Jenet Baribeau
CIS 445-01
Project 4

Prior to importing the data into SAS, I explored some of the outlier information available in the Excel format. Reviewing the data in Excel, proved to be helpful in identifying some of the outliers. The second node is partitioning data to create the training and validation sets. The data partition provided model weights for estimation prior to assessment. Three different types of statistical methods were used to model the real estate data. These ranged from non-flexible logistic regression through case/memory based reasoning (MBR) to a completely flexible scenario in neural networks (NN). With a total of 6 nodes, one set of each of the nodes (LR, MBR, NN) with minimal data normalization was created for comparison. In addition, after the data partition, a filter was added to mitigate the outliers seen in the Excel spreadsheet. Within the filter (see table below), a second transformation of the data included the following variables were changed to account for outliers and mistakes in the data.

Table 1 Variables Filtered

	Name of Variable	Level	Filtered
1	Age	Interval	Filtered to 0 because there can't be a negative age.
2	LotSize	Ordinal	Filtered to 2 because there is no 3 available in the description which rendered it rejected.
3	GarageType	Nominal	Filtered to 3 because 4 & 5 are outliers and there was only 1 of each in the list making it insignificant.
4	SalePrice	Interval	(Target) Filtered the min amount to $25000. There were 2 outlier amounts.

The Age filtered results initially I did incorrect because I changed the minimum age to 18 thinking it was the age of the owner not the age of the properties. However, once discovering the mistake, it was changed to 0 for the simple reason that there cannot be a negative age. I didn't really notice a difference in the outcomes either way. The 18 year difference had little impact.

The LotSize was filtered due to a mistaken entered data piece of 3. According to the lot size there is no nominal representation with a 3. With the 3 removed, the majority if not all of the LotSize data points were all the same (1). After the filter, the variable selection rejected the LotSize as irrelevant.

The GarageType was filtered to 3 because there was only 1 of each of the 4&5. With that said, the 4 was a "garage in the basement" and 5 was a "built-in garage". Logically, based on lower income neighborhood, neither of these items will be frequent and are an anomaly. Again the GarageType was rejected.

In the variable selection, I did notice the "TotalArea", "UpperArea", "SecondFloor", and "FirstFloor" were duplicate type information. After the filter, the variable selection rejected "FirstFloor" and "UpperArea" but not "SecondFloor" which gave me the idea that what mattered was actually having the second floor. I changed the variables to mark the measurement level of the second floor to a binary showing that either the house has a second floor or it doesn't. The results in Figure 1 Variable Selection after Filter demonstrates the effect of changing the second

floor to binary and rejecting the upper area and the first floor giving the initial Neural Network model without any transformation.

Figure 1 Variable Selection after Filter

Variable Selection

Variable Name	Role	Measurement Level	Type	Label	Reasons for Rejection
Age	Input	Interval	Numeric	Age	
Basement	Rejected	Interval	Numeric	Basement	Varsel:Small R-square val...
BasementType	Rejected	Ordinal	Numeric	BasementType	Varsel:Small R-square val...
Baths	Input	Interval	Numeric	Baths	
CentralAir	Input	Binary	Numeric	CentralAir	
ConstructionType	Input	Nominal	Numeric	ConstructionType	
FirePlace	Input	Interval	Numeric	FirePlace	
FirstFloor	Rejected	Interval	Numeric	FirstFloor	Varsel:Small R-square val...
G_GarageType	Input	Nominal	Numeric	Grouped Levels for Garage...	
GarageSize	Input	Interval	Numeric	GarageSize	
GarageType	Rejected	Nominal	Numeric	GarageType	Varsel:Small R-square val...
LotSize	Rejected	Ordinal	Numeric	LotSize	Varsel:Small R-square val...
Neighborhood	Rejected	Nominal	Numeric	Neighborhood	Varsel:Small R-square val...
SecondFloor	Input	Interval	Numeric	SecondFloor	
TotalArea	Input	Interval	Numeric	TotalArea	
UpperArea	Rejected	Interval	Numeric	UpperArea	Varsel:Small R-square val...
WallType	Input	Nominal	Numeric	WallType	

The initial results showing the Neural Network2 (Figure 2 Initial Results) after transformation being the best node to compare the data. The Maximum Absolute Error (MAE) is the second to lowest. The Line Regression with the filtered data had the lowest. If the object is to minimize the errors in prediction, if I read this correctly, then it appears that the errors are somewhat significant leading to minimal correlation.

Figure 2 Initial Results

Output

	tics	Neural2	MBR2	Reg	MBR	Neural	Reg2
61	Akaike's Information Criterion	2546.81	2502.51	2816.58	2814.81	2912.28	2482.88
62	Average Squared Error	131324168.10	180545034.87	198721384.11	238131035.99	185096845.60	135468715.16
63	Average Error Function	131324168.10	180545034.87	198721384.11	238131035.99	185096845.60	135468715.16
64	ion Criterion: Valid: Average Squared Error	266909508.37	291493772.69	293972679.83	298270928.51	302495823.01	334724012.27
65	Degrees of Freedom for Error	82.00	125.00	122.00	136.00	69.00	116.00
66	Model Degrees of Freedom	49.00	6.00	23.00	9.00	76.00	15.00
67	Total Degrees of Freedom	131.00	131.00	145.00	145.00	145.00	131.00
68	Divisor for ASE	131.00	131.00	145.00	145.00	145.00	131.00
69	Error Function	17203466021.54	23651399568.11	28814600695.55	34529000218.38	26839042612.17	17746401686.02
70	Final Prediction Error	288272564.13	197877358.22	273649119.10	269648378.99	592846418.52	170503727.70
71	Maximum Absolute Error	41096.26	41658.44	81245.96	81867.75	81788.67	34647.45
72	Misclassification Rate	.	.	.	.	.	.
73	Mean Square Error	209798366.12	189211196.54	236185251.60	253889707.49	388971632.06	152986221.43
74	Sum of Frequencies	131.00	131.00	145.00	145.00	145.00	131.00
75	Number of Estimate Weights	49.00	6.00	23.00	9.00	76.00	15.00
76	Root Average Sum of Squares	11459.68	13436.70	14096.86	15431.49	13605.03	11639.10
77	Root Final Prediction Error	16978.59	14066.89	16542.34	16420.97	24348.44	13057.71
78	Root Mean Squared Error	14484.42	13755.41	15368.32	15933.92	19722.36	12368.76
79	Schwarz's Bayesian Criterion	2687.69	2519.76	2885.04	2841.60	3138.51	2526.00
80	Sum of Squared Errors	17203466021.54	23651399568.11	28814600695.55	34529000218.38	26839042612.17	17746401686.02
81	Sum of Case Weights Times Freq	131.00	131.00	145.00	145.00	145.00	131.00
82	Number of Wrong Classifications	.	.	.	.	.	.

Figure 3 Model Comparison with Initial Filtered input

Results - Node: Model Comparison Diagram: Project_4

File Edit View Window

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function
Y	Neural2	Neural2	Neural Net...	SalePrice	SalePrice	2.6691E8	2546.807	1.3132E8	1.3132E8	82	49	131	131	1.72E1
	MBR2	MBR2	MBR (2)	SalePrice	SalePrice	2.9149E8	2502.505	1.8055E8	1.8055E8	125	6	131	131	2.365E1
	Reg	Reg	Regression	SalePrice	SalePrice	2.9397E8	2816.575	1.9872E8	1.9872E8	122	23	145	145	2.881E1
	MBR	MBR	MBR	SalePrice	SalePrice	2.9827E8	2814.808	2.3813E8	2.3813E8	136	9	145	145	3.453E1
	Neural	Neural	Neural Net...	SalePrice	SalePrice	3.025E8	2912.277	1.851E8	1.851E8	69	76	145	145	2.684E1
	Reg2	Reg2	Regression...	SalePrice	SalePrice	3.3472E8	2482.877	1.3547E8	1.3547E8	116	15	131	131	1.775E1

Figure 5 Second Results after changing 2nd Flr to binary

Output

	Neural	MBR2	Neural2	Reg	MBR	Reg2
Akaike's Information Criterion	2965.77	2324.16	2356.74	2770.29	2808.16	2309.37
Average Squared Error	162914515.77	172950577.75	121155967.57	122385124.26	237063867.80	134373362.52
Average Error Function	162914515.77	172950577.75	121155967.57	122385124.26	237063867.80	134373362.52
on Criterion: Valid: Average Squared Error	280828759.86	289744108.56	297490541.26	298541070.51	301194885.92	318537900.34
Degrees of Freedom for Error	33.00	117.00	79.00	110.00	139.00	109.00
Model Degrees of Freedom	112.00	5.00	43.00	35.00	6.00	13.00
Total Degrees of Freedom	145.00	122.00	122.00	145.00	145.00	122.00
Divisor for ASE	145.00	122.00	122.00	145.00	145.00	122.00
Error Function	23622604786.20	21099970485.52	14781028043.48	17745843017.87	34374260831.05	16393550227.97
Final Prediction Error	1268758501.58	187732678.41	253047274.04	200266566.97	257529813.22	166425724.23
Maximum Absolute Error	71639.92	41658.44	37540.73	39680.37	78252.63	34718.87
Misclassification Rate	.	.	.	.	.	.
Mean Square Error	715836508.67	180341628.08	187101620.80	161325845.62	247296840.51	150399543.38
Sum of Frequencies	145.00	122.00	122.00	145.00	145.00	122.00
Number of Estimate Weights	112.00	5.00	43.00	35.00	6.00	13.00
Root Average Sum of Squares	12763.80	13151.07	11007.09	11062.78	15396.88	11591.95
Root Final Prediction Error	35619.64	13701.56	15907.46	14151.56	16047.74	12900.61
Root Mean Squared Error	26755.12	13429.13	13678.51	12701.41	15725.67	12263.75
Schwarz's Bayesian Criterion	3299.16	2338.18	2477.31	2874.47	2826.02	2345.82
Sum of Squared Errors	23622604786.20	21099970485.52	14781028043.48	17745843017.87	34374260831.05	16393550227.97
Sum of Case Weights Times Freq	145.00	122.00	122.00	145.00	145.00	122.00
Number of Wrong Classifications	.	.	.	.	.	.

Figure 4 Model Comparison 2nd Filter

Results - Node: Model Comparison Diagram: Project_4

File Edit View Window

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function
Y	Neural	Neural	Neural Net...	SalePrice	SalePrice	2.8083E8	2965.767	1.6291E8	1.6291E8	33	112	145	145	2.362E1
	MBR2	MBR2	MBR (2)	SalePrice	SalePrice	2.8974E8	2324.159	1.7295E8	1.7295E8	117	5	122	122	2.11E1
	Neural2	Neural2	Neural Net...	SalePrice	SalePrice	2.9749E8	2356.736	1.2116E8	1.2116E8	79	43	122	122	1.478E1
	Reg	Reg	Regression	SalePrice	SalePrice	2.9854E8	2770.289	1.2239E8	1.2239E8	110	35	145	145	1.775E1
	MBR	MBR	MBR	SalePrice	SalePrice	3.0119E8	2808.157	2.3706E8	2.3706E8	139	6	145	145	3.437E1
	Reg2	Reg2	Regression...	SalePrice	SalePrice	3.1854E8	2309.368	1.3437E8	1.3437E8	109	13	122	122	1.639E1

I thought I saw a correlation in the "Second Floor" when the variable selection rejected all the other floors that essentially add up to the "Total area" of the house. By changing the second floor input to binary and rejecting the other floors, rerunning the entire module there was some statistical developments. The degree of errors is reduced significantly in comparison to the filtered results in the neural network (2) to more than half.

In manipulating and reviewing the data, I can see the how it is difficult to find a true pattern in lower income neighborhoods. The MBR2 was closely behind the NN with the change in input for the "Second Floor" demonstrating that perhaps the estimation of ranking with normalized distances can predict real estate pricing. It was difficult to see the efficacy of the MBR in comparison to the NN because the graphs were very close in arch and extension.

Figure 5 Workflow Diagram