Homework 3

Group 3
Nathalia Negri
Troy Yang

305-318-3446
781-812-4931

negri.n@husky.neu.edu
yang.tr@husky.neu.edu

Percentage of Effort Contributed by Student 1:________50%________

Percentage of Effort Contributed by Student 2:________50%________

Signature of Student 1:____________ Negri **________________**

Signature of Student 2:____________ **___________________**

Submission Date:________June 6, 2017__________

Problem 5.5

a. Classification Matrix

	Predicted Fraud	Predicted Non Fraud
Actual Fraud	310	90
Actual Non Fraud	130	270

b. Classification Matrix

	Predicted Fraud	Predicted Non Fraud
Actual Fraud	6.2	1.8
Actual Non Fraud	257.4	534.6

(257.4 + 1.8)/800 = .324

c. (6.2 + 257.4)/800 = .3295

Problem 5.6

a. Cost = 1000 * 2500 = 2500000
Revenue = 2128 * .5 * 1000 = 1064000
Profit = -2500000 + 1064000 = -1436000

b. According to the Decile-wise lift chart in Figure 5.17, we determined that in order for the average profit per sale to at least double the sales effort cost, the firm should include only the first decile, as it has a lift ratio greater than 2.

2128 * x > 2500 * 2   where x = lift ratio
x = 2.35, therefore the decile must have a lift ratio of at least 2.35

c. According to the Decile-wise lift chart in Figure 5.17, we determined that with a lower cutoff of $2500, the firm can include the first five deciles, as the fifth decile is just above the desired 1.17 lift ratio calculated below.

2128 * x > 2500   where x = lift ratio
x = 1.17, therefore the decile must have a lift ratio of at least 1.17

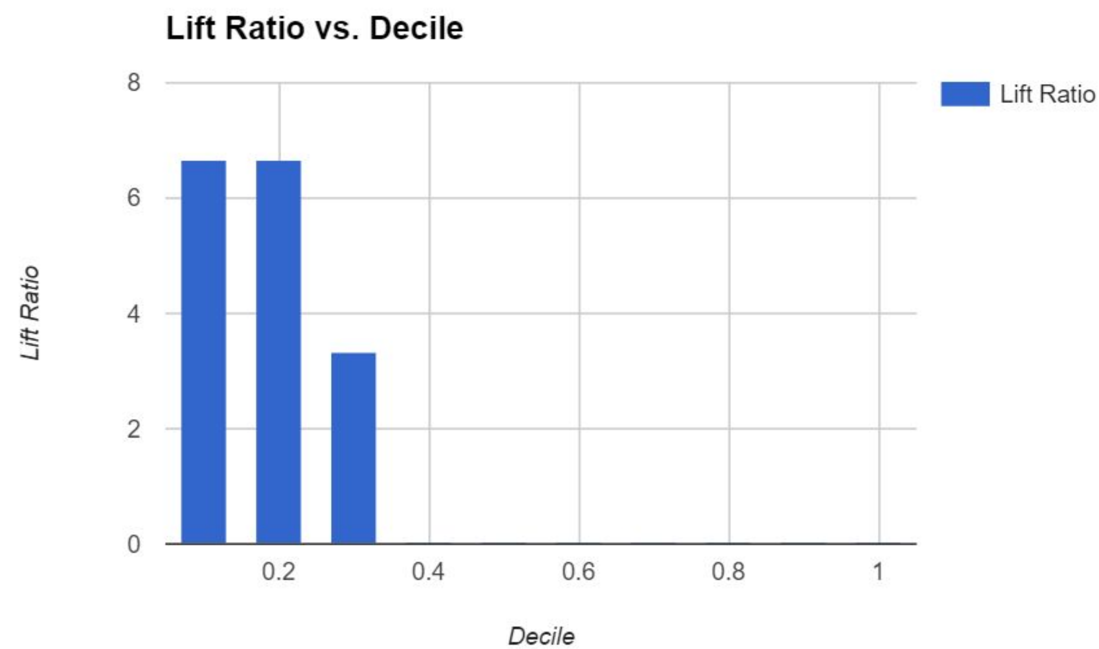
d. Creating a Decile-wise lift chart allows the firm to determine the range of customers they should include to meet their profit margin needs. Otherwise, they could reach out to too many customers and “waste” resources - an inefficient strategy.

Problem 5.7

a.

Cutoff	Error rate	Sensitivity	Specificity
0.25	0.4	1	0.5294117647
0.5	0.1	1	0.8823529412
0.75	0.05	0.6666666667	1

b.



Problem 6.1

a. The training data partition is what would be used to train the model(s). If there are multiple models in contention, the validation data is what would be used to determine which model to proceed with. Finally, once the model is selected, the test data is what would be used to determine the accuracy this model.

b. y = -28.81 - 0.26 * X1 + 3.76 * X2 + 8.28 * X3
where X1 = CRIM, X2 = CHAS, X3 = RM, Y = MEDV

c. 20.844 = -28.81 - 0.26 * 0.1 + 3.76 * 0 + 8.28 * 6

In the validation dataset, we found that there was one house that had a CRIM of 0.10153, CHAS of 0, and RM of 6.279 that had a MEDV of 20.0. That gives us a prediction error of approximately 0.844 = 20.844 - 20.

d.

i. INDUS, NOX and TAX appear to have a large chance of being highly correlated with one another. If a property has a high INDUS, it probably also has a high NOX as well as a low TAX. In other words, a highly industrial area might have high concentration of nitric oxide, which would result in lower property tax rates.

ii.

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	MEDV	CAT..MEDV
DIS	-0.37967009	0.66440822	-0.70802699	-0.099175780	-0.76923011	0.20524621	-0.74788054	1.00000000	-0.494587930	-0.53443158	-0.2324705	-0.4969958	0.2499287	0.1188865
ZN	-0.20046922	1.00000000	-0.53382819	-0.042696719	-0.51660371	0.31199059	-0.56953734	0.66440822	-0.311947826	-0.31456332	-0.3916785	-0.4129946	0.3604453	0.3652962
MEDV	-0.38830461	0.36044534	-0.48372516	0.175260177	-0.42732077	0.69535995	-0.37695457	0.24992873	-0.381626231	-0.46853593	-0.5077867	-0.7376627	1.0000000	0.7897888
RM	-0.21924670	0.31199059	-0.39167585	0.091251225	-0.30218819	1.00000000	-0.24026493	0.20524621	-0.209846668	-0.29204783	-0.3555015	-0.6138083	0.6953599	0.6412654
CAT..MEDV	-0.15198696	0.36529623	-0.36627559	0.108631150	-0.23250184	0.64126541	-0.19119589	0.11888651	-0.197924023	-0.27368672	-0.4434247	-0.4699108	0.7897888	1.0000000
CHAS	-0.05589158	-0.04269672	0.06293803	1.000000000	0.09120281	0.09125123	0.08651777	-0.09917578	-0.007368241	-0.03558652	-0.1215152	-0.0539293	0.1752602	0.1086312
PTRATIO	0.28994558	-0.39167855	0.38324756	-0.121515174	0.18893268	-0.35550149	0.26151501	-0.23247054	0.464741179	0.46085304	1.0000000	0.3740443	-0.5077867	-0.4434247
CRIM	1.00000000	-0.20046922	0.40658341	-0.055891582	0.42097171	-0.21924670	0.35273425	-0.37967009	0.625505145	0.58276431	0.2899456	0.4556215	-0.3883046	-0.1519870
LSTAT	0.45562148	-0.41299457	0.60379972	-0.053929298	0.59087892	-0.61380827	0.60233853	-0.49699583	0.488676335	0.54399341	0.3740443	1.0000000	-0.7376627	-0.4699108
RAD	0.62550515	-0.31194783	0.59512927	-0.007368241	0.61144056	-0.20984667	0.45602245	-0.49458793	1.000000000	0.91022819	0.4647412	0.4886763	-0.3816262	-0.1979240
TAX	0.58276431	-0.31456332	0.72076018	-0.035586518	0.66802320	-0.29204783	0.50645559	-0.53443158	0.910228189	1.00000000	0.4608530	0.5439934	-0.4685359	-0.2736867
AGE	0.35273425	-0.56953734	0.64477851	0.086517774	0.73147010	-0.24026493	1.00000000	-0.74788054	0.456022452	0.50645559	0.2615150	0.6023385	-0.3769546	-0.1911959
INDUS	0.40658341	-0.53382819	1.00000000	0.062938027	0.76365145	-0.39167585	0.64477851	-0.70802699	0.595129275	0.72076018	0.3832476	0.6037997	-0.4837252	-0.3662756
NOX	0.42097171	-0.51660371	0.76365145	0.091202807	1.00000000	-0.30218819	0.73147010	-0.76923011	0.611440563	0.66802320	0.1889327	0.5908789	-0.4273208	-0.2325018

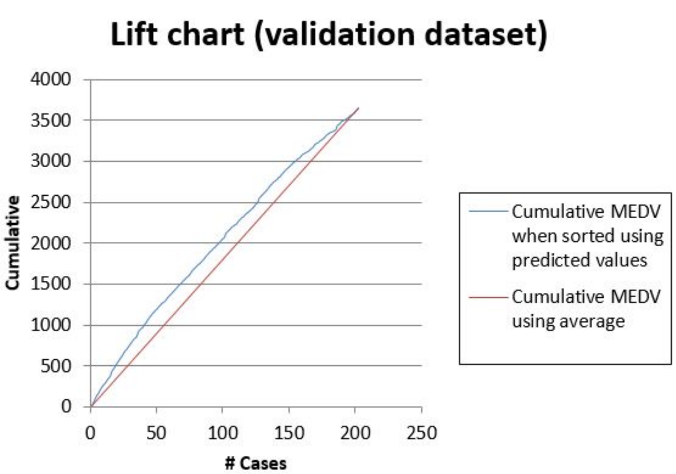
According to this correlation matrix, we would say that we should get rid of the following variables due to potential redundancy. NOX because it shares the same relationship with MEDV as INDUS, and RAD because it shares the same relationship with MEDV as TAX.

iii.

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Model 1	2	3	4	5	6	7	8	9	10	11	12
Choose Subset	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Intercept					CHAS	RM	AGE	DIS	TAX	PTRATIO	LSTAT
Choose Subset	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Intercept			ZN		CHAS	RM	AGE	DIS	TAX	PTRATIO	LSTAT
Choose Subset	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Intercept						RM	AGE	DIS	TAX	PTRATIO	LSTAT

These are the three best models according to XL Miner.

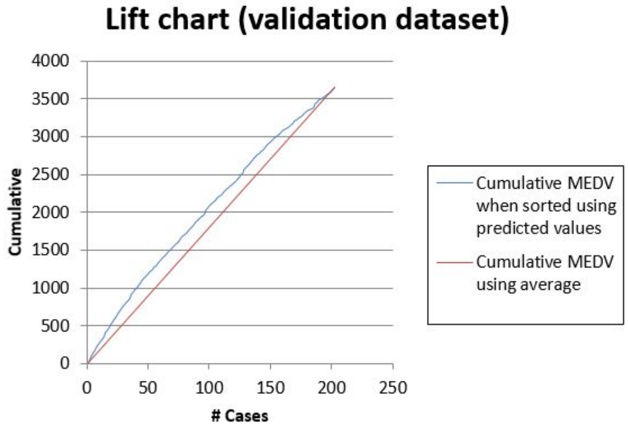
Best Model:



Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
13946.26	8.309084	0.981899637

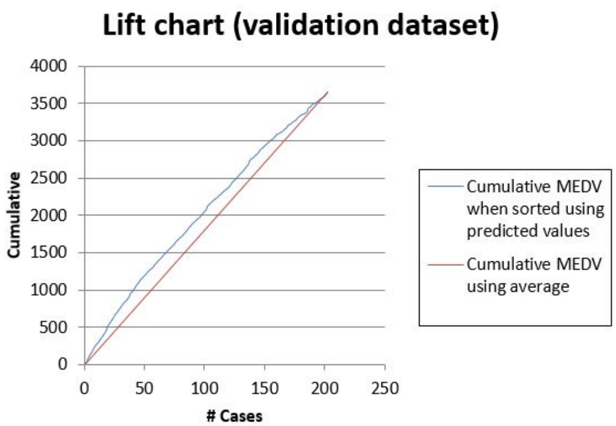
Second Best Model:



Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
13848.25	8.279837	0.997445117

Third Best Model:



Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
14117.14	8.359836	1.012257623

All three models had RMS errors of around 8.3, and average errors of around 1. All three models also had very similar looking lift charts. The best model has 8 coefficients and an adjusted R-squared of 0.864, which is close to 1. The following equation describes the best model:

y = -12.39 + 0.84 * X1 + 9.41 * X2 - 0.05 * X3 - 0.76 * X4 - 0.01 * X5 - 0.60 * X6 - 0.17 * X7
where X1 = CHAS, X2 = RM, X3 = AGE, X4 = DIS, X5 = TAX, X6 = PTRATIO, X7 = LSTAT, and y = MEDV

Problem 6.2

a.

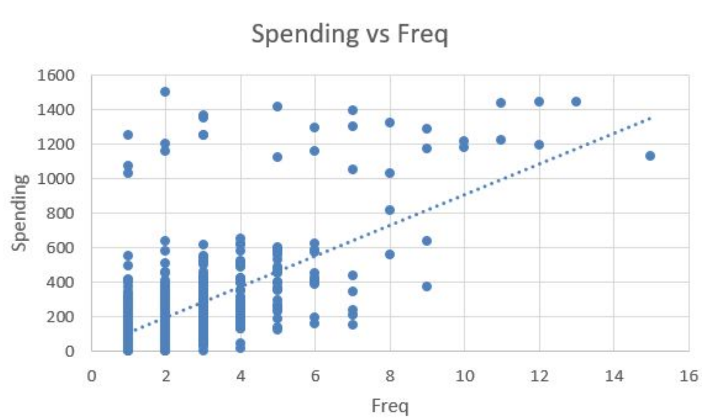
US		
Row Labels	**Average of Spending**	**StdDev of Spending**
0	212.6875449	201.4660834
1	203.567539	224.5243399
Grand Total	**205.09058**	**220.7716387**

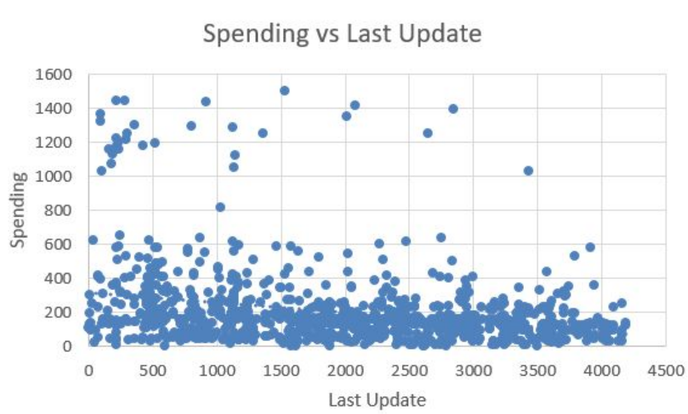
Web Order		
Row Labels	**Average of Spending**	**StdDev of Spending**
0	208.5571711	222.5168702
1	202.184761	219.4608468
Grand Total	**205.09058**	**220.7716387**

Gender = male		
Row Labels	**Average of Spending**	**StdDev of Spending**
0	209.8800617	223.001074
1	200.5620039	218.7635353
Grand Total	**205.09058**	**220.7716387**

Address_is_res		
Row Labels	**Average of Spending**	**StdDev of Spending**
0	210.9939897	239.8158704
1	184.5213004	133.2359847
Grand Total	**205.09058**	**220.7716387**

b.





c.
i. Data was partitioned in this fashion:
60% Training
40% Validation

ii. y = 77.67 - 16.24 * X1 + 91.23 * X2 - 0.021 * X3 + 3.94 * X4 + 14.81 * X5 - 93.56 * X6
where X1 = US, X2, = FREQ, X3= LAST_UPDATE, X4= WEB ORDER, X5 = GENDER_MALE, X6= ADDRESS_RES, Y = SPENDING

iii. Based on the coefficients, we know that US residents that have a high purchasing frequency and are male are most likely going to be the ones that spend the most money.

iv. Using backwards elimination, we would eliminate Web Order. That is because it is the variable with the highest P-Value, which was 0.78.

v. 71.69 = 1.93 - 8.68 * 1 + 98.8 *6 - 0.008 * 380 + 15.38 * 1 + 0.68 * 1 - 98.89 * 0
Predicted Value = 619.60
Actual Value = 450.07
Prediction Error = 450.07 - 619.60 = -169.53

vi. **Training Data Scoring - Summary Report**

Total sum of squared errors	RMS Error	Average Error
17476929.3	170.6699	-1.40214E-14

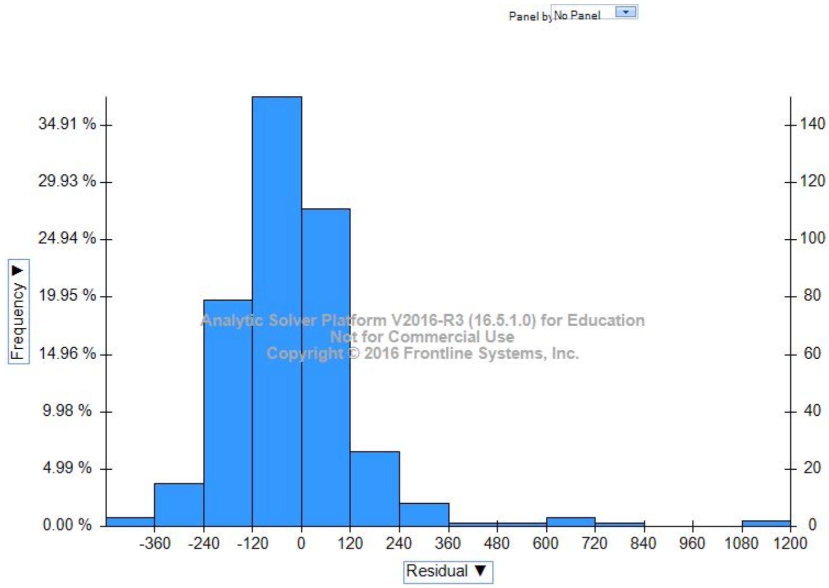
Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
9478471.57	153.9356	10.135082

According to XL Miner, the RMS Error on the validation data is 153.94 and the average error is 10.14. On the training data, 170.67 was the RMS error and the average error was close to 0.

Based on the summary reports, we would say that based on the fact that the RMS error was lower for the validation data than even the training data, that the model has pretty decent predictive accuracy.

vii.

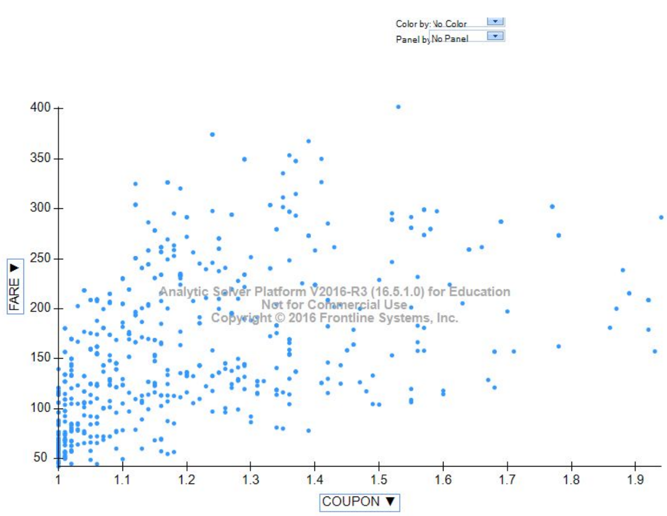


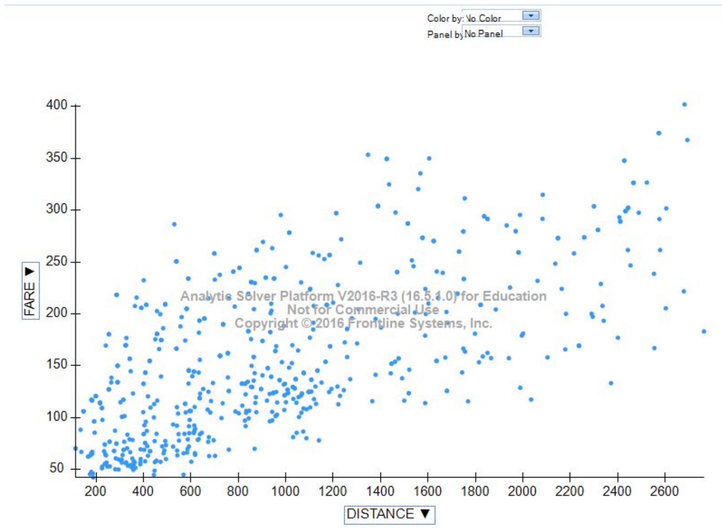
The distribution of the residuals for our model appears to be fairly normal. The vast majority of the data is right around the center, and there are not too many outliers. We believe that since the distribution of the residuals is fairly normal, that using a linear regression is valid in this situation.

Problem 6.3

a. According to the scatter plots and the correlation matrix below, it appears that DISTANCE is the best predictor of fare.

	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.

VACATION

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

SLOT

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

SW

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

GATE

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

It appears that SW is the best categorical variable for predicting fares. Whether or not Southwest serves the route has a large influence on the average fare. As you can see in the pivot table above, the presence of Southwest significantly drops the price of the average fare, whereas the other predictors don't have such a large impact on average fare.

c.

i. We converted categorical variables (Vacation, SW, Slot, and Gate) into dummy variables and partitioned the data into training (60%) and validation (40%) sets.

ii. $y = 10.25 * X1 \; -1.76 * X2 + 0.009 * X3 + 0.001 * X4 + 0.002 * X5 + 4.41605E\text{-}06 * X6 + 4.31136E\text{-}06 * X7 + 0.074 * X8 + -0.001 * X9 - 33.76 * X10 - 23.21 * X11 - 58.32 * X12 -17.46 * X13 - 17.4 * X14$

where X1 = COUPON, X2, = NEW, X3= HI, X4= S_INCOME, X5 = E_INCOME, X6= S_POP, X7 = E_POP, X8, = DISTANCE, X9= PAX, X10= VACATION_YES, X11 = SW_NO, X12=SW_YES, X13 = SLOT_FREE, X14 = GATE_FREE, y = FARE

iii. y = -13.11 + 0.009 * X1 + 0.001 * X2 + 0.002 * X3 + 4.423E-06 * X4 + 4.395E-06 *X5 + 0.076 * X6 - 0.0009 * X7 - 33.88 * X8 - 35.52 * X9 - 17.59 * X10 - 17.18 * X11

where X1 = HI, X2, = S_INCOME, X3= E_INCOME, X4=S_POP, X5 = E_POP, X6= DISTANCE, X7 = PAX, X8 = VACATION_YES, X9= SW_YES, X10= SLOT_FREE, X11 = GATE_FREE, y = FARE

The new model eliminated the predictors COUPON, NEW, VACATION_NO, SW_NO, SLOT_CONTROLLED, GATE_CONSTRAINED.

iv. Below we compare the two models.

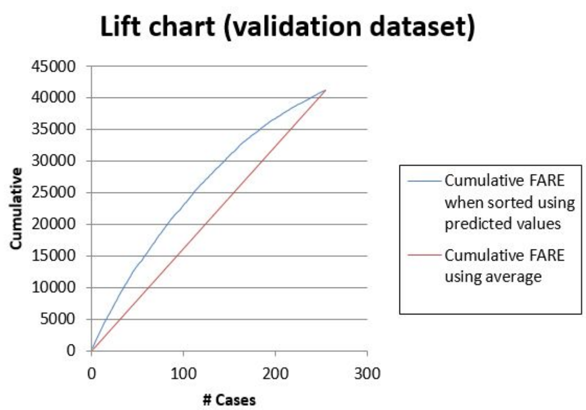
Stepwise Regression:

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
476204.638	35.26122724	5.0981E-14

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
473821.1311	43.10594023	-22.40596



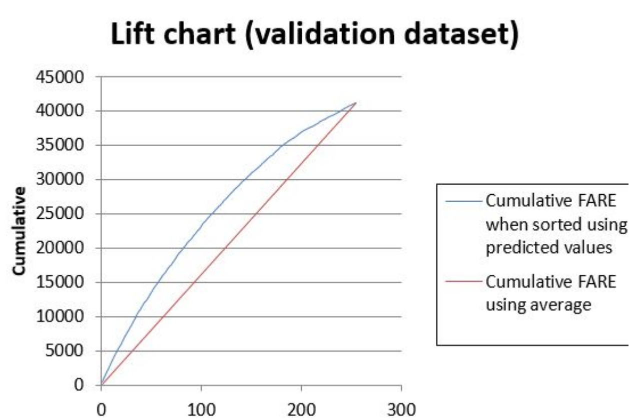
Exhaustive Search:

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
477462.0978	35.307752	6.53774E-14

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
321287.7696	35.495803	3.963915552



The RMSE for the stepwise regression model was larger than the RMSE for the exhaustive search. This is expected because the exhaustive search eliminated a few predictors, creating a more efficient and accurate model. Similarly, the lift chart for the exhaustive search model looks better than the lift chart for the stepwise regression model - however, the differences were not significant.

v. With the given inputs, the predicted fare is $249.06.

vi. If Southwest decides to cover this route, the predicted fare becomes $213.54.

vii. COUPON, NEW, PAX, SW, SLOT, and GATE probably would not be available. Of those predictors, COUPON could possibly be calculated based on DISTANCE, as they are highly correlated. A domain expert might also be able to offer some insight regarding GATE and SLOT information to estimate those predictors.

viii. Using an exhaustive search we found that the best model includes the remaining 7 predictors that we decided would be available before flights begin operating on a new route: HI, S_INCOME, E_INCOME, S_POP, E_POP, DISTANCE, VACATION.

Regression Model

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	-137.64799	24.09610042	-5.71245922	2.27E-08	-185.0284	-90.267584	9869485.5
HI	0.0111289	0.00141025	7.891433756	3.316E-14	0.0083559	0.0139019	4586.4551
S_INCOM	0.0029631	0.000693925	4.270042307	2.479E-05	0.0015986	0.0043276	99951.868
E_INCOM	0.0019803	0.000534449	3.705338074	0.0002429	0.0009294	0.0030312	351602.28
S_POP	4.262E-06	8.58287E-07	4.96541764	1.043E-06	2.574E-06	5.949E-06	18534.967
E_POP	5.936E-06	9.64538E-07	6.153927855	1.942E-09	4.039E-06	7.832E-06	91344.291
DISTANCE	0.0839553	0.00354925	23.65437428	2.344E-76	0.0769764	0.0909342	922586.54
VACATIOI	-35.653642	5.288027395	-6.74233312	5.906E-11	-46.051544	-25.25574	79159.45

Residual DF	375
R^2	0.7059571
Adjusted R^2	0.7004683
Std. Error Estim	41.729307
RSS	653000.66

ix. After plugging in the given values into the regression model in part viii, the predicted value for the fare is $238.60.

x. The RMSE for the newest model (excluding predictors that would not be available before flights start operating on the new route), is very slightly worse than the model in part iii. Since the RMSE is only worse by about $9, and fares are in the hundreds, we don't think it is significantly worse than model iii. Therefore, it is not necessary to revisit the model unless any major changes occur.

d. If the goal of the analysis was to evaluate the impact of Southwest's presence on the airline industry, we do not think it would be necessary to exclude predictors that could not be estimated prior to new routes. Instead, domain experts might invest in collecting data on Southwest and how they operate. Technically speaking, this would result in the data miner focusing on different predictors that might be involved after the expanded data collection. Conceptually, different methods might be necessary as general linear regression might not be the best tool for generating the best model anymore.

Problem 6.4

a.

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	9658.29213	758.7481444	12.72924646	1.53E-33	8168.601	11147.98	84477260374
Age_08_04	-106.71561	4.082567121	-26.13934	1.1E-105	-114.731	-98.7001	7477930646
KM	-0.02089499	0.001779037	-11.7451175	3.33E-29	-0.02439	-0.0174	220209306.1
HP	37.47969865	4.0715304	9.205309788	3.86E-19	29.48584	45.47355	172898133.8
Automatic	438.6578275	202.6700936	2.164393472	0.030771	40.74472	836.5709	18646493.24
Doors	104.5171611	50.33839326	2.076291163	0.038231	5.685082	203.3492	10751906.57
Quarterly_Ta	15.76205668	2.374424626	6.638263649	6.36E-11	11.10022	20.42389	290748125.8
Mfr_Guarante	194.6413277	100.2121476	1.942292749	0.052502	-2.11058	391.3932	108.5356443
Guarantee_P	63.85650868	15.39298003	4.148417562	3.76E-05	33.63464	94.07838	3160528.34
Airco	133.6504502	118.2838596	1.129912827	0.258899	-98.5826	365.8835	19214466.68
Automatic_ai	3049.588727	239.1272601	12.7529949	1.2E-33	2580.097	3519.08	330862930.8
CD_Player	233.8708788	131.4519621	1.779135702	0.075651	-24.2158	491.9576	5962338.249
Powered_Wi	395.0633454	113.5110263	3.480396206	0.000532	172.201	617.9257	18304136.86
Sport_Model	415.7743878	108.9756908	3.815294813	0.000148	201.8165	629.7322	25008364.63
Tow_Bar	-268.148481	107.0006284	-2.50604585	0.012434	-478.229	-58.0684	10188824.85
Fuel_Type_CN	0	0	N/A	N/A	0	0	0
Fuel_Type_Di	2511.749109	485.3681571	5.174935917	2.98E-07	1558.8	3464.699	17195757.39
Fuel_Type_Pe	1977.800463	501.9089595	3.940556201	8.94E-05	992.3756	2963.225	23645611.8

Based on their p-values being the smallest, Age_08_04, KM, Automatic_airco appear to be the most influential when it comes to the price.

b. **Training Data Scoring - Summary Report**

Total sum of squared errors	RMS Error	Average Error
1067464698	1219.311131	-4.94015E-13

Residual DF	701
R^2	0.89009024
Adjusted R^2	0.8875816
Std. Error Estimate	1234.00737
RSS	1067464698

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
2042682287	2177.01678	1672.005908

Test Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1335596401	2157.231095	1607.71665

Based on the adjusted R squared value of .888 RMS error of 2157 and average error of 1607.72, it appears that the model has fairly good predictive accuracy. In the context of tens of thousands,, 1607 is pretty small.

Problem 7.2

a. Matching all predictors except for age, income, experience and CC avg, we narrowed it down to around 20 cases. All of these cases were classified were 0, which means they were not accepted for the loan, so the true nearest neighbor to this specific case is a 0, which would mean this case would be classified as a 0 as well.

b. Low K values tend to overfit, but are able to capture the local structure. High K values provide more smoothing, but are susceptible to missing the local structure. As K approaches n, the model basically becomes majority rules. The best K values tend to be between 1 and N. However, in XLMiner 20 is the maximum value for K, so in XLMiner, the best values will be between 1 and 20, with the sample principles for low and high K values still applying.

c. **Validation Data Scoring - Summary Report (for k = 5)**

Cutoff probability value for success (UPDATABLE)	0.5	Updating the value here will NOT update value in detailed report

Confusion Matrix		
	Predicted Class	
Actual Clas	1	0
1	65	111
0	58	1766

d. Using K=5 that row was classified as a 0.

e. **Training Data Scoring - Summary Report (for k = 5)**

Cutoff probability value for success (UPDATABLE)	0.5	Updating the value here will NOT update value in detailed report

Confusion Matrix		
	Predicted Class	
Actual Clas	1	0
1	126	119
0	52	2203

Test Data Scoring - Summary Report (for k = 5)

Cutoff probability value for success (UPDATABLE)	0.5	Updating the value here will NOT update value in detailed report

Confusion Matrix		
	Predicted Class	
Actual Clas	1	0
1	38	59
0	28	875

Validation Data Scoring - Summary Report (for k = 5)

Cutoff probability value for success (UPDATABLE)	0.5	Updating the value here will NOT update value in detailed report

Confusion Matrix		
	Predicted Class	
Actual Clas	1	0
1	52	86
0	49	1313

As expected, the sensitivity on the training data for this model was somewhat higher than the sensitivity on the validation data as well as the testing data. This however was not all that surprising, given the fact that the training data was used to build the model in the first place. Comparing the training and validation matrices against the test data matrix is a little bit tough, because the validation and test matrices have more in common with each other than the training and validation ones. The reason for that is because in this case, we are only using one model, therefore the validation and test data are used for the same purpose.

Problem 7.3

a. **Validation error log for different k**

Value of k	Training RMS Error	Validation RMS Error	
1	0	4.643584	
2	0	4.549499	
3	1.41995E-15	4.41328	
4	0	4.39613	
5	0	4.286175	<- Best k

The best K is 5, which means that the model looks at the 5 nearest neighbors for a given record/case, and also had the lowest error.

b.

Workbook	7.3_BostonHousing.xlsx
Worksheet	Sheet8
Range	$A$1:$L$2

Predicted Value	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT
18.91836	0.2	0	7	0	0.538	6	62	4.7	4	307	21	10

c. For any training data scenario, the nearest neighbor for a given record is always going to be itself.

d. The validation data error is overly optimistic because the validation data was used to select the best K, therefore fitting the validation data better than new data.

e. Predicting MEDV for several thousands of new tracts using K-NN prediction will cost a lot in terms of processing power. For K-NN for each prediction, the distance (Statistical, Euclidian or Manhattan) must be calculated to other records in the training data, and the closest K neighbors will be used to help predicted. For one record that is not that bad, but once you start scaling up it gets worse and worse.