**SCORING THE TOP 20 MOST LUXURIOUS HOUSES**

**USING PREDICTIVE ANALYTICS**

Anushka Nair

University of Houston

Spring 2020

**Phase 1:**

**Business/Research Understanding Phase**

* The goal of the project is to apply predictive analytics techniques to predict and rank the sale price and recommend 20 of the most luxurious homes out of 100 candidate properties for which the sales prices are not provided.
* We are provided with 4 raw data files containing information from an anonymous United States city assessor’s office located in the North West region. The values in the data files are for individual residential real estate properties sold in that city over a time period of 4 years .
* We are also provided with a Score data set which consists of many of the variables from the raw datasets except for the Sale Price and a few other variables. A few properties might be missing for a few values.
* We are acting as a set of investors who plan to purchase 20 residential properties in the city where the raw data files were collected. The broker of the properties gave us a list of 100 properties and their information as your Score data set, consisting of many assessed values for them. We need to develop models and appropriate procedures to predict and rank the sale price and recommend the 20 most luxurious houses. The reason for investing in the top luxury houses is that their potential investment growth is believed to be the greatest.

**Phase 2**

**Data Understanding Phase**

* **Variables Description**

1. PID (Property Identification) – a unique number to identify each property.

2. Lot Area – The size of the lot, measured in square feet, on which the house is located.

3. Lot Shape – The general shape of the lot. A lot with a regular shape has a value of 1, and another with not a regular shape has a value of 0.

4. Bldg Type (i.e., Building Type) – This describes the type of home in terms of its footprint. A single-family detached type of home is indicated by a value of 1, and a townhouse type of home is indicated by a value of 0.

5. Overall Quality – This is a rating of the overall material and finish of the house. The numeric scale of this rating is as follows. 10 - Very Excellent 9 - Excellent 8 - Very Good 7 -Good 6 - Above Average 5 - Average 4 - Below Average 3 - Fair 2 - Poor 1 - Very Poor

6. Overall Condition: This is a rating of the overall condition of the house. The numeric scale of this rating is as follows. 10 - Very Excellent 9 - Excellent 8 - Very Good 7 - Good 6 - Above Average 5 - Average 4 - Below Average 3 - Fair 2 - Poor 1 - Very Poor

7. Exterior Quality – This is a rating of the quality of the material on the exterior. A good quality is indicated by a 1, and an average quality is indicated by a 0.

8. Foundation – This describes the type of foundation upon which the house is built. A concrete foundation is indicated by a value of 2; a cinder-block foundation by a value of 1; and brick foundation by a value of 0.

9. Year Built – This describes the year when the house was constructed.

10. Year Remodel – This describes the year when the house was remodeled. If the house was never remodeled, then the “year remodel” is the same as the “year built.”

11. Veneer Area of Exterior Wall – This describes the area in square feet of the exterior wall that is veneer.

12. Bsmt Finish Type (Basement Finished Type ) – This indicates whether a home’s basement is finished or not in the sense that it can be lived in or not. When it is finished, it has a value of 1, and a value of 0 otherwise.

13. Basement Finished Sqr ft – This is the measure of the area of a finished basement.

14. Basement Unfinished Sqr ft – This is the measure of the area of an unfinished basement.

15. Total Bsmt Sqr ft – This is the measure of the total basement area.

16. Heating QC (Heating Quality Condition) – This is a measure of the rating of how well the heating unit is for a house. The rating scale is as follows. 3 - Excellent 2 - Good 1 - Average 0 - Fair

17. 1st Flr Sqr ft (First floor Sqt ft) – This is a measure of the living space on the first floor of a house.

18. 2nd Flr Sqr ft (Second floor Sqt ft) – This is a measure of the living space on the second floor of a house.

19. Above Ground Living Area – This is a measure of the living space of the entire house, excluding the basement.

20. Number Full Bath Bsmt - This indicates the number of full bathrooms in the basement of a house. A value of 1 indicates that there is a full bathroom and a value of 0 indicates that there is not a full bathroom in the basement.

21. Half Bath House - This indicates whether there is a half bathroom in the house (excluding the basement). A value of 1 indicates that there is a half bathroom and a value of 0 indicates that there is not a half bathroom in the house.

22. Number Full Bath House - This indicates the number of full bathrooms there are in the house, not including bathroom in the basement.

23. Bedroom Above Ground - This indicates the number of bedrooms there are in the house, not including the basement.

24. Number Room Above Ground - This indicates the number of rooms there are in the house, not including the basement.

25. Fireplaces – This indicates the number of fireplaces there are in the house, not including the basement.

26. Garage Type – Whether there is a garage of a given type is described and indicated as follows. 3 - Attached to house 2 - Built-In (Garage part of house - typically has room above garage) 1 - Detached from home 0 - No garage

27. Garage Cars – This indicates the number of cars that can be accommodated in the garage of the house.

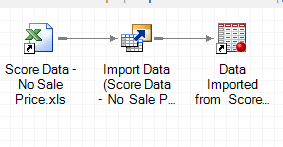
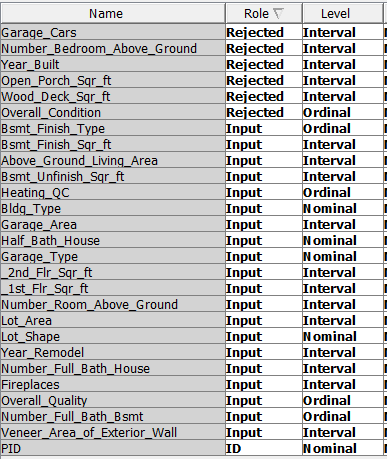
28. Garage Area – This is the size of garage in square feet.

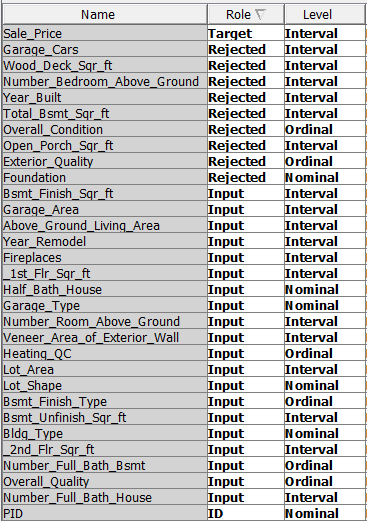
29. Wood Deck Sqr ft – This is the size of the wood deck area in square feet for a house.

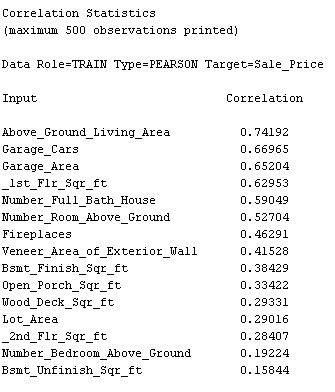
30. Open Porch Sqr ft - This is the size of the open porch area in square feet for a house.

31. Sale Price – This is the sale price of a house (not included in the Score data set)

* **A close up of a map

  Description automatically generatedData Management Procedures**:
  + I appended the Property Survey 1 and 2 tables, followed by joining it with the Quality Assessment and House feature on PID.
  + I went ahead and converted the Score dataset on SAS Enterprise Guide and also changed the Roles and levels to match the original dataset.
  + 
* **Final Variables-Names & Configuration**



* I have explained why I chose to drop certain variables in the next couple pages.
* I set PIDs role to ID since it is a unique identifier, and Sale\_Price to target since that is what we are seeking to predict.
* I changed the levels of Heating QC, Overall Quality, Number Full Bath Basement and Basement finish type to ordinal, because when I looked at their graphs and descriptions, they were ranked.
* I changed the levels of Building type, Garage Type, Half Bath House, and Lot shape to Nominal because from the variable descriptions, they were categorical.
* I rejected Open Porch sqr ft, Overall Condition, Wood Deck Sqr ft, Exterior quality, Foundation, Garage Cars, Number bedroom above ground, Year Built andTotal Bsmt sqr ft, so ignore the levels for those variables.
* **Exploratory Data Analysis**
* After comparing the score dataset and the raw datasets, I decided to drop the Exterior quality, foundation and total bsmt columns because the score dataset does not contain these columns.
* I decided to reject year\_built, since year\_remodeled would be a better indicator of sale price and having both variables would lead to collinearity.
* Looking at the correlations, Above\_Ground\_Living\_Area, Garage\_Cars, Garage\_Area, \_1st\_Flr\_Sr\_ft, Number\_full\_bath\_house, Number\_Room \_Above\_Ground seem to be the most important, and I will definitely choose to keep them in the model.
  + However, I will drop Garage\_Cars because Garage\_cars and Garage\_area depend on each other and hence would be colinear. The histograms below show the distribution of mean Sale price by Number of garage cars and Garage area respectively. As we see, The garage area seems to have more of a clear consistent trend compared to number of cars, which is why I chose to keep Garage Area.

A picture containing screenshot

Description automatically generated A picture containing fence

Description automatically generated

* I generated a few plots to visualize and look for the predictors that seem to have the most obvious trends. Above Ground living area, Number of rooms above ground, Overall quality, Heating QC, Number of full bath houses, and bsmt finish sqr ft seemed to have the strongest trends, so they will be part of my final model.

A picture containing fence

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

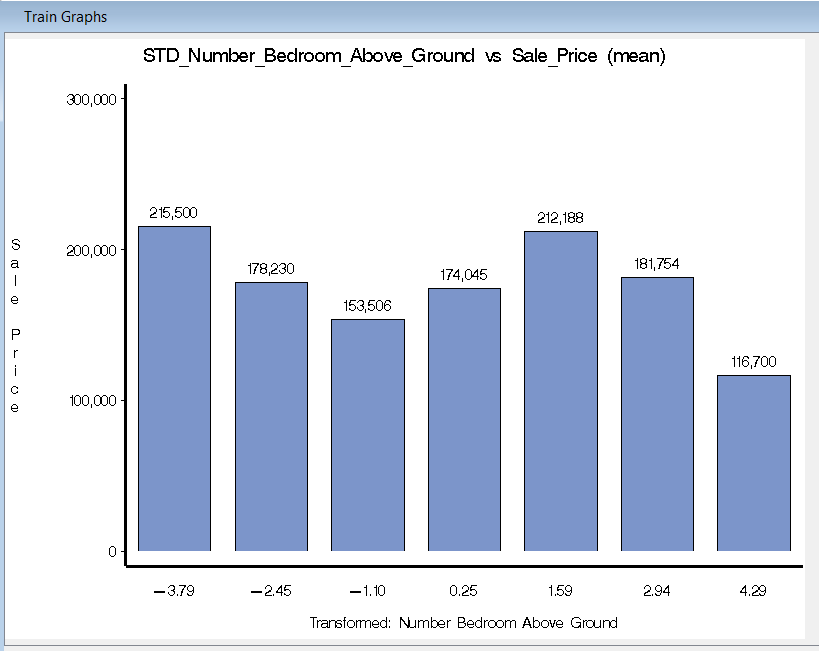
A picture containing screenshot

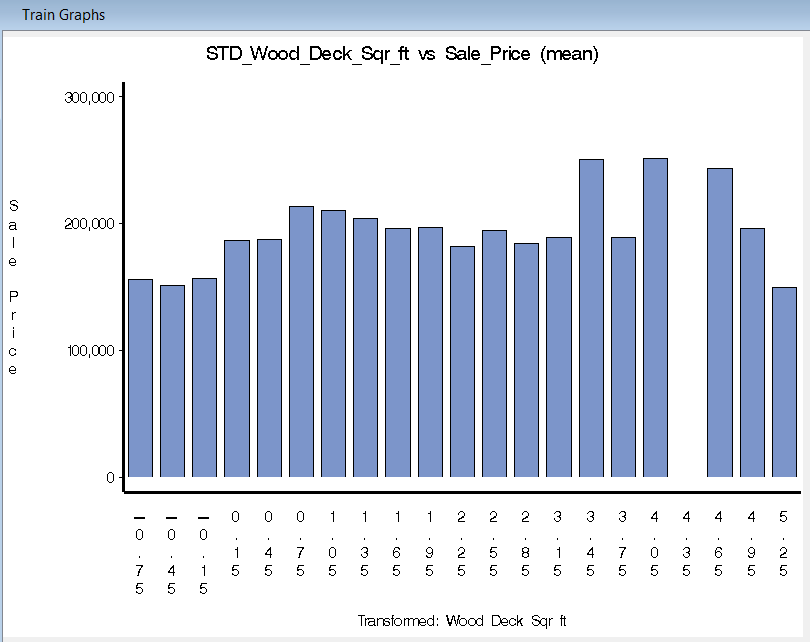
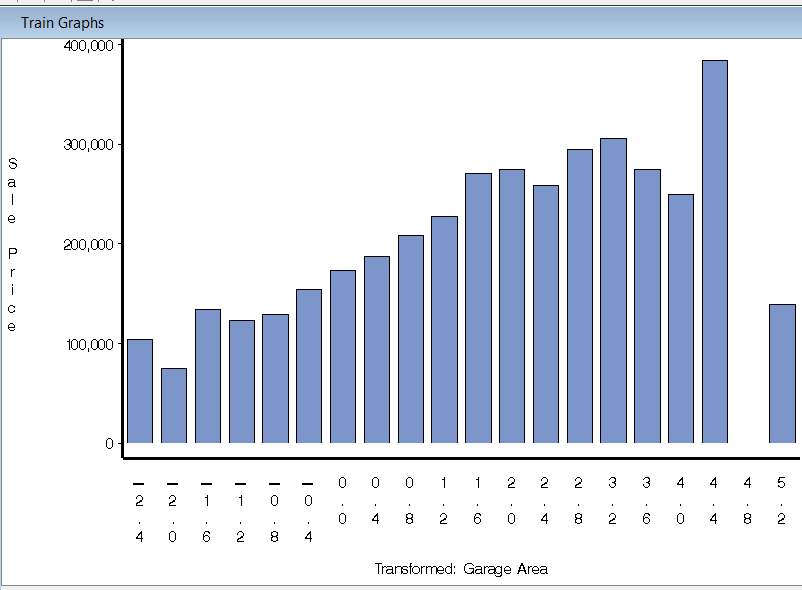
Description automatically generatedA picture containing fence

Description automatically generated

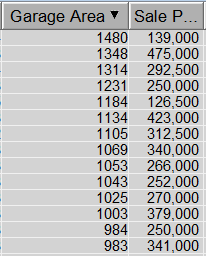
* A few predictors didn’t seem to have any obvious trends, for example, Open Porch Sqr ft, Number bedroom above ground, and wood deck sqr ft, so I dropped them. Similarly, garage trend seems to have an outlier towards the end, so it might also be taken out of the final model.
  + However, Number of rooms above ground would contain number of bedrooms above ground, so I decided to drop number of bedrooms above ground. The sale price by bedroom above ground doesn’t seem to match intuition, it seems as though more bedrooms cause a lower sale price which does not make sense, especially because number of rooms above ground seemed to tell us the opposite.

A screenshot of a cell phone

Description automatically generated

* The Garage Area seemed to have a couple outliers, as we can see in the picture above. To fix it, I went ahead and took a look at the Highest garage prices.

As we see, for the highest garage prices, the sale price values were consistently above >$250k, except for the sale price of 139,000 and 126,500. I decided to manually replace those values with the approximate mean.

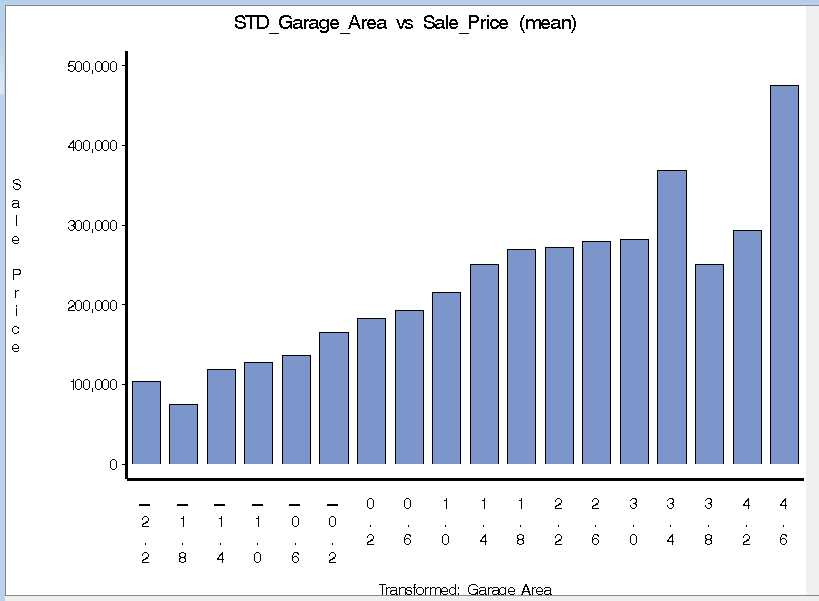
I replaced the 1480 garage area with 404, and 1184 with 294.

I used the garage area values that corresponded with the sale price to calculate the mean. Pictures of the values used are shown below. (I did not include the observation that needed to be changed to calculate the mean, just the values around it)

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

* As we can see, in the before versus after graph below the trend is more consistent.

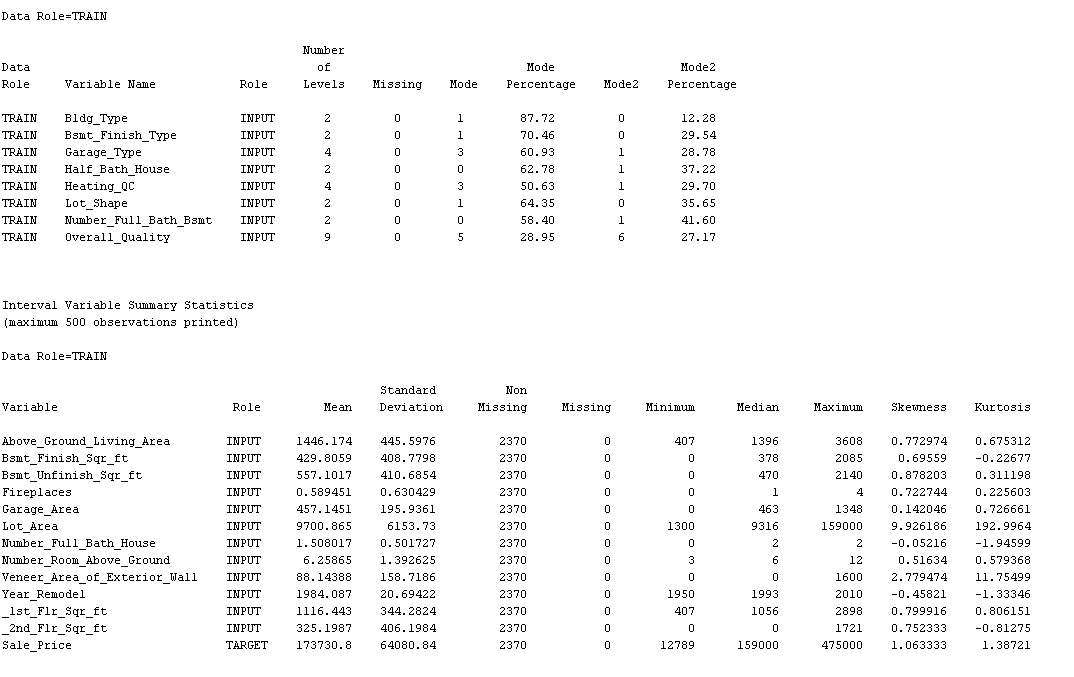
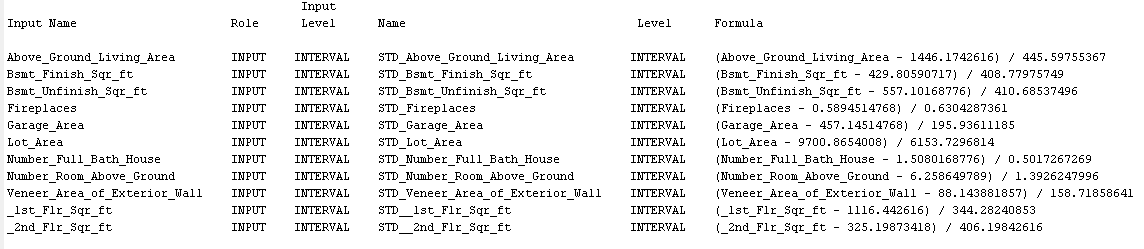
A picture containing fence

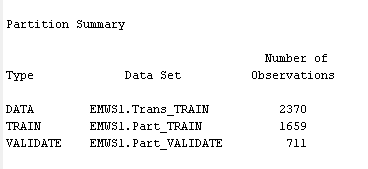
Description automatically generated

* I realize that the way I implemented it might be problematic, but I could not figure out how to use SAS EM for this, I apologize.
* When I ran a basic regression model on the data, I noticed that the top 20 predicted seemed to mostly have overall condition = 5 (screenshot shown below). After investigating further, I saw that the overall condition vs sale price seemed to be concentrated at 5, which does not look like a good measure to use in our final model since overall condition is slightly arbitrary. I decided to reject it.

**Phase 3**

**Data Preparation Phase**

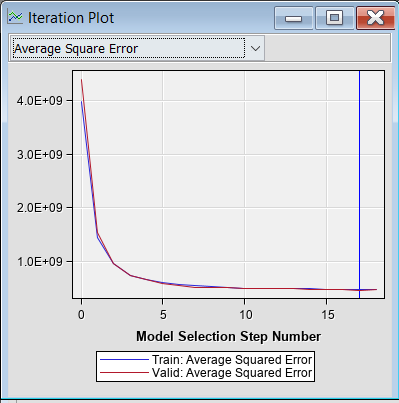
* I used the Stat explore node to check for missing values, but the results showed no missing values, so there was no need for data imputation. (screenshot below)
* I added a Transform variables node. Under Default methods, I went ahead and standardized all interval inputs, except for year\_remodeled.(screenshot below)
* Under Sample properties, I chose method to be random and size to be max.
* I added a data partition node, with training as 70% and testing as 30%.



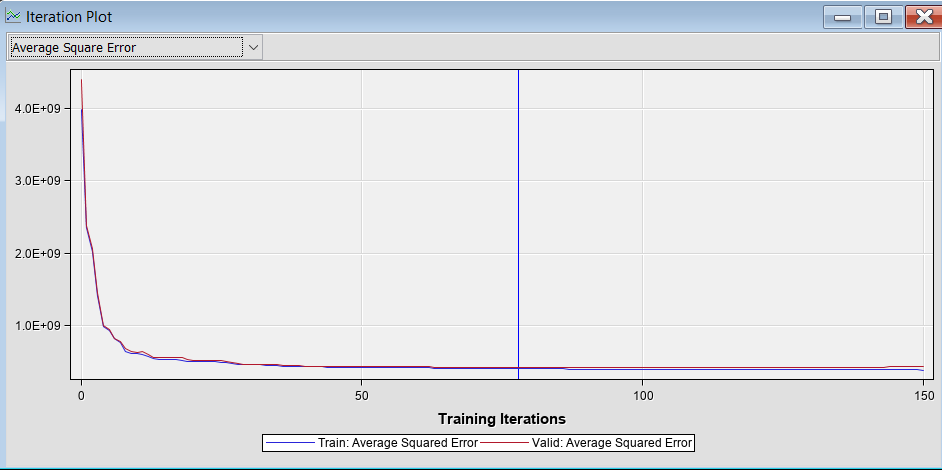
**Phase 4**

**Modeling Phase:**

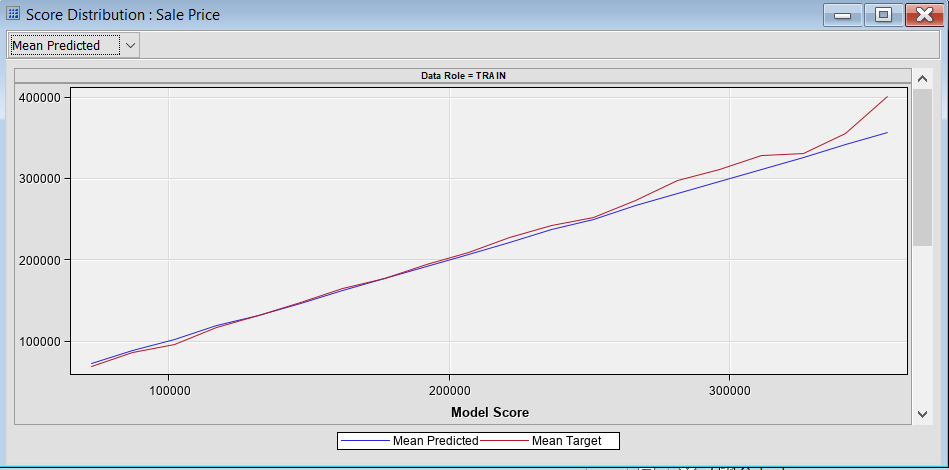
* **Regression**:
* Under regression, selected forward under model selection.
* I changed the selection criterion to Validation error, changed selection defaults to No and under selection options, I changed the entry and stay significance levels to 0.1.
* The average squared error for this model is 4.6424x10^8
* I went ahead and changed two factor interactions to Yes, but that increased the Average squared error on the validation set= 4.857x10^8
* Changing the entry and significance levels to 0.5 seemed to make no difference, so I did not use them.



* My final regression model had entry and stay significance levels of 0.1 and two factor interactions set to No, with a validation average squared of 4.6424x10^8
* **Neural network:**
  + I added the neural network node from models.
  + Under train, for model selection criterion, I picked Average error. I clicked to open the optimization window and changed maximum iterations to 150. Under preliminary training, I changed enable to no.
  + Under network, I chose number of hidden units to be 8, and the validation average square error was 4.479x10^8.
  + I went ahead and increased the number of hidden units to be 25, but the error increased to 4.8849x10^8.
  + I finally changed the number of hidden units to 10, and the Validation ASE reduced to 4.2008x10^8, which I used as my final Neural Network model.



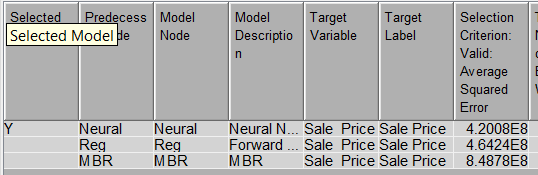
* **K-nearest-neighbors**
* I dragged the MBR node.
* I started with K = 4, which gave me a Validation ASE of 8.769x10^8.
* I increased my K to 8, which reduced the Validation ASE to 8.58x10^8.
* Changed K=6 which further reduced the Validation ASE to 8.488x10^8 so that is the final KNN model.
* I couldn’t find an ASE plot for KNN so I used the score matrix plot.



**PHASE 5**

**Evaluation Phase**

* I dragged the model comparison node under Assess and connected the three competing models to the node.
* Under selection statistic, I picked Average Squared error, and for selection table I picked Validation.



* Running the node provided the output above. The Neural Network is the optimal model in terms of Average Squared error with an ASE of 4.2008x10^8.
* K-nearest neighbors has the worst Average squared error, and it makes sense because the KNN model in SAS does not use categorical variables, so it probably lost out on a lot of useful data.
  + For example, Overall Quality and Heating QC were both categorical and were strongly related to Sale Price, but K-nearest neighbors lost that valuable information.
  + KNN however is easier to implement in general, since all you mainly have to do is choose the optimal number of neighbors.
* The Regression Model did significantly better than K-nearest neighbors, and that makes sense because it had all important inputs. Regression in this context makes sense and is easier to understand since a lot of the relationships were linear, i.e the higher the square footage, fireplaces, Garage Area etc, the higher the Sale price. However, regression wouldn’t easily capture any hidden relationships or non-linear relationships as easily.
* Neural Networks did the best because it can detect complex relationships between the different input variables and the target. I decided to move forward with the neural network model as the final model.

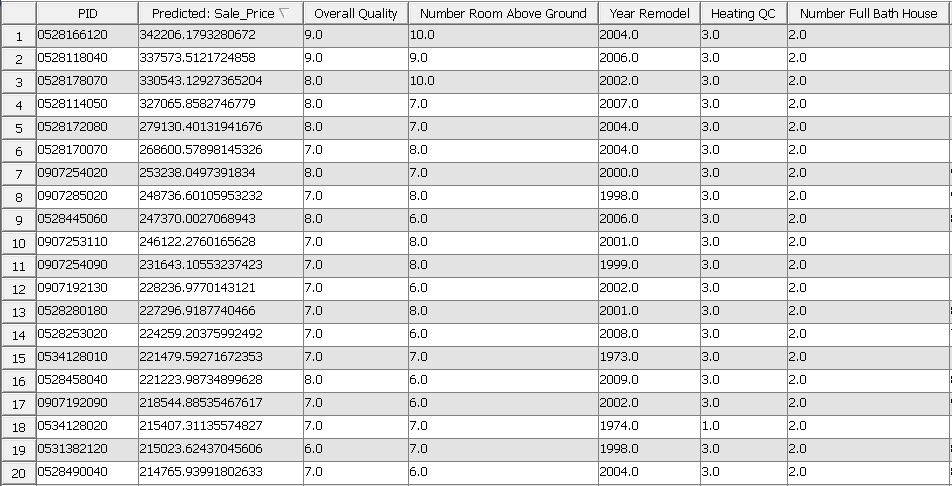
**Phase 6**

**Deployment Phase:**

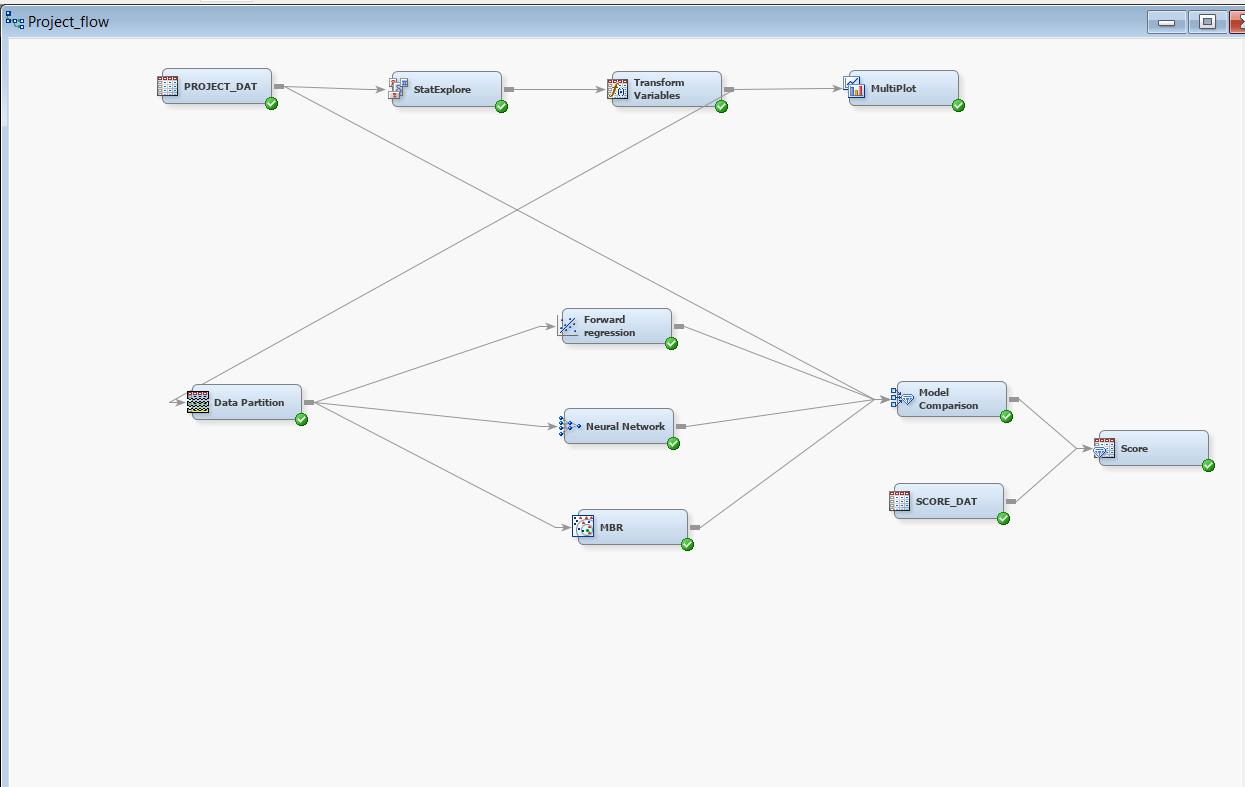
* I dragged my converted Score dataset and placed it next to the model comparison node.
* I dragged the Score node from Assess, and connected both, the model comparison node and the score dataset node to the score node.
* To finally score the dataset, under the score node I clicked on exported data, and then EMWS1.Score\_SCORE, and then browse.

**RESULTS: TOP 20 LUXURIOUS HOMES**

* Here is what the best model predicted the top 20 most luxurious houses. I added a few of the input variables in here in order to comment on a few observations I made.



* I added some of the variables to the screenshot so I can talk about some of the observations I made.
* The top 20 houses based on the neural network seemed to all have overall quality>5, which makes sense because in the original dataset, overall quality had a strong linear trend.
* The Number of rooms above ground also are >5, which makes sense intuitively.
* The heating QC is also all equal to 3, which makes sense because the value 3 meant excellent.
* Most of the year remodels seemed to be in the 2000s, which would make sense because a house remodeled recently would have a better price.
* All the values for Full Bath House are equal to 2, which was the highest value possible.

**This is what my Process Flow looked like at the end of the project.**