

Machine Learning with SAS Enterprise Miner		Group No	3
Student Name <i>(as per record)</i>	Adekola Oluwatade Job	Student Nos	215383256
	Yuan Yue		215398617
	Carolyn Pimentel		214443106

	Exceptional	Meets expectations	Issues noted	Improve	Unacceptable
Prepare Exec Report					
Prepare Data					
Discover Relationships					
Create Models					
Evaluate & Improve					
Provide Solution					
Research & Extend					
Brief Comments	This table has been provided for self-assessment only.				
	All research and SAS targets were distributed evenly (3 targets as per 3 group members) with each team member contributing evenly and in parallel to working towards one goal. The report was well organised and written overall due to each member's good contributions.....				
					Total

Executive summary (one page limit)

With the recent economic and population growth, it appears that the property market at Ames Iowa has been very active. Considering a limited number of properties advertised for auction, potential buyers would like to bid for the property with a reasonable price expectation in mind. However, besides the price, factors such as affordability and value for money also impact on the decision-making process.

As an independent online business, Best Iowa Buys have collected a sample data comprising 2930 records of properties sold between 2006 and late 2010 in Ames, each described with 79 variables such as location, house style, land size etc. Even though the dataset is available for analysis, predicting house prices is not an easy process.

Best Iowa Buys has employed myself as a data analyst to better serve our members. I am motivated to develop a predictive model to estimate the price of Ames properties advertised for auction. Further, this will assist members with information such as affordability and value for money based on different groups of houses, and will hopefully make it easier for members to identify a suitable property to purchase.

Using different machine learning algorithms i.e supervised and unsupervised, we'll build different models, evaluate them and use the best algorithm to make predictions and suggest solutions. Deep analysis will be carried out with up to eighty attributes of the properties that hugely influences the price of houses to be put on auction.

Cluster analysis (an unsupervised learning mechanism) helps to naturally group properties that belong to the same group based on data points. This will help us reduce dimensionality of the dataset and even help create variables that combines several attributes thus making model creation simpler and easier to understand. The creation of clusters and segments like affordable house, houses that represents good value for money etc. will ultimately guide potential customers in their selection choices.

These sorts of questions as above are aimed to be solved using cluster analysis and segmentation. This gives an overall guide to different categories of potential customers. Also, this allows for flexibility in house selection based on several options. The cluster analysis will provide a guide to insightful unique grouping of variables to provide additional information to analysts as well.

After going through several articles and online web materials on what the term affordability and value for money thoroughly means and how they influence price of real estates properties, these insights were discovered about the Ames dataset. Using the **house style** variable, properties with less than 1.5 stories are categorised as **small** while other above 1.5 are categorised as **large**. This makes sense as grouping the properties into two categories will help narrow down classification. Next is using sales price to further group them. From the sales price distribution summary, the mean and median value will be used to estimate expensive small houses, cheap small houses etc. The segment variable created as a result of this rule can further be used as target for affordability.

Also, for good value for money, variables such as neighbourhood location, overall quality, overall condition, etc. are huge determining factors. Using rules builder node, these variables will be combined as a segment and eventually set as target to predict houses that have great potentials of being good value for money. These segments will greatly reduce the dimensionality of the Ames Dataset.

More so, models will be built and several evaluations will be done on these models to determine the champion model that'd be used by Best Iowa Buys to predict the sales price of new properties that would be advertised for auction in the future. Selection criteria will hugely be based on models with lowest average squared error, lowest misclassification rate and other selection criterion. Another technique called cross validation can be used to evaluate the models and determine the final model as well.

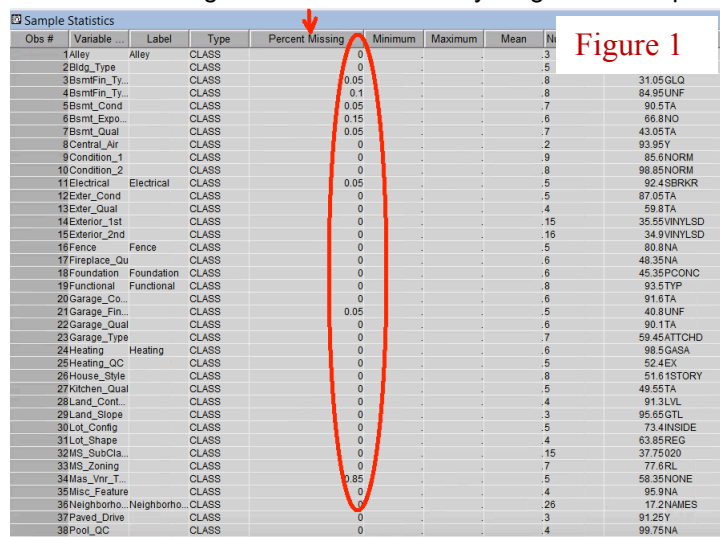
In conclusion, all output will be exported using the score node for operational use by the clients. This can help extensively in easy running and understanding of the process flow of the analytics solution. The target output can also be used to compare with the current economic situation and can be used to guide the prediction accuracy.

Data exploration and preparation in SAS EM (one page limit)

In the data preparation stage, all sorts of exploratory analysis were done to ensure that the data is ready for model building. Firstly, using the Stat Explore node, all the variables were explored to identify their properties, data type, frequency plots, skewness or normality, correlation with the target variable, cleaning of missing values using the impute node etc. Also, the transformation node was used to correct data that were highly skewed. In the transformation node, the log transformation and optimal binning methods were selected to correct and help choose good interval boundaries for interval variables.

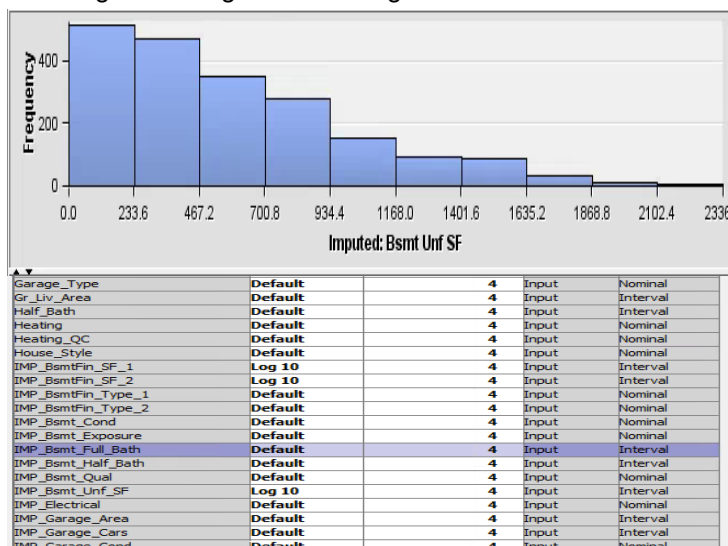
The below diagrams show a summary diagram of the process undergone during data cleaning.

Sample Statistics



Obs #	Variable	Label	Type	Percent Missing	Minimum	Maximum	Mean	Std Dev
1	Alley	Alley	CLASS	0				
2	Bldg_Type		CLASS	0				
3	BsmtFin_Ty_1		CLASS	0.05			31.05	GLO
4	BsmtFin_Ty_2		CLASS	0.1			84.95	UNF
5	Bsmt_Cond		CLASS	0.05			90.5	TA
6	Bsmt_Expo		CLASS	0.15			66.8	NO
7	Bsmt_Qual		CLASS	0.05			43.65	TA
8	Central_Air		CLASS	0			93.95	Y
9	Condition_1		CLASS	0			85.6	NORM
10	Condition_2		CLASS	0			98.85	NORM
11	Electrical	Electrical	CLASS	0.05			92.45	BRKR
12	Exter_Cond		CLASS	0			87.65	TA
13	Exter_Qual		CLASS	0			59.8	TA
14	Exterior_1st		CLASS	0			35.55	VINYLS
15	Exterior_2nd		CLASS	0			34.9	VINYLS
16	Fence	Fence	CLASS	0			80.8	NA
17	Fireplace_Ou		CLASS	0			48.35	NA
18	Foundation	Foundation	CLASS	0			45.35	PCONC
19	Functional	Functional	CLASS	0			93.5	TYP
20	Garage_Co		CLASS	0			91.6	TA
21	Garage_Fin		CLASS	0.05			40.8	UNF
22	Garage_Qual		CLASS	0			90.1	TA
23	Garage_Type		CLASS	0			59.45	ATCHD
24	Heating	Heating	CLASS	0			98.5	GASA
25	Heating_QC		CLASS	0			52.4	EX
26	House_Style		CLASS	0			51.6	1STORY
27	Kitchen_Qual		CLASS	0			49.55	TA
28	Land_Cont		CLASS	0			91.3	VLY
29	Land_Slope		CLASS	0			95.65	GTL
30	Lot_Config		CLASS	0			73.4	INSIDE
31	Lot_Shape		CLASS	0			63.85	REG
32	MS_SubCla		CLASS	0			37.75	020
33	MS_Zoning		CLASS	0			77.6	RL
34	Mas_Vnr_T		CLASS	0.85			58.35	NONE
35	Misc_Feature		CLASS	0			95.9	NA
36	Neighborhood	Neighborhood	CLASS	0			17.2	NAMES
37	Paved_Drive		CLASS	0			91.25	Y
38	Pool_QC		CLASS	0			99.75	NA
39	Pool_Lot		CLASS	0			69.6	COMPLET

Figure 1

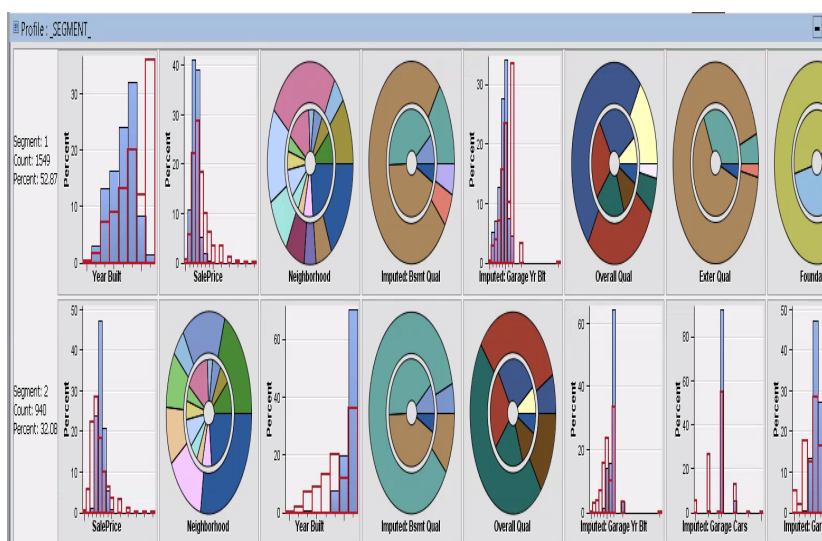


From clustering, three targets are identified which are Sales Price, Overall Quality and Affordability. A new variable called Affordability was created through clustering.

There are 8 House Styles - 1.5Fin, 1.5n Unf, 1Story, 2.5Fin, 2.5Unf, 2Story, SFoyer and SLvl. The media prices for these styles are 129675, 113000, 155000, 194000, 160950, 189000, 143000 and 165000.

If the sales price is lower than the median price within its house style, we will define the affordability High, otherwise, the affordability is Low.

Variables of Sales Price, Neighbourhood, Exter_Qual, Year_Built and Bsmt_Qual are the predictors for the target of Overall Quality.



The plot below shows the correlation between all selected variables and the target variable. As seen in the diagram below, the top five most related predictors are **Overall_Quality**, **Bsmt_Quality**, **Garage_Cars**, **Exter_Qual** and **Gr_Liv_Area** which are shown in the below Pearson Correlation Plot.



Create Models in SAS EM (two page limit / page 1)

After all data pre-processing procedures have been carried out (data preparation cleaning and performing cluster analysis), the variable selection node was implemented where all unimportant variables were dropped using criteria like Rsquared values etc. All the selected variables were then used to build predictive models that would predict affordable houses. Some of the models built include, decision trees, Neural network, Logistic Regression, gradient boosting, Ensemble models and High performance models like HP forest. The targets being Sales Price, Value for money and EM_Outcome (Affordable houses) which were both created using the rules builder node. The combinations are as a result of some variables which include house style and its categorical values used in creating segments. The below diagram shows the selected models and how the SAS diagram was implemented.

Rules Builder for Affordability

Creating Affordability variable with rules builder

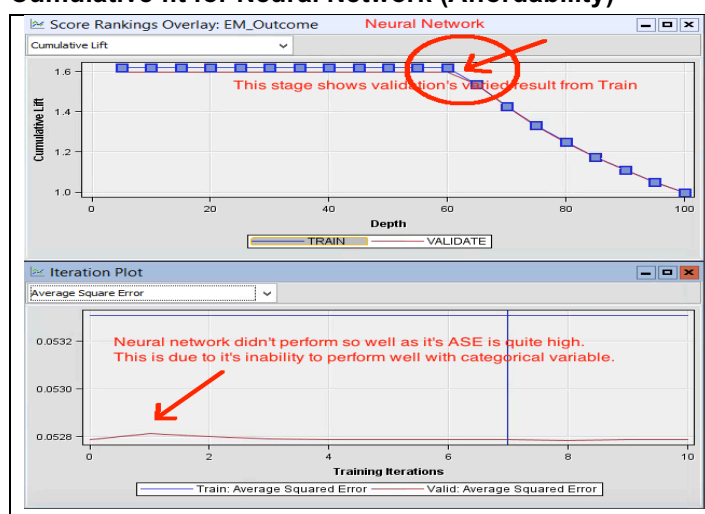
Value	Used	Result
0	Yes	
Larger_Houses	Yes	
Cheap_Larger_Houses	Yes	
Expensive_Larger_Houses	Yes	
Expensive_Smaller_Houses	Yes	

Rules Builder for Value for money

```
IF Overall_Qual IN ("3.0", "4.0", "5.0", "6.0") THEN DO;
  EM_Outcome = "MORE VALUE";
  IF SalePrice < 120000.0 THEN DO;
    EM_Outcome = "MORE VALUE";
  END;
END;
ELSE IF Overall_Qual NOT IN ("3.0", "4.0", "5.0", "6.0") THEN DO;
  EM_Outcome = "LESS VALUE";
  IF SalePrice > 200000.0 THEN DO;
    EM_Outcome = "LESS VALUE";
  END;
END;
```

The two formulas represents the formulas used to derive Affordability and Value for money respectively.

Cumulative fit for Neural Network (Affordability)

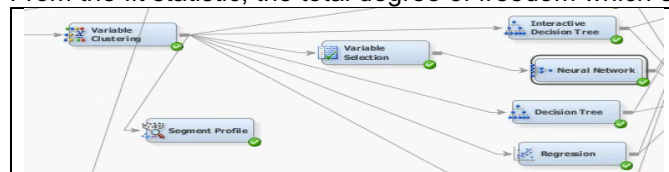


Fit statistics for Neural Network (Affordability)

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
EM_Outco...		DFT	Total Degrees of Freedom	8204		
EM_Outco...		DFE	Degrees of Freedom for Error	8182		
EM_Outco...		DFM	Model Degrees of Freedom	22		
EM_Outco...		NW	Number of Estimated Weights	22		
EM_Outco...		AIC	Akaike's Information Criterion	2041.505		
EM_Outco...		SBC	Schwarz's Bayesian Criterion	2195.777		
EM_Outco...		ASE	Average Squared Error	0.053306	0.052787	
EM_Outco...		MAX	Maximum Absolute Error	0.892127	0.892127	
EM_Outco...		DIV	Divisor for ASE	10255	4395	
EM_Outco...		NOBS	Sum of Frequencies	2051	879	
EM_Outco...		RASE	Root Average Squared Error	0.230881	0.229754	
EM_Outco...		SSE	Sum of Squared Errors	546.6557	231.9983	
EM_Outco...		SUMW	Sum of Case Weights Times Freq	10255	4395	
EM_Outco...		FPE	Final Prediction Error	0.053593		
EM_Outco...		MSE	Mean Squared Error	0.05345	0.052787	

From the cumulative lift plot, out of the 100% data that were used to train and validate the data, the model performed well after 60% where the train and validate gives the same measure. Also, in the Average squared error plot, it's seen that there's minimal variation between the train's ASE and the validation's ASE. This

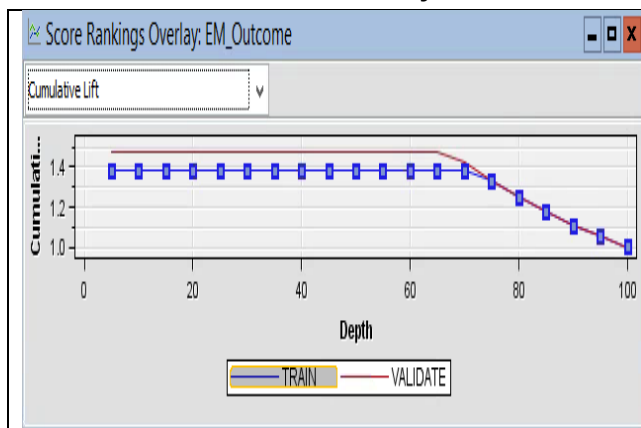
From the fit statistic, the total degree of freedom which explains the freely selected numbers that are used for training.



The diagram shows how the output of the cluster analysis node fed into the predictive algorithms used in prediction.

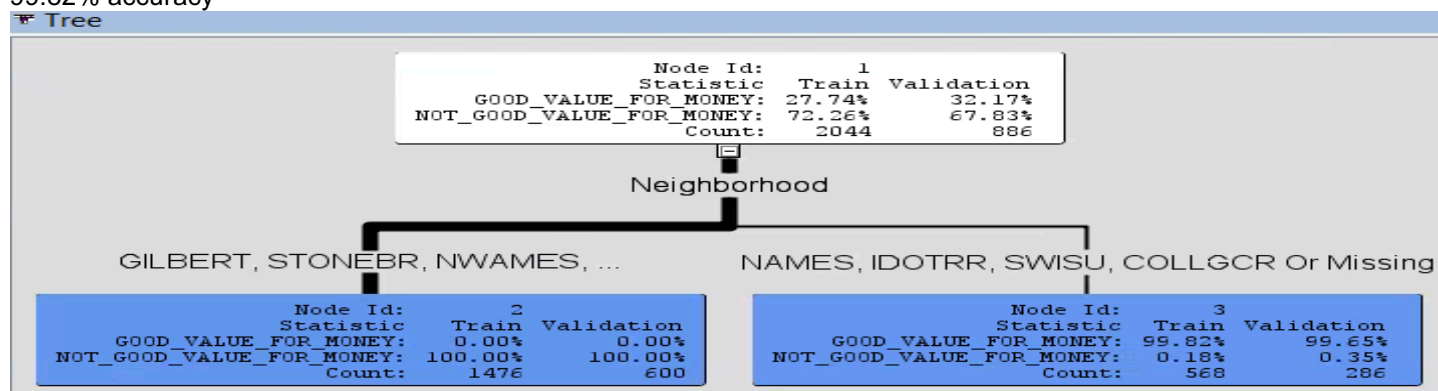
Create Models in SAS EM (two page limit / page 2)

Decision Tree for Value for money and other summary plots



The cumulative lift plot which explains the relationship between Depth of training and validation and also the cumulative lift where the validation performed better until after 70%. Then it started plummeting. Likewise, in the Decision tree diagram below, the data was split according to the rules in the rules builder node and used to determine the most important node (Neighborhood). From the fit statistics, it's also discovered that the misclassification rate for both train and validation are **0.00048** and **0.0001**. This implies a good model and good accuracy for predicting houses with good value for money.

The decision tree diagram below shows the Neighbourhood variable as the most important variable with a very high entropy value and beneath it are the roots. This variable was created as a segment in the rules builder node to identify locations that can potentially bring in good returns if properties located in those areas are invested on. From the tree, it's evident that area codes with NAMES, IDOTTR, SWISU etc are a good value for money as the training statistics gives a 99.82% accuracy



Also, from the fit statistics plot, it's seen that the misclassification rate and maximum absolute error are the major criterion for evaluation for the decision tree model. The misclassification rate means that the model, at worse case scenario, when it misclassifies, it will only do that at the rate of 0.00048

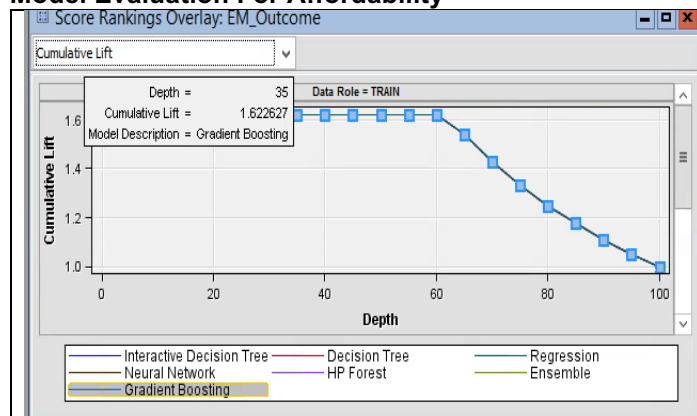
Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
EM_Outcome		_NOBS_	Sum of Frequ...	2044	886	.
EM_Outcome		_MISC_	Misclassificati...	.0004892	0.001129	.
EM_Outcome		_MAX_	Maximum Abs...	0.998239	0.998239	.
EM_Outcome		_SSE_	Sum of Squar...	1.996479	1.994731	.
EM_Outcome		_ASE_	Average Squa...	.0004884	0.001126	.
EM_Outcome		_RASE_	Root Average ...	0.022099	0.033551	.
EM_Outcome		_DIV_	Divisor for ASE	4088	1772	.
EM_Outcome		_DFT_	Total Degrees...	2044	.	.

Evaluate and Improve the Models in SAS EM (two page limit / page 1)

Model evaluation is the final stage of predictive model building which involves the use of some model evaluation criterion to check which predictive algorithm works best to predict the target variables. In this case where we have target variables as both categorical and numerical. Therefore, misclassification rate and Averaged squared error will be very important criterion in evaluating our model. This is event in the SAS EM fit statistic plot below.

Model Evaluation For Affordability



In the score ranking overlay plot shown, it's evident that the **gradient boosting** model is the **best model**. At depth 35, the cumulative fit score is 1.622.

Also, from the fit statistic table below, the gradient boosting has the lowest misclassification rate. Which means that during classification of house into affordable or not affordable, it can only misclassify affordable houses at 0.097 rate and the average error is even way lesser with 0.019.

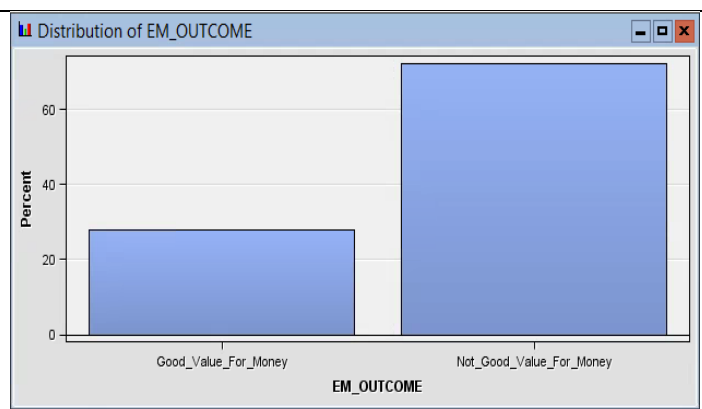
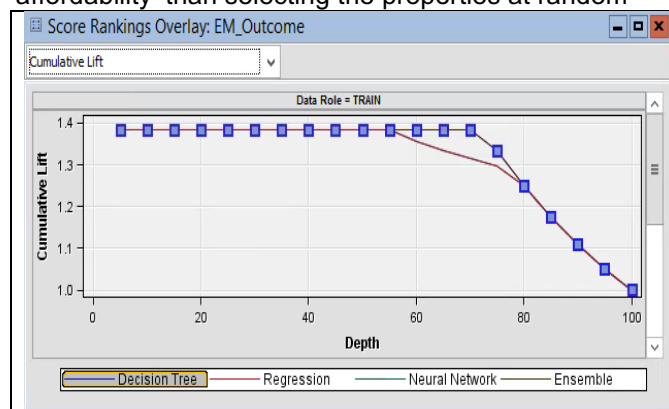
Fit Statistics for Affordability

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Average Squared Error	Train: Divisor for ASE	Train: Maximum Absolute Error	Train: Sum of Frequencies	Train: Root Average Squared Error	Train: Sum of Squared Errors	Train: Frequency of Classified Cases	Train: Misclassification Rate	Train: Number of Wrong Classifications	Train: Average Squared Error
Y	Boost	Boost	Gradient Bo...	EM_Outcome		0.097838	0.01969	10255	0.973857	2051	0.140321	201.9221		0.064846		
	Ensmbl	Ensmbl	Ensemble	EM_Outcome		0.102389	0.018721	10255	0.861933	2051	0.136824	191.9816	2051	0.044369	91	
	Tree	Tree	Decision Tr...	EM_Outcome		0.114903	0.030116	10255	0.990854	2051	0.173539	308.8361		0.10629		
	Tree2	Tree2	Interactive ...	EM_Outcome		0.114903	0.030116	10255	0.990854	2051	0.173539	308.8361		0.10629		
	HPDMLForest	HPDMLForest	HP Forest	EM_Outcome		0.117179	0.028776	10255	0.919998	2051	0.169636	295.103	2051	0.068747	141	
	Reg	Reg	Regression	EM_Outcome		0.117179	0.001649	10255	0.965183	2051	0.040603	16.90668		0.003413		
	Neural	Neural	Neural Net...	EM_Outcome		0.243458	0.053306	10255	0.892127	2051	0.230881	546.6557		0.238908	490	

Model Evaluation for Value for Money

After the model evaluation was done for **Value for money** target. it's evident that the **Decision Tree** model is the **best model**. This is because it has the smallest Misclassification rate of 1.33%.

Also, the cumulative lift plot for the decision tree is 19.06205. This is identifying that the decision tree model makes it 19 times more likely that we can identify houses that have a particular style, and coupled with a sales price shows 'affordability' than selecting the properties at random



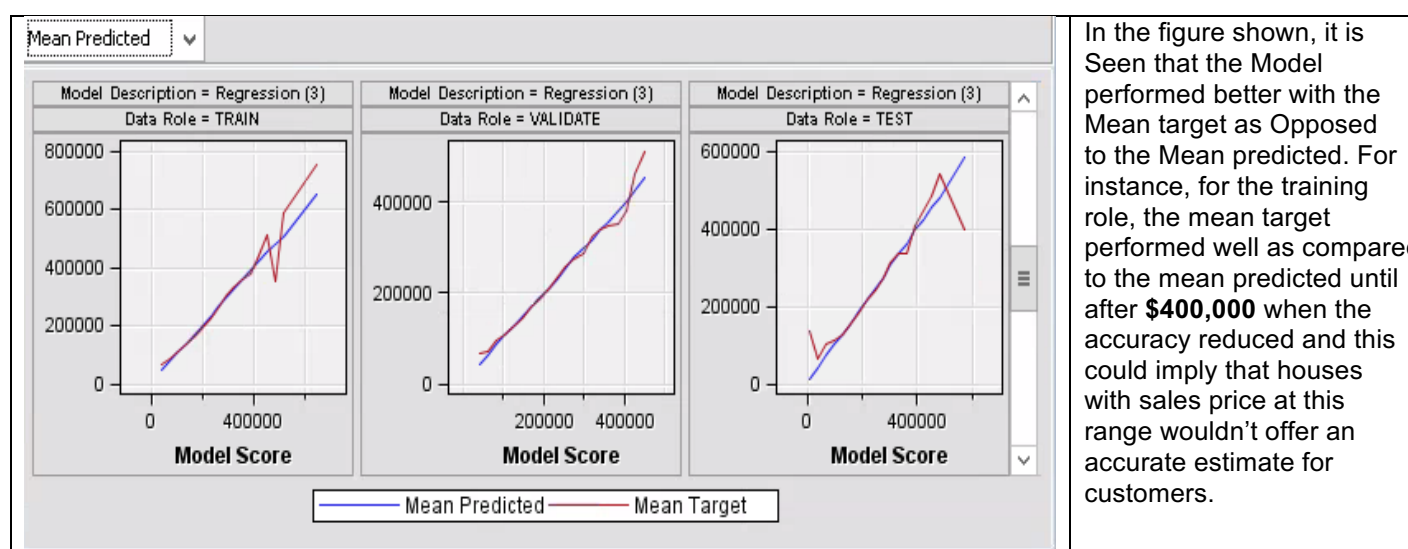
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error
Y	Tree	Tree	Decision Tr...	EM_Outcome		0.001129		.0004884				2044	4088			0
	Reg	Reg	Regression	EM_Outcome		0.068849	1024.434	0.041296	0.125351	1788	256	2044	4088	512.434	0.053122	0
	Ensmbl	Ensmbl	Ensemble	EM_Outcome		0.069977		0.026675					4088			0
	Neural	Neural	Neural Net...	EM_Outcome		0.071106	2214.017	0.041427	0.127206	1197	847	2044	4088	520.0172	0.100054	0

Evaluate and Improve the Models in SAS EM (two page limit / page 2)

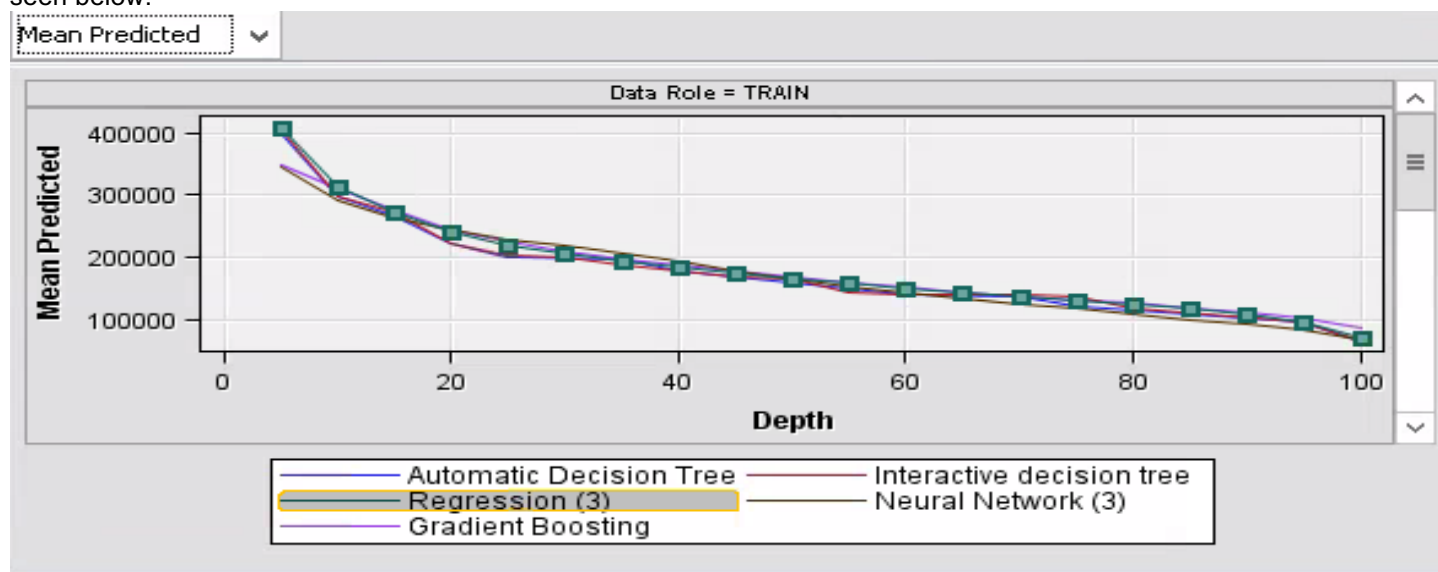
Model evaluation for sale price

After the model evaluation was done for **Sale Price** target, it's evident that the **Regression** model is the **best model**. As seen in the fit statistics table below, the Average Squared error is very low with 5.9735E8 as compared to that of gradient boosting with 8.0516E8 and others below. We could conclude that when using regression modelling for prediction for sales price, the margin of error for predicted sales estimates will be $\pm 5.9735E8$.

Fit Statistics								
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Sum of Frequencies	Train: Mean Absolute Error
Y	Reg3	Reg3	Regression...	SalePrice	SalePrice	5.9735E8	1172	
	Boost	Boost	Gradient Bo...	SalePrice	SalePrice	8.0516E8	1172	
	Tree3	Tree3	Interactive d...	SalePrice	SalePrice	1.0866E9	1172	
	Tree4	Tree4	Automatic ...	SalePrice	SalePrice	1.2026E9	1172	
	Neural3	Neural3	Neural Net...	SalePrice	SalePrice	1.3927E9	1172	

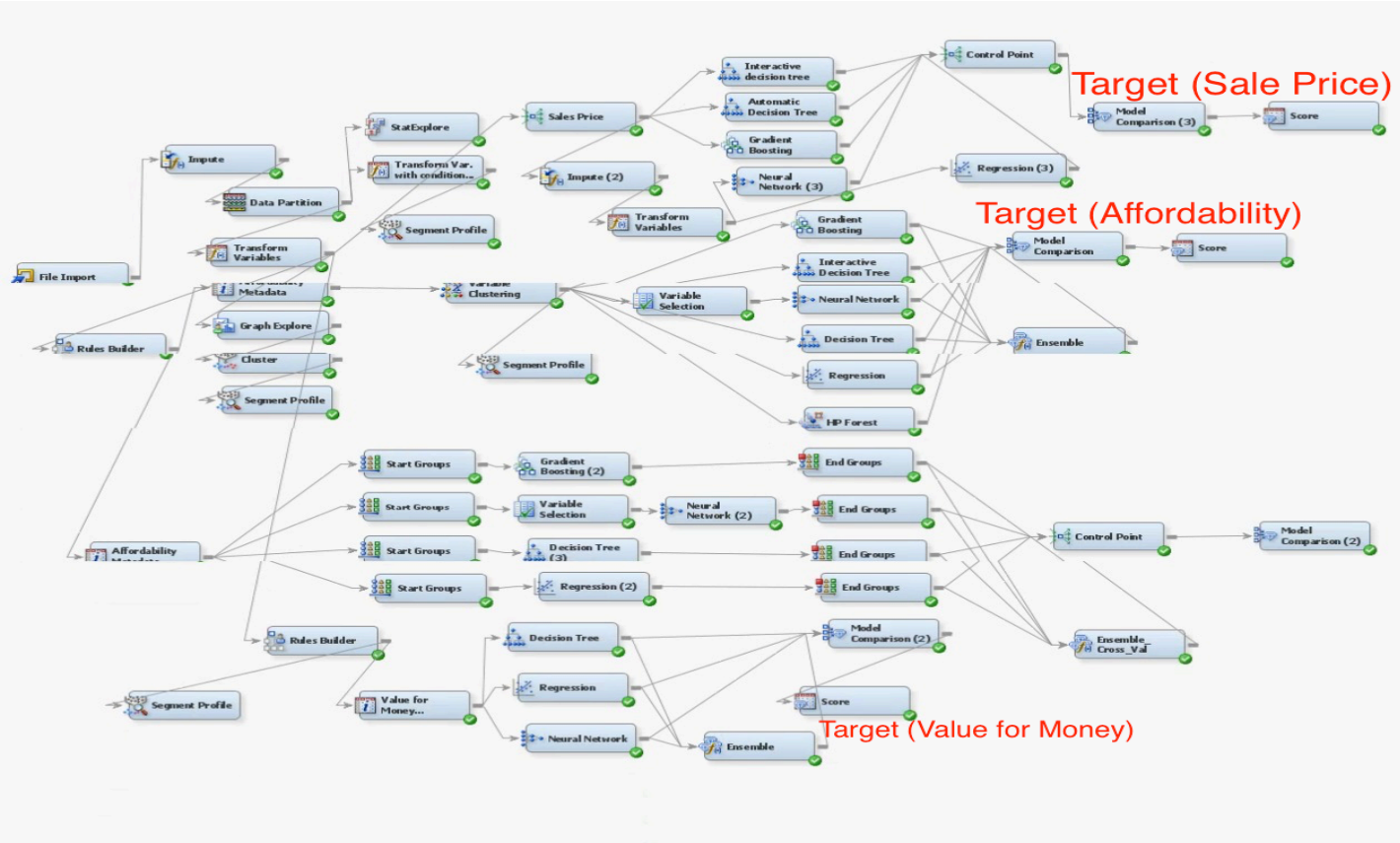


In the score ranking overlay for Sale price, during training, regression performance has been substantially better than other models by some margin even though all predictive models used appear to be closely related in their output as seen below.



Provide an Integrated Solution in SAS EM (two page limit / page 1)

Integrating all analytic process into a process can be seen in the single diagram below where all analytics process for all three target are carried out, starting from a single node (input node). Starting from the top where the target is sales price, when the score node is run by client, it automatically runs for target as Sale price and utilizes all the models in the SAS diagram below to build predictive models based on each algorithm, do a model evaluation to rank the best models and then score the final outcome. This is done for targets Affordability and Value for money respectively. This whole diagram can be seen below.



Highlighting some very important steps leading to the final model in order to provide easy to implement steps to client is very important. In order to get the best out of the model used, I'd recommend the use of cross validation to ensure a much more accurate result. This is because of the working principle of cross validation works in such a way that data is split into several k parts and for validation, while the remaining k-1 parts are used to fit the model. This ensure a more accurate predictive model as the k validation is continually carried out for all k parts until all the parts have been utilized for validation. Also, prediction error is greatly reduced.

From the outcomes below, the comparison between the outcome of the models built out with cross validation and the outcome of the models built without cross validation can be seen using the misclassification rate as a criterion. Without cross validation, gradient boosting has a misclassification rate of **0.0978** but with cross validation there's a huge reduction in misclassification as the misclassification rate went down to **0.0819**.

Without Cross-Validation

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
High misclassification without Cross validation						
Y	Boost	Boost	Gradient Bo...	EM_Outcome		0.097838
	Ensmbl	Ensmbl	Ensemble	EM_Outcome		0.102389

Provide an Integrated Solution in SAS EM (two page limit / page 2)

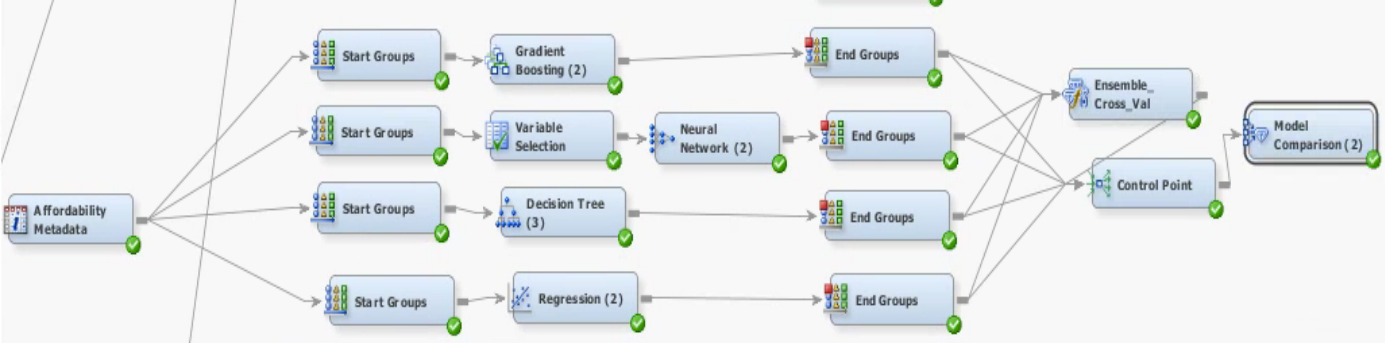
With Cross-Validation

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
	EndGrp	EndGrp	End Groups	EM_Outcome		0.081911
	Ensmbl2	Ensmbl2	Ensemble...	EM_Outcome		0.09215

In addition, implementing ensemble model to cross validation boost its efficiency. Ensemble model is a very robust model that plays a huge role in model building. Its flexibility is in such a way that it trains several similar models and combines their result to improve accuracy, reduce bias and produce very robust models. It's evident in the model comparison as seen above.

The figure below shows how cross validation was carried out in the SAS diagram and how ensemble was included thereafter.

Cross-Validation Screenshot



Finally, the **score node** generates a score code on the trained data diagram path. After the score node has been run, the output will be used for operational purpose.

Output						
31						
32	Variable Name	Role	Creator	Comment	Label	Variable Hidden
33						
34	Alley	INPUT			Alley	N
35	Bedroom_AbvGr	INPUT			Bedroom_AbvGr	Y
36	Bldg_Type	INPUT			Bldg_Type	N
37	BsmtFin_SF_1	INPUT			BsmtFin_SF_1	Y
38	BsmtFin_SF_2	INPUT			BsmtFin_SF_2	Y
39	BsmtFin_Type_1	INPUT			BsmtFin_Type_1	Y
40	BsmtFin_Type_2	INPUT			BsmtFin_Type_2	Y
41	Bsmt_Cond	INPUT			Bsmt_Cond	Y
42	Bsmt_Exposure	INPUT			Bsmt_Exposure	Y
43	Bsmt_Full_Bath	INPUT			Bsmt_Full_Bath	Y
44	Bsmt_Half_Bath	INPUT			Bsmt_Half_Bath	Y
45	Bsmt_Qual	INPUT			Bsmt_Qual	Y
46	Bsmt_Unf_SF	INPUT			Bsmt_Unf_SF	Y
...						
Output Variables						
	Variable Name	Creator	Variable Label	Function	Type	
	Clus1	VarClus		TRANSFORM	N	
	Clus10	VarClus		TRANSFORM	N	
	Clus11	VarClus		TRANSFORM	N	
	Clus2	VarClus		TRANSFORM	N	
	Clus3	VarClus		TRANSFORM	N	
	Clus4	VarClus		TRANSFORM	N	
	Clus5	VarClus		TRANSFORM	N	
	Clus6	VarClus		TRANSFORM	N	
	Clus7	VarClus		TRANSFORM	N	
	Clus8	VarClus		TRANSFORM	N	
	Clus9	VarClus		TRANSFORM	N	
	EM_CLASSIFICATION Score		Prediction for Afford...	CLASSIFICATION	C	
	EM_EVENTPROBA...	Score	Probability for level ...	PREDICT	N	
	EM_PROBABILITY	Score	Probability of Classi...	PREDICT	N	
	EM_SEGMENT	Score	Node	TRANSFORM	N	
	G_Condition_1	Ensmbl	Grouped Levels for ...	TRANSFORM	N	

From the score node outputs seen in the figure, the output window shows all the variables used in running the predictive models, variable name, their roles, and status.

Just below the output window is the output variables window which shows all the inputs. it's seen that the clusters were added to the input during model building, then the function used for modifying the data to specified requirements is known as transformation. Other variables used include segments, conditions etc. The functions used are prediction and transformations well.

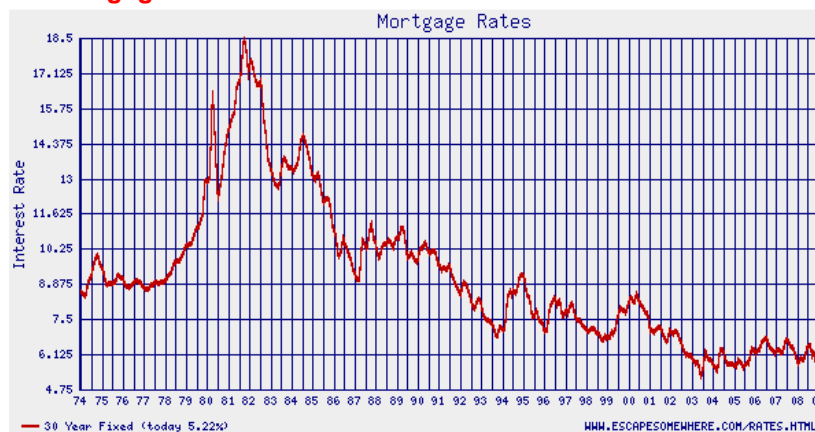
Then the type of variables they are is represented as N and C which means numeric and categorical respectively. The list continues downwards until all variables are used for model building.

Further Research and Extensions in SAS EM (one page limit)

Apart from the variables available in the dataset, there are other factors that could impact on the house prices in Ames.

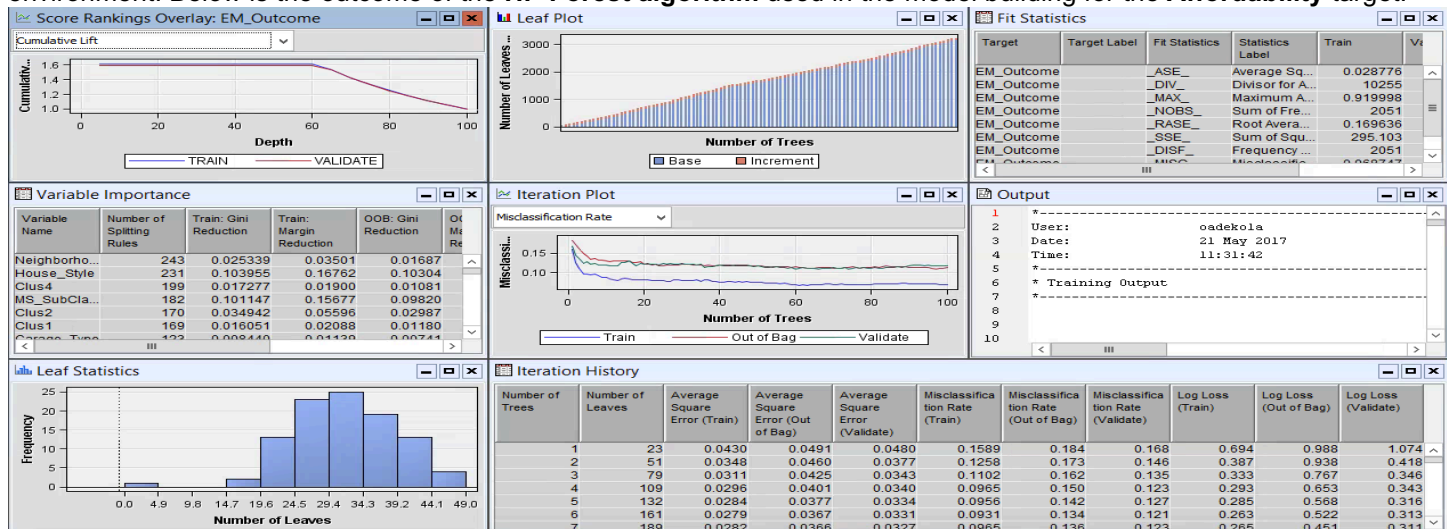
- Strong economic growth and rising incomes could help people to spend more on houses, and the rising demand will push up the house prices. The low economic growth, for example, in 2007-2008 financial crisis will reduce the demand to purchase properties. Analysis of data of economic growth will certainly help to predict the sales prices, affordability and value for money in the future.
- Rising unemployment rate and even the fear of unemployment may discourage people from entering the property market, which will then influence the sales prices of property market. Data of unemployment rate from 2006 to 2010 could be a variable in the model.
- Mortgage rate could affect demand for properties and affordability. Referring to the chart below, the interest rates were between 6 and 7 percent during 2006-2008, and there was a sharp drop after Mid-2008. Due to the financial crisis around that period, US might lower interest rate to boost property market. However, it's interesting to see if lowering mortgage rate influences the house prices in Ames.
- Also, reduction in currency value as at the time the data was gathered as compared to the prediction. All these factors need to be included to ensure accuracy prediction accuracy (SAS, 2017).

The time series below shows how increase in years from the period of the 70s influences interest rates on mortgages



Another approach can be getting external data that adds the influence of inflation to the target and therefore makes prediction accuracy increase.

In addition, talking about another solution by SAS to improve predictive modelling, after further research was done, these are solutions provided by **SAS (High Performance Data Mining)**. These performance analytics are designed especially for big data by reducing response time of predictive models. These HPDM models are faster in terms of computation and storage than normal models because they make use of parallel processing technique in a distributed computing environment. Below is the outcome of the **HP Forest algorithm** used in the model building for the **Affordability** target.



The **HP forest** result above, which is a high performance random forest, shows the leaf plot, cumulative lift, the leaf statistics (with a slightly skewed plot) and also the list of variables by importance. It's seen that its average squared error (**0.0430**) is lower than the averaged squared error of the normal random forest algorithm.

References

1. Propertylogy.com. (2017). *10 Timeless Factors That Affect Property Price | Propertylogy*. [online] Available at: <http://www.propertylogy.com/knowledge/10-timeless-factors-that-affect-property-price/>
2. Tax.ny.gov. (2017). *How to Estimate the Market Value of your Home*. [online] Available at: https://www.tax.ny.gov/pubs_and_bulls/orpts/mv_estimates.htm
3. SAS. (2017). *Base SAS(R) Software fact sheet*. [online] Available at: <http://www.sas.com/technologies/bi/appdev/base/factsheet.pdf>
4. Perry, G.L., Schultze, C., Solow, R. and Gordon, R.A., 1970. Changing labor markets and inflation. *Brookings Papers on Economic Activity*, 1970(3), pp.411-448.
5. Communities.sas.com. (2017). *SAS High-Performance Analytics tip #1: How it differs from SAS Grid & SAS In-Memory Analytics*. [online] Available at: <https://communities.sas.com/t5/tkb/articleprintpage/tkb-id/library/article-id/1243>