	0.000239298
Subset 8	8	526164.153	40.71247683	0.76307093	0.75864825	1.75221E-06
Subset 7	7	543402.9711	52.07043972	0.75530837	0.75140372	2.23341E-08
Subset 6	6	592082.8967	87.79139226	0.73338805	0.72985208	3.06367E-14
Subset 5	5	635796.2919	119.6638948	0.71370413	0.710674539	4.44793E-19
Subset 4	4	774907.1993	225.4577312	0.65106318	0.64830115	1.12847E-33
Subset 3	3	990311.191	390.3693204	0.55406785	0.551720837	3.70364E-52
Subset 2	2	1225068.302	570.2771957	0.4483579	0.446910022	3.24941E-68
Subset 1	1	1682579.844	922.7923461	0.24234275	0.242342753	1.30673E-92

The above images can be referred from sheets ending ‘STEPWISE’

The stepwise regression is done we have the results as shown above. Here we need to choose the best model from the available subsets based on the values of R^2, Adj. R^2 and Mallow’s Cp value(#predictor +1) . From the above picture we can say that we have 13 predictors and the best subset is we found is 11th subset. Using

Regression Summary

Metric	Value
Residual DF	371
R2	0.783198061
Adjusted R2	0.776769971
Std. Error	36.0243436
RSS	481466.4861

the subset we generate the stepwise regression.

Predictor Screening

Predictor	Criteria	Included
Intercept	1.413624255	TRUE
COUPON	3.454731617	TRUE
HI	33486.08864	TRUE
E_INCOME	219391.7811	TRUE
S_POP	106685127.2	TRUE
E_POP	69411958.24	TRUE
DISTANCE	11654.40403	TRUE
PAX	215642.951	TRUE
VACATION_Yes	8.177527008	TRUE
SW_Yes	8.912910299	TRUE
SLOT_Free	8.091224105	TRUE
GATE_Free	7.2470046	TRUE

Tolerance for e	9.07264E-06

Coefficients

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	17.54126708	-32.56929476	67.65182893	25.48367678	0.688333447	0.491672994
COUPON	6.134342987	-25.440304	37.70898998	16.05725557	0.382029355	0.702658322
HI	0.009110268	0.006551273	0.011669262	0.001301374	7.000497671	1.2018E-11
E_INCOME	0.001690551	0.000746929	0.002634174	0.000479878	3.522874609	0.000480158
S_POP	4.74243E-06	3.13197E-06	6.35289E-06	8.18999E-07	5.790521138	1.49673E-08
E_POP	3.80038E-06	1.94407E-06	5.6567E-06	9.44027E-07	4.02571463	6.88897E-05
DISTANCE	0.075100447	0.06616863	0.084039031	0.004545708	16.52117686	2.39141E-46
PAX	-0.000854001	-0.00124043	-0.000467571	0.000196518	-4.345654137	1.79521E-05
VACATION_Yes	-35.96542866	-45.02843996	-26.90241735	4.608985455	-7.803328738	6.19899E-14
SW_Yes	-38.1736666	-47.31314188	-29.03419131	4.647871137	-8.213150811	3.62111E-15
SLOT_Free	-19.08883186	-28.78608509	-9.391578635	4.931528561	-3.870773864	0.000128168
GATE_Free	-17.44426557	-27.96703789	-6.921493252	5.35134548	-3.259790577	0.00121801

The above pictures can be referred from sheets ending 'STEPSS'
The stepwise regression was calculated to predict the fare based on the given input variables. The regression equation was found to be

FARE=17.54126708+6.134342987*COUPON+0.009110268*HI+0.001690551*E_INCOME+4.74243E-06*S_POP+3.80038E-06*E_POP+0.075100447*DISTANCE–0.000854001*PAX–35.96542866*VACATION_Yes-38.1736666*SW_Yes-19.08883186*SLOT_Free–17.44426557*GATE_Free

iv. Repeat (iii) using exhaustive search instead of stepwise regression. Compare the resulting best model to the one you obtained in (iii) in terms of the predictors that are in the model.

Regression Summary

Metric	Value
Residual DF	370
R2	0.785238308
Adjusted R2	0.778273064
Std. Error	35.90285596
RSS	476935.5744

Predictor Screening

Predictor	Criteria	Included
Intercept	1.53278E-15	FALSE
COUPON	2.877009329	TRUE
NEW	15.83731177	TRUE
HI	32251.39024	TRUE
S_INCOME	99174.25422	TRUE
E_INCOME	219391.7811	TRUE
S_POP	106685127.2	TRUE
E_POP	69411958.24	TRUE
DISTANCE	11519.3443	TRUE
PAX	215642.951	TRUE
VACATION_No	5.23769E-15	FALSE
VACATION_Yes	8.149752609	TRUE
SW_No	1.01429635	TRUE
SW_Yes	8.779115831	TRUE
SLOT_Controlle	3.95078E-15	FALSE
SLOT_Free	8.020632568	TRUE
GATE_Constrain	2.93261E-15	FALSE
GATE_Free	6.943145248	TRUE

Tolerance for en	9.07264E-06

Coefficients

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	0	0	0	0	N/A	N/A
COUPON	10.24575172	-21.54811806	42.03962149	16.1684543	0.633687768	0.526677195
NEW	-1.75953839	-6.356993349	2.837916569	2.33798971	-0.752586028	0.452178542
HI	0.009122908	0.006570801	0.011675014	0.001297848	7.029257093	1.01014E-11
S_INCOME	0.001254741	-4.88504E-05	0.002558332	0.000662928	1.892724869	0.059176837
E_INCOME	0.001846147	0.00089259	0.002799704	0.000484922	3.807101403	0.000164596
S_POP	4.41605E-06	2.77757E-06	6.05453E-06	8.33233E-07	5.299897973	1.99832E-07
E_POP	4.31136E-06	2.37782E-06	6.2449E-06	9.83282E-07	4.384666615	1.51709E-05
DISTANCE	0.074125826	0.065057053	0.083194599	0.004611834	16.07296164	1.93279E-44
PAX	-0.000893952	-0.001281595	-0.00050631	0.000197132	-4.534799904	7.80584E-06
VACATION_No	0	0	0	0	N/A	N/A
VACATION_Yes	-33.74914644	-43.11228096	-24.38601192	4.761528358	-7.087881014	6.96628E-12
SW_No	-23.21350588	-92.85910078	46.43208901	35.4175703	-0.655423443	0.512603565
SW_Yes	-58.32119611	-125.2494242	8.607032006	34.03568062	-1.713531066	0.087454746
SLOT_Controlle	0	0	0	0	N/A	N/A
SLOT_Free	-17.45963666	-27.33574137	-7.583531945	5.022394215	-3.476357273	0.000568827
GATE_Constrain	0	0	0	0	N/A	N/A
GATE_Free	-17.39758297	-27.89963716	-6.895528783	5.340714557	-3.257538441	0.001228019

The exhaustive search regression is done on the partition data gives the values of subsets with Max. CP, R^2 and Adj R^2 values. The best subset is chosen on the basis of high adj R^2 value and Max. CP value Which is one more than the number of predictor. The above picture can be referred from the sheets ending with ‘Exhaustive’.

Best Subsets Details

Subset ID	#Coefficients	RSS	Mallows's Cp	R2	Adjusted R2	Probability
Subset 12	12	477731.47	11.1831072	0.78488	0.778501697	0.55399075
Subset 13	13	476759.02	12.42957989	0.785318	0.778355143	0.512603565
Subset 14	14	476204.64	14	0.785567	0.778012911	0
Subset 11	11	482081.37	12.55374061	0.782921	0.777085732	0.209454471
Subset 10	10	495258.12	20.76410101	0.776988	0.771606767	0.005819331
Subset 9	9	507772.24	28.46100804	0.771353	0.766461861	0.000239298
Subset 8	8	526164.15	40.71247683	0.763071	0.75864825	1.75221E-06
Subset 7	7	543402.97	52.07043972	0.755308	0.75140372	2.23341E-08
Subset 6	6	577407.54	76.41979667	0.739996	0.736547962	2.27957E-12
Subset 5	5	617365.94	105.3826434	0.722003	0.719061453	7.35628E-17
Subset 4	4	700714.43	167.9674629	0.684472	0.68197423	6.31806E-26
Subset 3	3	928232.97	342.2663383	0.582021	0.579821471	4.28717E-47
Subset 2	2	1225068.3	570.2771957	0.448358	0.446910022	3.24941E-68
Subset 1	1	1682579.8	922.7923461	0.242343	0.242342753	1.30673E-92

From the above table we have to choose the best subset and further run regression on that subset to get the summary of that model. Here we select subset 14 as it has high Adjusted R^2 value and also the Mallow's Cp value is 1+ predictor value(14+1 =15). By looking at the regression summaries of both the models we can say that, they have very slight differences in their values and both of them are considered to be good. The model with high Adjusted R^2 value and low Std. Error is considered to be the better among the models. By looking at the values we can say that Exhaustive search regression model is slightly better than stepwise as they have higher R2 values.

Regression Summary

Metric	Value
Residual DF	369
R2	0.785567445
Adjusted R2	0.778012911
Std. Error	35.92391229
RSS	476204.638

Predictor Screening

Predictor	Criteria	Included
Intercept	3.10765E-15	FALSE
COUPON	2.877009329	TRUE
NEW	15.83731177	TRUE
HI	32251.39024	TRUE
S_INCOME	99174.25422	TRUE
E_INCOME	219391.7811	TRUE
S_POP	106685127.2	TRUE
E_POP	69411958.24	TRUE
DISTANCE	11519.3443	TRUE
PAX	215642.951	TRUE
VACATION_Ye	8.149752609	TRUE
SW_No	1.01429635	TRUE
SW_Yes	8.779115831	TRUE
SLOT_Free	8.020632568	TRUE
GATE_Free	6.943145248	TRUE

Tolerance f	9.07264E-06

Coefficients

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	0	0	0	0	N/A	N/A
COUPON	10.24575172	-21.54811806	42.03962149	16.1684543	0.633687768	0.526677195
NEW	-1.75953839	-6.356993349	2.837916569	2.33798971	-0.752586028	0.452178542
HI	0.009122908	0.006570801	0.011675014	0.001297848	7.029257093	1.01014E-11
S_INCOME	0.001254741	-4.88504E-05	0.002558332	0.000662928	1.892724869	0.059176837
E_INCOME	0.001846147	0.00089259	0.002799704	0.000484922	3.807101403	0.000164596
S_POP	4.41605E-06	2.77757E-06	6.05453E-06	8.33233E-07	5.299897973	1.99832E-07
E_POP	4.31136E-06	2.37782E-06	6.2449E-06	9.83282E-07	4.384666615	1.51709E-05
DISTANCE	0.074125826	0.065057053	0.083194599	0.004611834	16.07296164	1.93279E-44
PAX	-0.000893952	-0.001281595	-0.00050631	0.000197132	-4.534799904	7.80584E-06
VACATION_Ye	-33.74914644	-43.11228096	-24.38601192	4.761528358	-7.087881014	6.96628E-12
SW_No	-23.21350588	-92.85910078	46.43208901	35.4175703	-0.655423443	0.512603565
SW_Yes	-58.32119611	-125.2494242	8.607032006	34.03568062	-1.713531066	0.087454746
SLOT_Free	-17.45963666	-27.33574137	-7.583531945	5.022394215	-3.476357273	0.000568827
GATE_Free	-17.39758297	-27.89963716	-6.895528783	5.340714557	-3.257538441	0.001228019

The above pictures can be referred from sheets ending with 'EXSS'.

The regression equation is

Fare=10.24575172*COUPON+1.75953839*NEW+0.009122908*HI+0.001254741*S_INCOME+0.001846147*E_INCOME+4.41605E-06*S_POP+3.31136E-06*E_POP+0.074125826*DISTANCE–0.000893952*PAX–33.74914644*VACATION_Yes-23.21350588*SW_No-58.32119611*SW_Yes-17.45963666*SLOT_Free – 17.39758297*GATE_Free

v. Compare the predictive accuracy of both models (iii) and (iv) using measures such as RMSE and average error and lift charts.

Exhaustive

Training: Prediction Summary

Metric	Value
SSE	476204.6
MSE	1243.354
RMSE	35.26123
MAD	27.59104
R2	0.785567

Validation: Prediction Summary

Metric	Value
SSE	320573.7
MSE	1257.152
RMSE	35.45633
MAD	27.79736
R2	0.780519

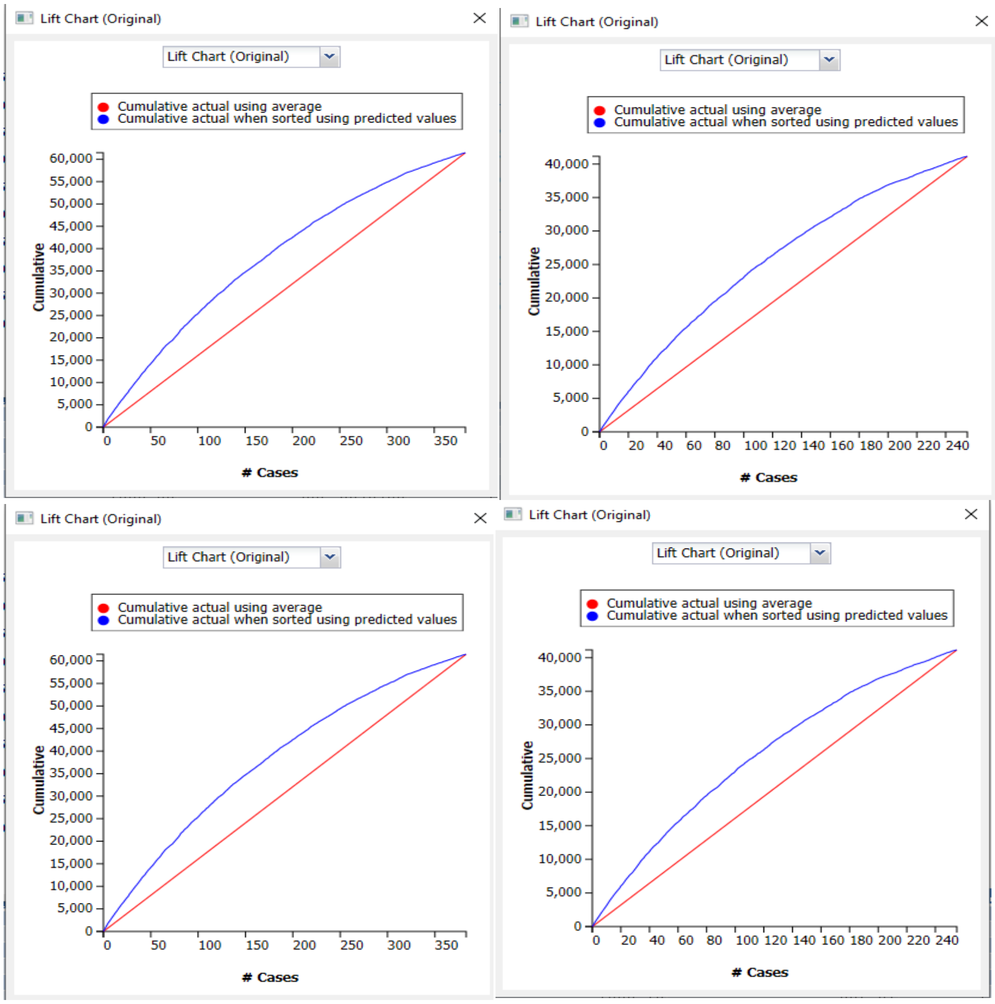
Stepwise

Training: Prediction Summary

Metric	Value
SSE	481466.5
MSE	1257.093
RMSE	35.4555
MAD	27.75791
R2	0.783198

Validation: Prediction Summary

Metric	Value
SSE	323795.2
MSE	1269.785
RMSE	35.63405
MAD	27.78419
R2	0.778314



After comparing both the models stepwise and exhaustive across RMSE, average error and lift charts we can say that there is no significant difference between the values of the given measures. We can say that Exhaustive models have slightly better values when compared to stepwise when we see at the RMSE value. Also from regression summary, the values are almost same when we see at the standard error values. The lift charts also do not have any significant difference in their values. So it is difficult to find the best among the models.

vi. Using model (iv), 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 = $ 28,760, E INCOME = $ 27,664, S POP = 4,557,004, E POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12,782, DISTANCE = 1976 miles.

Equation after substituting the given values -

Fare=10.24575172*1.202–1.75953839*3+0.009122908*4442.141+0.001254741*28760+ 0.001846147*27664+4.41605*0.000001*4557004+3.31136*0.000001*3195503+ 0.074125826*1976–0.000893952*12782–33.74914644*0-23.21350588*1-58.32119611*0-17.45963666*1–17.39758297*1 = 242.40

vii. Using model (iv), predict the reduction in average fare on the route in (vi) if Southwest decides to cover this route.

Equation after substituting the given values -

Fare=10.24575172*1.202–1.75953839*3+0.009122908*4442.141+0.001254741*28760+ 0.001846147*27664+4.41605*0.000001*4557004+3.31136*0.000001*3195503+ 0.074125826*1976–0.000893952*12782–33.74914644*0-23.21350588*0-58.32119611*1-17.45963666*1 –17.39758297*1 = 207.29

viii. In reality, which of the factors will not be available for predicting the average fare from a new airport (i.e., before flights start operating on those routes)? Which ones can be estimated? How?

Southwest airlines have nominal effect on the average fare of flights as there used to be a regulation fee for the new flights. Also, Southwest airlines are starting service on new routes, so it Southwest Airlines shouldn’t be included in the dataset. Moreover we are not aware of SW airlines future price. so, it won’t be available.

Factors that can be estimated –
HI – Herfindal index can be estimated from the historical data. It is the measure of market concentration. Which fluctuates

Other factors –
We can get the POP and Income from the census bureau after the routes are selected.
Vacation and distance will be know after we know the route
Slot and gate will the airport’s designing.
PAX – we can estimate this from the population of the city

ix. Select a model that includes only factors that are available before flights begin to operate on the new route. Use an exhaustive search to find such a model.

Regression Summary

Metric	Value
Residual DF	370
R2	0.755767107
Adjusted R2	0.74784604
Std. Error Estima	38.287117
RSS	542384.2314

Predictor Screening

Predictor	Criteria	Included
Intercept	1.70079E-15	FALSE
COUPON	2.913425572	TRUE
NEW	15.83731177	TRUE
HI	32251.39024	TRUE
S_INCOME	99174.25422	TRUE
E_INCOME	219391.7811	TRUE
S_POP	106685127.2	TRUE
E_POP	69411958.24	TRUE
DISTANCE	11519.3443	TRUE
PAX	215642.951	TRUE
VACATION_No	1.080891882	TRUE
VACATION_Yes	8.09539647	TRUE
SLOT_Controlle	6.06634E-15	FALSE
SLOT_Free	8.448433171	TRUE
GATE_Constrain	4.35641E-15	FALSE
GATE_Free	7.367259056	TRUE

Tolerance for en	9.07264E-06

Coefficients

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	0	0	0	0	N/A	N/A
COUPON	22.44003269	-11.25659684	56.13666221	17.13623612	1.309507673	0.191174995
NEW	-2.116371823	-7.015107409	2.782363764	2.491225113	-0.849530543	0.396135481
HI	0.010648161	0.00796507	0.013331251	0.001364471	7.803874215	6.21128E-14
S_INCOME	0.002800674	0.00148707	0.004114278	0.000668026	4.192462474	3.45606E-05
E_INCOME	0.002488071	0.001489309	0.003486832	0.000507915	4.898599842	1.44562E-06
S_POP	4.56963E-06	2.82396E-06	6.3153E-06	8.87751E-07	5.14741899	4.29759E-07
E_POP	5.41197E-06	3.37659E-06	7.44736E-06	1.03508E-06	5.228531639	2.86392E-07
DISTANCE	0.076491667	0.066851236	0.086132099	0.004902589	15.60230158	1.57412E-42
PAX	-0.001007935	-0.001419726	-0.000596145	0.000209414	-4.813132536	2.16841E-06
VACATION_No	-110.8629158	-180.516174	-41.20965772	35.42178237	-3.129794957	0.001888199
VACATION_Yes	-138.7123701	-206.5146333	-70.91010695	34.48046904	-4.022925847	6.97114E-05
SLOT_Controlle	0	0	0	0	N/A	N/A
SLOT_Free	-22.98601604	-33.38671291	-12.58531918	5.289217343	-4.345825582	1.79508E-05
GATE_Constrain	0	0	0	0	N/A	N/A
GATE_Free	-25.55872353	-36.49373934	-14.62370773	5.56094231	-4.596113771	5.9146E-06

Best Subsets Details

Subset ID	#Coefficients	RSS	Mallows's Cp	R2	Adjusted R2	Probability
Subset 12	12	543442.18	11.72170214	0.755291	0.748035188	0.396135481
Subset 13	13	542384.23	13	0.755767	0.74784604	0
Subset 11	11	559733.56	20.83525041	0.747955	0.74117938	0.002952945
Subset 10	10	570019.44	25.85199754	0.743323	0.737129836	0.000361621
Subset 9	9	597376.17	42.51402647	0.731005	0.7252506	3.14014E-07
Subset 8	8	615146.99	52.63680411	0.723002	0.717831781	6.34519E-09
Subset 7	7	636784.13	65.3970803	0.713259	0.708683659	5.21599E-11
Subset 6	6	664772.17	82.48977187	0.700656	0.696686361	9.9442E-14
Subset 5	5	738450.39	130.7510851	0.66748	0.663960765	3.43937E-21
Subset 4	4	845498.44	201.7763952	0.619276	0.616262659	4.7228E-31
Subset 3	3	951351.39	271.9864435	0.571611	0.569356574	1.29227E-39
Subset 2	2	1248828.7	472.9175125	0.437659	0.436182744	2.51581E-60
Subset 1	1	1682579.8	766.8109172	0.242343	0.242342753	3.1149E-83

We remove the column SW from the Data and perform exhaustive search and get the results as shown above. Now we have to choose the best subset and do the regression again to get the best fit model. We can see that subset 13 has better adj R^2 and Mallow's Cp value(13) which are better than any other subset. We choose and subset and do the regression again to get the following results. The above pictures can be referred from sheets 'BNEX'.

Regression Summary

Metric	Value
Residual DF	370
R2	0.755767107
Adjusted R2	0.74784604
Std. Error Estima	38.287117
RSS	542384.2314

Predictor Screening

Predictor	Criteria	Included
Intercept	1.70079E-15	FALSE
COUPON	2.913425572	TRUE
NEW	15.83731177	TRUE
HI	32251.39024	TRUE
S_INCOME	99174.25422	TRUE
E_INCOME	219391.7811	TRUE
S_POP	106685127.2	TRUE
E_POP	69411958.24	TRUE
DISTANCE	11519.3443	TRUE
PAX	215642.951	TRUE
VACATION_No	1.080891882	TRUE
VACATION_Yes	8.09539647	TRUE
SLOT_Controlle	6.06634E-15	FALSE
SLOT_Free	8.448433171	TRUE
GATE_Constrain	4.35641E-15	FALSE
GATE_Free	7.367259056	TRUE

Tolerance for en	9.07264E-06

Coefficients

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	0	0	0	0	N/A	N/A
COUPON	22.44003269	-11.25659684	56.13666221	17.13623612	1.309507673	0.191174995
NEW	-2.116371823	-7.015107409	2.782363764	2.491225113	-0.849530543	0.396135481
HI	0.010648161	0.00796507	0.013331251	0.001364471	7.803874215	6.21128E-14
S_INCOME	0.002800674	0.00148707	0.004114278	0.000668026	4.192462474	3.45606E-05
E_INCOME	0.002488071	0.001489309	0.003486832	0.000507915	4.898599842	1.44562E-06
S_POP	4.56963E-06	2.82396E-06	6.3153E-06	8.87751E-07	5.14741899	4.29759E-07
E_POP	5.41197E-06	3.37659E-06	7.44736E-06	1.03508E-06	5.228531639	2.86392E-07
DISTANCE	0.076491667	0.066851236	0.086132099	0.004902589	15.60230158	1.57412E-42
PAX	-0.001007935	-0.001419726	-0.000596145	0.000209414	-4.813132536	2.16841E-06
VACATION_No	-110.8629158	-180.516174	-41.20965772	35.42178237	-3.129794957	0.001888199
VACATION_Yes	-138.7123701	-206.5146333	-70.91010695	34.48046904	-4.022925847	6.97114E-05
SLOT_Controlle	0	0	0	0	N/A	N/A
SLOT_Free	-22.98601604	-33.38671291	-12.58531918	5.289217343	-4.345825582	1.79508E-05
GATE_Constrain	0	0	0	0	N/A	N/A
GATE_Free	-25.55872353	-36.49373934	-14.62370773	5.56094231	-4.596113771	5.9146E-06

The regression model formed after removing the factors before flights begin to operate on new airports is:

FARE=22.44003269*COUPON-2.116371823*New+0.010648161*HI+0.002800674*S_INCOME+0.002488071*E_INCOME+4.56963E-06*S_POP+5.41197E-06*E_POP+0.076491667*DISTANCE-0.001007935*PAX-110.8629158*VACATION_No-138.7123701*VACATION_Yes-22.98601604*SLOT_Free -25.55872353*GATE_Free

Using the exhaustive search similar to (iv) we get the above equation.

The above pictures can be referred from the sheets 'BNEXSS'.

x. Use the model in (ix) to predict the average fare on a route with characteristics COUPON = 1.202, NEW = 3, VACATION = No, SW = No, HI = 4442.141, S INCOME = $ 28,760, E INCOME = $ 27,664, S POP = 4,557,004, E POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12,782,
DISTANCE = 1976 miles.

Equation after substituting the values

Fare=22.44003269*1.202-2.116371823*3+0.010648161*4442.141+0.002800674*28760+0.002488071*27664+4.56963*0.000001*4557004+5.41197*0.000001*3195503+0.076491667*1976-0.001007935*12782-110.8629158*1-138.7123701*0-22.98601604*1 -25.55872353*1 = 234.27

xi. Compare the predictive accuracy of this model with the model (iv). Is this model good enough, or is it worthwhile re-evaluating the model once flights begin on the new route?

Important

Exhaustive (before new flight)

Training: Prediction Summary

Metric	Value
SSE	542384.2
MSE	1416.147
RMSE	37.63173
MAD	29.92301
R2	0.755767

Validation: Prediction Summary

Metric	Value
SSE	402329.5
MSE	1577.763
RMSE	39.72106
MAD	31.60865
R2	0.724545

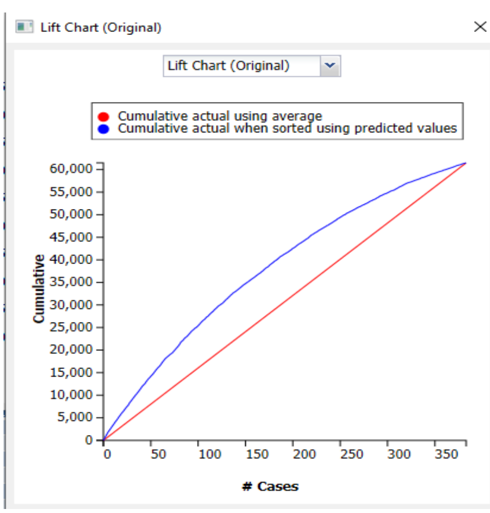
Exhaustive (new flights)

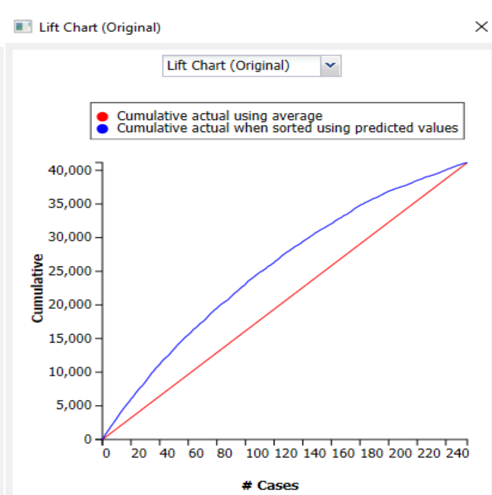
Training: Prediction Summary

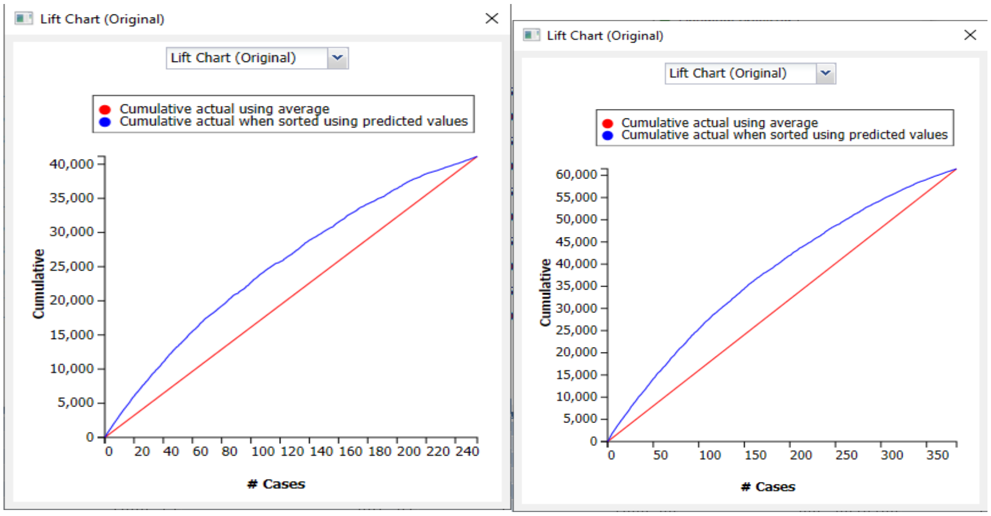
Metric	Value
SSE	476204.6
MSE	1243.354
RMSE	35.26123
MAD	27.59104
R2	0.785567

Validation: Prediction Summary

Metric	Value
SSE	320573.7
MSE	1257.152
RMSE	35.45633
MAD	27.79736
R2	0.780519







If we check the predictive accuracy of the models (iv) and (ix) we can say that model (iv) is more accurate as it has less RMSE value when compared to (ix) and also has higher R^2 values when compared to the other model. We need to re-evaluate the model once flights begin on the new route.