House Prices

Analysis and Prediction

Case of house sales in King County

Predicting house prices and examining factors by analyzing differences in various properties.

**Executive Summary**

# Overview

The project is to helps customers to estimate house values, and to identify the attributes which affect the house values. Our SAS model is especially useful to those clients who are buying, selling, and remodeling their houses. With this model, we can predict the value and the variance of the value of a house according to the location, size, and features of a property.

***Problem***

The current house prices have not been transparent, and there is no solid standards that allow house owners or buyers to objectively evaluate house values. At the time, users can only evaluate house values by accessing house price index, which cover a huge region and does not specify different attributes. They can solely depend on their own judgement built upon historical house prices, economic environment, and their own preference.

***Solution***

Our SAS model provide an easy-to-interpret formula that allows user to calculate, by inputting location, size and features of the house, the actual value of the house, and to interpret the factors that concludes the price. Beside determining the price, house owners or constructors can also interpret the best features for the houses to fit the needs of the residents or to create highest profit.

***Opportunity***

As the internet develops, the access to information easier. The customers now expect the house price to be transparent, just as goods on the shelf of supermarkets. The project meets the needs under the trend. To acquire new markets, our SAS model can be applied to different regions, in which people should have different requirements and preference on houses.

**Project Motivation/Background**

**Data Description**

***Data Overview***

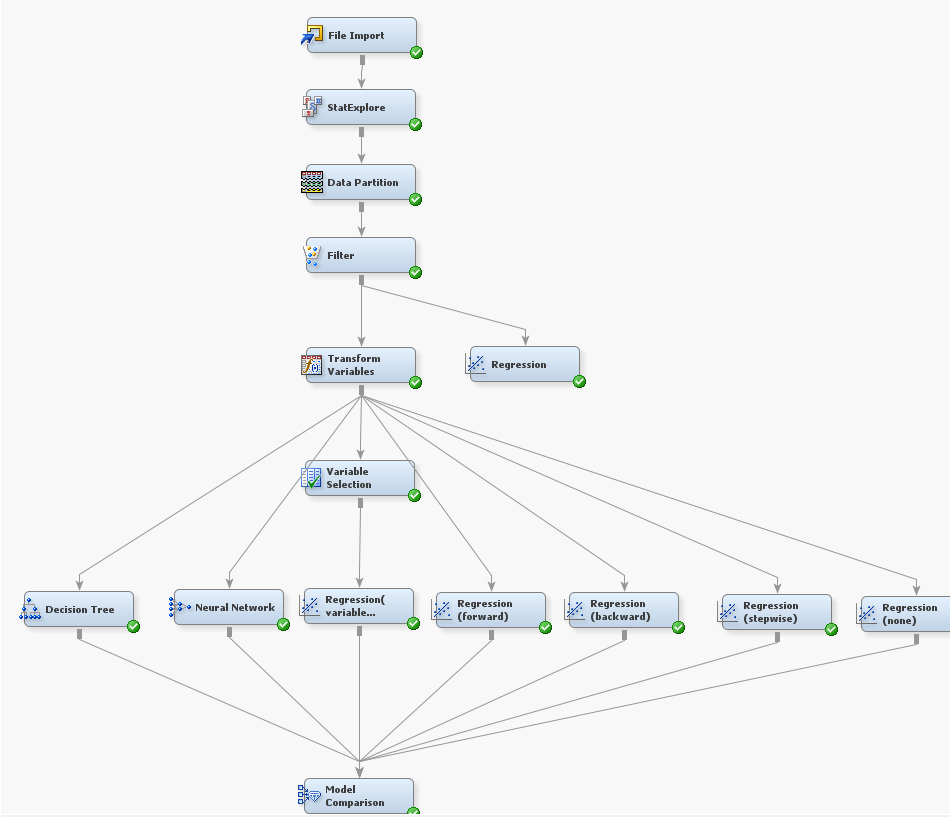
The data is second-handed data set called “House sale in King County” on Kaggle, and contains 21613 data sets, which represent houses sold between May 2014 and May 2015 in King County. Each of the data sets has 21 variables, including our target input ”price” and other contributes that determine “price”.

***Explanation of the Variables***

|  |  |
| --- | --- |
| id | Unique ID for each home sold |
| date | Date of the home sale |
| price | Price of each home sold |
| bedrooms | Number of bedrooms |
| bathrooms | Number of bathrooms, where .5 accounts for a room with a toilet but no shower |
| sqft\_living | Square footage of the apartments interior living space |
| sqft\_lot | Square footage of the land space |
| floors | Number of floors |
| waterfront | A dummy variable for whether the apartment was overlooking the waterfront or not |
| view | An index from 0 to 4 of how good the view of the property was |
| condition | An index from 1 to 5 on the condition of the apartment |
| grade | An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design |
| sqft\_above | The square footage of the interior housing space that is above ground level |
| sqft\_basement | The square footage of the interior housing space that is below ground level |
| yr\_built | The year the house was initially built |
| yr\_renovated | The year of the house’s last renovation |
| zipcode | What zipcode area the house is in |
| lat | Lattitude |
| long | Longitude |
| sqft\_living15 | The square footage of interior housing living space for the nearest 15 neighbors |
| sqft\_lot15 | The square footage of the land lots of the nearest 15 neighbors |

\*Weird number of bathroom(s): There are no official standard to “count” the number of bathrooms. Conventionally, a full bathroom with toilet, sink and tub/shower is called “1 bath”; a bathroom with only toilet and sink is called “0.5”. There are some houses with 0.25, 1.25 or 1.75 bathrooms. From our own research, we believe that a 0.25 bathroom is a toilet without sink. This weird situation happens in old houses with extended bathroom that built years after the main construction.

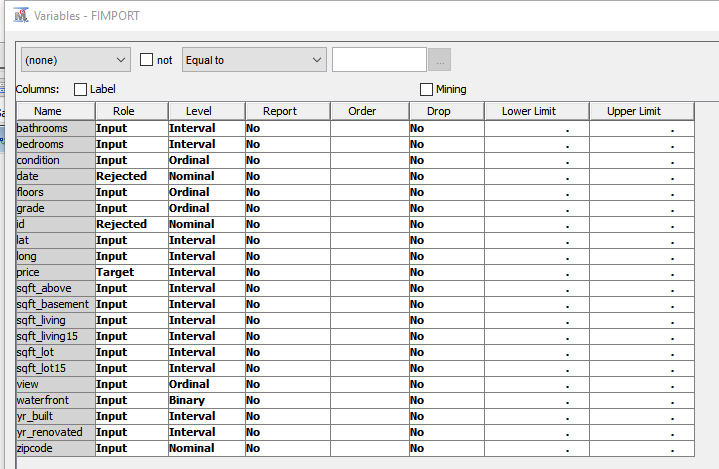
**BI model/Enterprise Miner Diagrams**

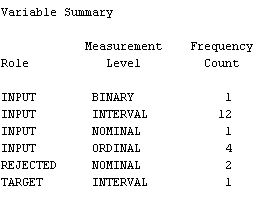


1. File Importing.

(1) Using a File Import node to import data that is stored in external formats into a data source that SAS Enterprise Miner can interpret. Here our house price dataset is a CVS flat file.

(2) Set the model role and measurement level for each variable.

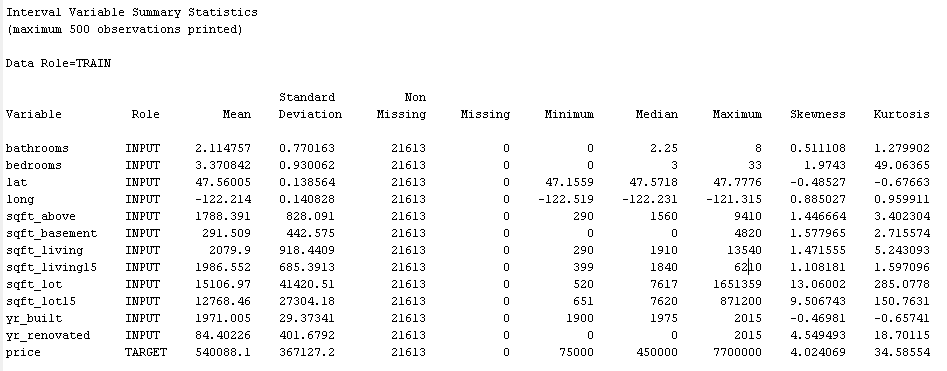


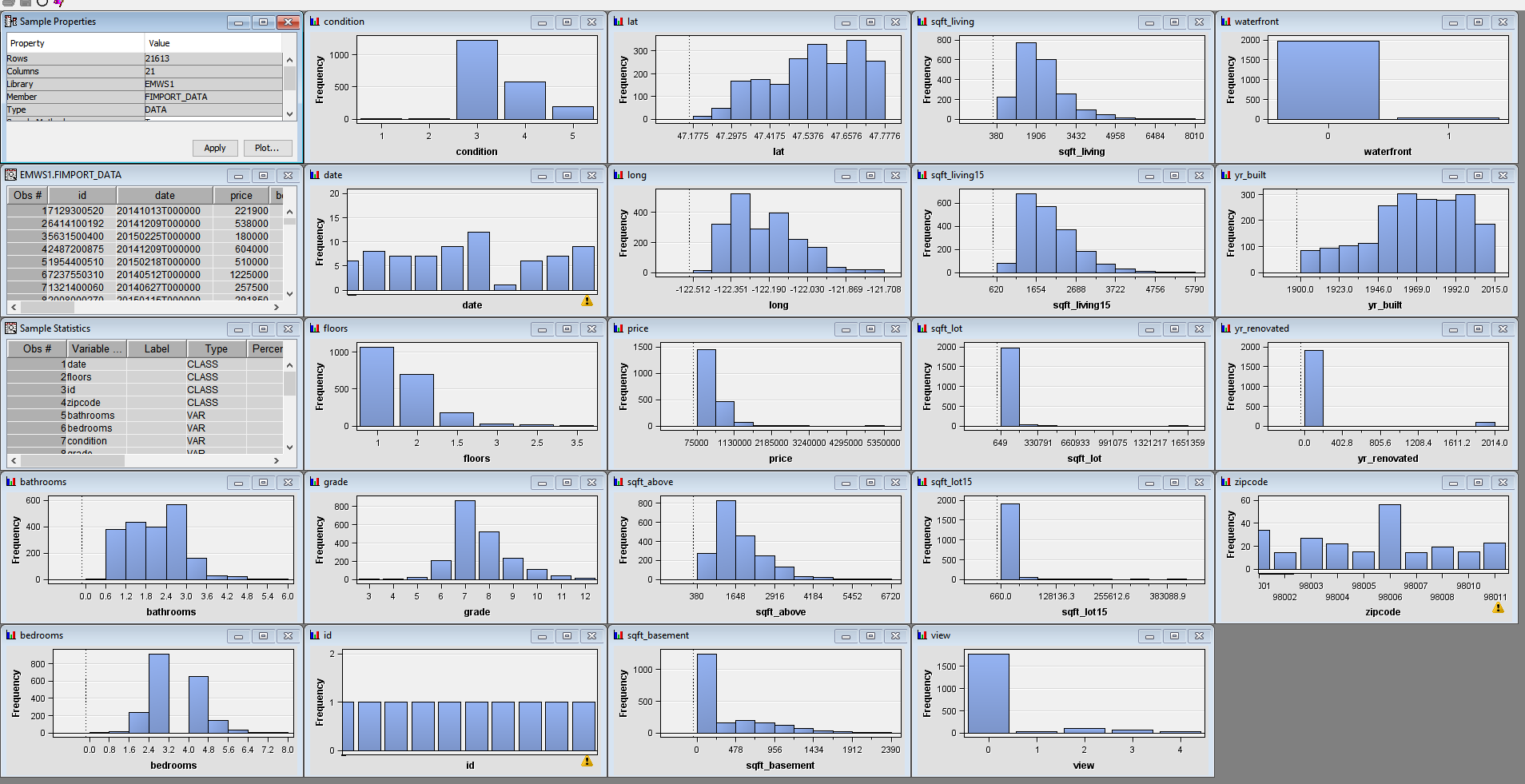


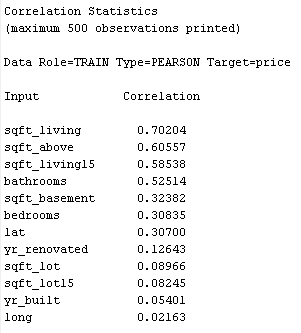
Here, we rejected 2 variables (date, ID), because I think there is no relationship between them and the target variable.

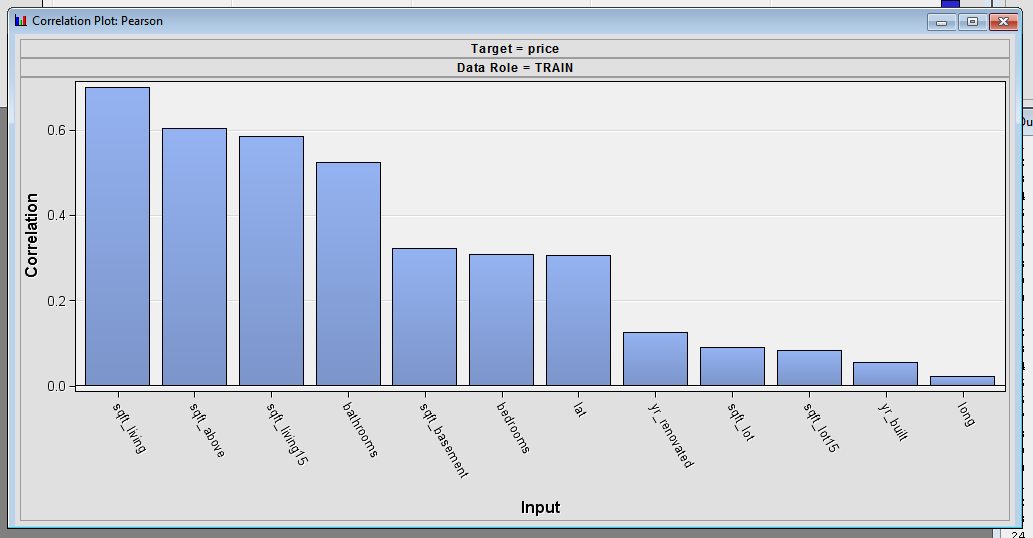
2. Exploring variable statistics

Using the StatExplore node, we can examine variable distribution and statistics. We can also compute correlation statistics for interval variables by interval input and target.





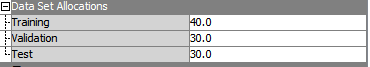




We can see that there is no missing value in all of the variables; and there are four variables (sqft\_lot, sqft\_lot15, yr\_renovated, price) with a skewness more than 2. We needn’t to use Impute Node to deal with missing values. We need to use transformation node to deal with the skewness. And we will use Filter node to deal with the outliers.

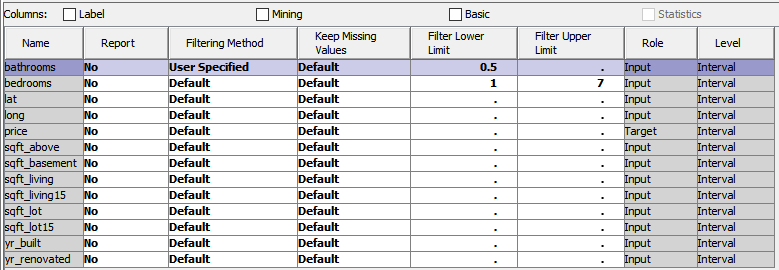
3. Data partitioning

We use Data Partition node to specify what percentage of the input data set is included in each partition.

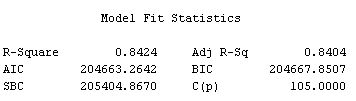


4. Dealing with outliers

After observing the dataset, we found there are some houses with 0 bedrooms or bathrooms or extremely high number of bedrooms. Because outliers can greatly affect modeling results and, subsequently, the accuracy and reliability of trained models, we use the Filter node to exclude outliers, by setting the upper/lower limit, 40 observations are excluded.

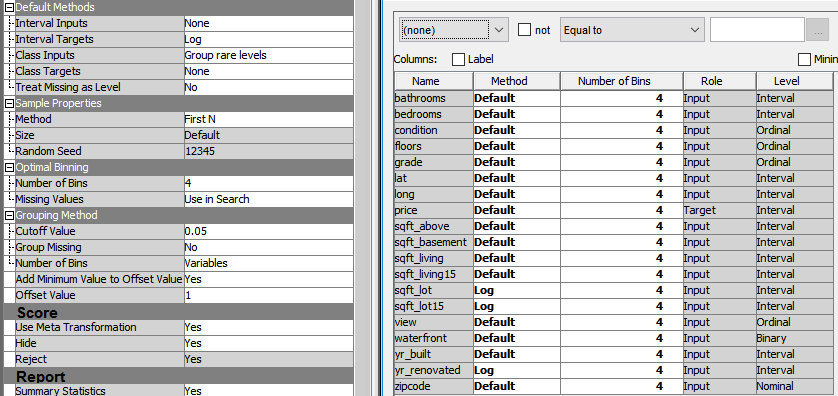


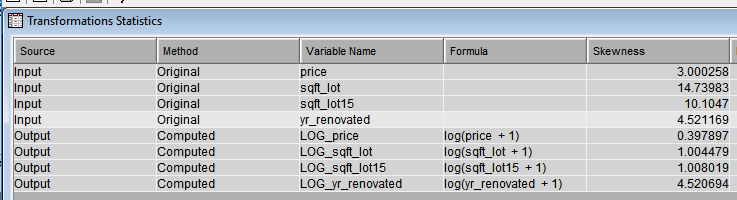
5. First Regression Node (without transformation for skewness)



6. Performing Variable Transformation

After we have viewed the sample statistics and variable distributions, it is obvious that some variables have highly skewed distributions. In highly skewed distributions, a small percentage of the data points can have a large amount of influence on the final model. Sometimes, performing a transformation on an input variable can yield a better fitting model. We use the Transform Variables node and apply a log transformation to the input and target variables with skewness more than 2. We can see the skewness of output variables are smaller than that of the original variables.





7. Fitting a Regression Model

Because the target is an interval variable, we use the multiple linear regression. There are four types of model (backward, forward, stepwise, none), and we tried all of them to see which one performed better. In the final step--Model Comparison, we know that the Regression (forward) and Regression (stepwise) both performed best among the all 9 models.

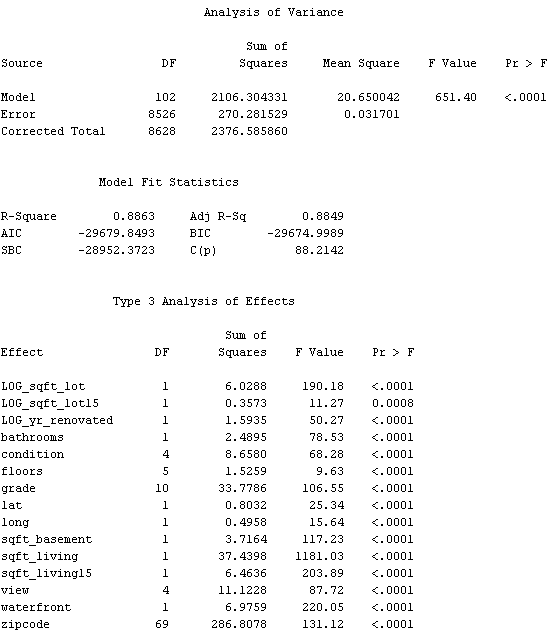
Backward — Variable selection begins with all candidate effects in the model and systematically removes effects that are not significantly associated with the target. This continues until no other effect in the model meets the Stay Significance Level, or until the Stop Variable Number is met. This method is not recommended for binary or ordinal targets when there are many candidate effects or when there are many levels for some classification input variables.

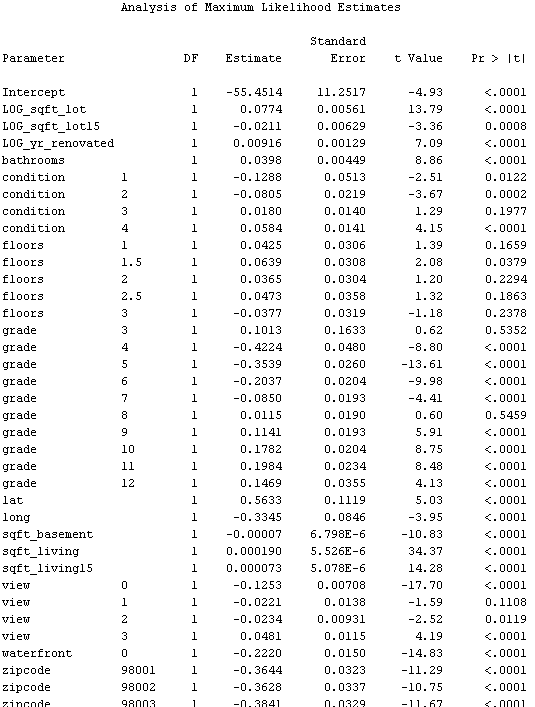
Forward — Variable selection begins with no candidate effects in the model and then systematically adds effects that are significantly associated with the target. This proceeds until none of the remaining effects meet the Entry Significance Level, or until the Stop Variable Number is met.

Stepwise — Variable selection begins with no candidate effects in the model and then systematically adds effects that are significantly associated with the target. After an effect is added to the model, it can be removed if it is deemed that the effect is no longer significantly associated with the target.

None — No model selection is performed and all candidate effects are included in the final model.

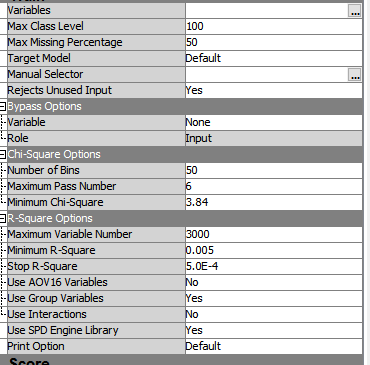
Here are some results of stepwise regression model.

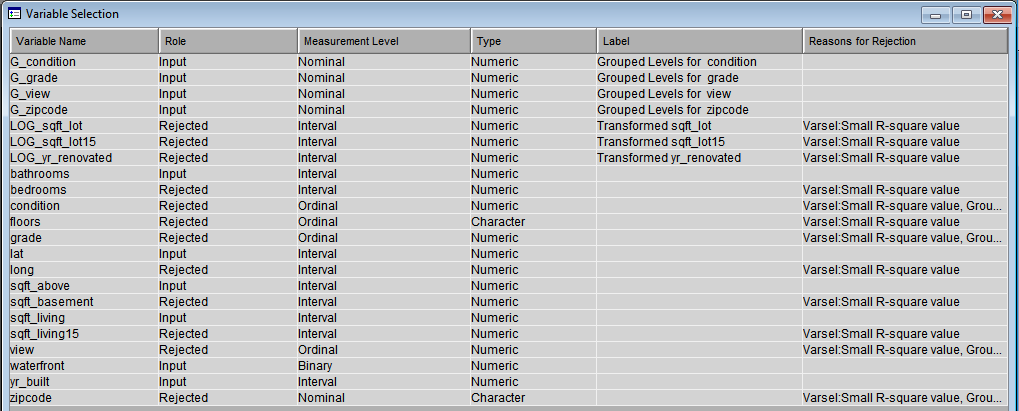




8. Selecting useful input variables

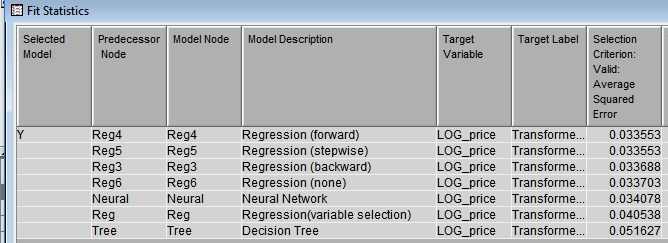
Using the Variable Selection node, we can quickly identify input variables that are useful for predicting the target variable. Here, a least squares regression was used to maximize the model R-square value. We can see 7 variables are rejected due to small R-square value, and 4 class variable are grouped.

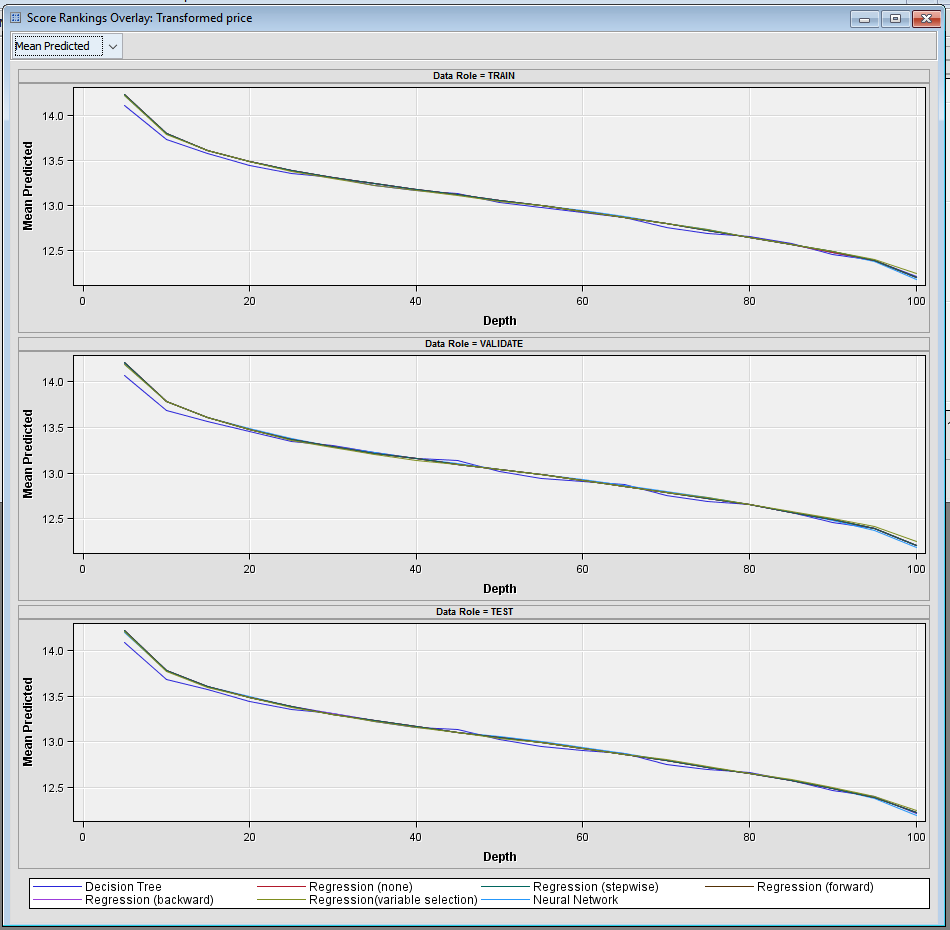


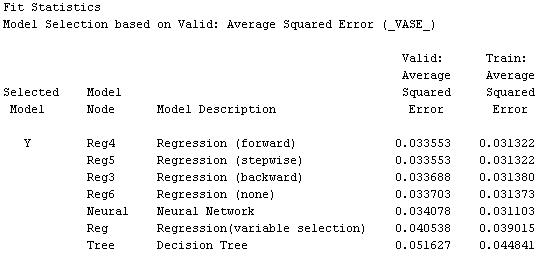


9. Evaluating the Models

From the results of Model Comparison node, we can see Regression (forward) and Regression (stepwise) both performed best in all seven models for they have the smallest Average Squared Error 0.033553.







**Findings and Managerial Conclusion**

***Best Model***

In the Model Comparison node, although SAS chose Forward Regression as the best model, we can see that the statistics on both Forward and Stepwise Regression are the same, which means they are doing equally well for our data set. With the regression equation, now we can predict more precisely the price of a house, given the location, size and features of the house; and we can predict the change of the price before and after modifications on a property. (\*Due to the a huge size of nominal data, the equation contains over a hundred variables.)

***Equation***

*Log(price)=-55.4514+0.0774\*(log(sqft\_lot))-0.0211\*(log(sqft\_lot15))+0.00916\*(log(yr-renovated))+0.0398\*bathrooms-0.1288\*condition(1)-0.0805\*condition(2)=0.0180\*condition(3)+0.0584\*condition(4)+0.0425\*floors(1)+0.0639\*floors(1.5)+0.0365\*floors(2)+0.0473\*floors(2.5)-0.0377\*floors(3)+0.1013\*grade(3)-0.4224\*grade(4)-0.3539\*grade(5)-0.2037\*grade(6)-0.0850\*grade(7)+0,0115\*grade(8)+0.1141\*grade(9)+0.1782\*grade(10)+0.1984\*grade(11)+0.1469\*grade(12)+0.5633\*lat-0.3345\*long-0.00007\*sqft\_basement+0.000190\*sqft\_living=0.000073\*sqft\_living15-0.1253\*view(0)-0.0221\*view(1)-0.0234\*view(2)+0.0481\*view(3)+0.2220\*waterfront-0.3644\*zipcode(98001)-0.3628\*zipcode(98002)+……*

*Where*

*condition(k)=1, if k=1, otherwise condition(k)=0.*

*floors (k)=1, if k=1, otherwise floors(k)=0.*

*grade(k)=1, if k=1, otherwise grade(k)=0.*

*view(k)=1, if k=1, otherwise view(k)=0.*

*Zipcode(k)=1, if k=1, otherwise waterfront(k)=0.*

***Application and Examples***

Given the input:

Sqft\_living=xx, sqft\_lot=xx,bedroom=xx,….

We can calculate log(price)=xxx, price=xxx.

**For example:**

1.If Sqft\_living increase 100 unit, keeping other input constant, how will the price change?

Estimates of coefficient of sqft\_living:0.000190

The new price will be exp(0.000190\*100)=1.019 times of the original one. We can say the price increases 1.9%.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2. If the number of bathrooms increase 1 unit, keeping other input constant, how will the price change?

Estimates of coefficient of bathrooms:0.0398

The new price will be exp(0.0398)=1.041 times of the original one. We can say the price increases 4.1%.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3. . If the condition changes from 1 to 2, keeping other input constant, how will the price change?

Estimates of coefficient of condition(1)=-0.1288, Estimates of coefficient of condition(2)=-0.0805

The new price will be exp(-0.0805-(-0.1288)))=1.049 times of the original one. We can say the price increases 4.9%.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4. If the yr\_renovated changes from 0 to 2015, keeping other input constant, how will the price change?

Estimates of coefficient of log\_yr\_renovated=0.00916

The new price will be exp(0.00916\*log(2015))=1.031 times of the original one. We can say the price increases 3.1%.

***Reference***

Harlfoxem. "House Sales in King County, USA”

Kaggle, https://www.kaggle.com/harlfoxem/housesalesprediction