**CIS 575 Final Project Report**

**December 1, 2020**

**PREDICTING THE PROPERTY PRICES IN KING COUNTY USING PREDICTIVE ANALYSIS AND SAS MINER**

By,

Ashish Arun

**Table of Contents**

Project Summary ……………………………………………………………………………………………………………………4

Background …………………………………………………………………………………………………………………………….4

Business Objective …………………………………………………………………………………………………………………..5

Dataset Description and Summary ..…...…………………………………………………………………………….…….5

Dataset Preprocessing …………………………………………………………………………………………………………….7

Description of Data Modeling and Assessment Techniques .…………………………………………………….7

Processing Map ……………………………………………………………………………………………………….………………8

Data Partition …………………………………………………………………………………………………………….……………10

Data Modelling ……………………………………………………………………………………………………………………….12

Decision tree …………………………………………………………………………………………………………………...12

Regression …………………………………………………………………………………………………………...………….13

Regression Model Result ………………………………………………………………………………………………….14

Standalone Neural Network ..…………………………………………………………………………………………..15

Auto Neural ..……………………………………………………………………………………………………………………16

Neural Network After Regression .…………………………………………………………………………………...17

Model Comparison Node ……………………………………………………………………………………………………….18

Conclusion ………………………………………………………………………………………………………………………………20

Appendix ………………………………………………………………………………………………………………………………..22

Appendix 1: Formulas Tab in transform variable ………………………………………………………………22

Appendix 2: Variable and method …….……………………………………………………………………………..23

Appendix 3: Correlation plot …………………………………………………………………………………………...23

Appendix 4: Data Partition Summary ……………………………………………………………………………….24

Appendix 5: Decision tree variable importance ……………………………………………………………….24

Appendix 6: Variables considered for the Regression model …………………………………………..25

Appendix 7: Final Weight of standalone Neural network ………………………………………………..25

Appendix 8: Fit statistics of standalone Neural Network …………………………………………………26

Appendix 9: Final Weight of AutoNeural network …………………………………………………………..26

Appendix 10: Final Weight of Neural network after regression node ……………………………...27

Appendix 11: Final project diagram ………………………………………………………………………………..27

Appendix 12: Decision Tree figure …………………………………………………………………………………...28

**Project Summary**

With a commendable growth in the earnings of people in the United States, There has been a significant rise in the middle class population who aspire to purchase a decent home in a decent locality. With people relocating very often, investing in real estate and even purchasing a house to retire, it is now more of a necessity to know and predict the price of houses depending on localities and amenities.

**Background**

In a real time scenario, picture yourself arriving in a new city and looking for a house to purchase. With no knowledge of the place and the range, you might be lost with a handful of choices which are overpriced and not enough value for your money . This vulnerability piqued our interest and we decided to focus our research on the area of house-hunting aka real-estate. To arrive on a subject matter to fully utilize our knowledge in data mining, we have taken into consideration King County in the state of Washington, for our analysis. With the advent of many companies providing the bread and butter to many people, King County is one of the choices for people to head towards and earn a livelihood. With a splurging population of 2,291,030 as reported by Census in 2020, King County has been our area of interest. It is also the 12th most populous county in the United States. As of a survey in July 2019, there were about 970,301 housing units and the owner occupancy rate stands at a staggering 57.1%. We will use predictive analysis to conclude on the price variation checklist of the available houses in King County based on the different attributes pertaining to a house.

**Business objective**

The ever changing house prices poses a problem, however there are attributes which help in determining the house prices. Knowing the insights of factors plays a major role in determining the house price and also predicting the future house prices of king county help both customers and realtors alike. An individual can estimate the price of a house in king county depending on attributes like construction date, Number of bedrooms, floor size and various other attributes. Realtors can also use the same model to predict the house prices and evaluate the investment they are making and predict the profit they may achieve over a period of time to eventually help them obtain profits.

**Dataset description and Summary**

The dataset is obtained from Kaggle.com and contains the data of the house sale prices in King county, which is located in Seattle, USA between May 2014 to May 2015. The various attributes and description of those attributes are given as follows-

Id - a unique set of characters and numbers used to identify the each property.

Price - The market value of the property mentioned.

Bedrooms - the number of bedrooms in that property.

Bathrooms - number of bathrooms in the property.

Sqft\_living - total land area of the living room measured in square feet.

Sqft\_lot - total land area of the parking lot measured in square feet.

Floors - Total number of floors in the property

Waterfront - Number of waterfront available in the property

View - rating (1 - 5) for the view from the property

Condition - rating (1 - 5) for the condition of the property where 1 represents that the property is in poor condition and 5 represents that it is in excellent condition.

Grade - represents the grade of the property based on king county’s grading system.

Sqft\_above - total land area of the property excluding the basement, measured in square feet.

Sqft\_basement - total land area of the basement of the property, measured in square feet.

Yr\_built - represents the year of construction of the property.

Yr\_ renovated - represents the year at which the property was renovated.

Zip code - contains the zip code of the locality where the property is located.

Lat - contains the latitude coordinates of the property.

Long - contains the longitude coordinates of the property.

Sqft\_living15 - total living area in 2015

Sqft\_lot15 - total area of the parking lot in 2015.

There are about 21613 observations in the dataset, with 2 binary variables, 5 nominal and 14 interval variables , the target variable is “Price” which is also an interval variable.

**Dataset Pre-Processing**

The raw data was collected from Kaggle.com. To further suit our goal, we imported the csv file for further scrutinization of the dataset to filter the required fields. The main business goal of this project is to predict the pricing of houses in King County to enable the consumer/buyer or a real estate agent to calculate the best suited budget to aid in choosing the house. The year of build is of utmost importance to assess the feasibility to buying a house in a specific budget curve. Hence, we transformed the raw data collected from Kaggle.com by adding a variable/field - Sold\_Year (date format) to the dataset which we have further designated as the new data to analyze and predict our model. The new field - Sold\_Year - is a nominal variable and it’s format was changed to YYYY. Prior to uploading the csv file to SAS Enterprise Miner, we made sure that the csv file was clean of missing values and was ready for further processing. The clean data was then ready for further analysis and the measures for the nominal and interval variables were fixed.

**Description of Data Modeling and Assessment Techniques:**

Any business decision or goal is achieved by having strategic models compete with each other to prove which one is a better model to get the desired result. So five different models were chosen to compare namely - decision tree, regression tree and standalone neural networks, Auto Neural and Neural network after regression node.

**Processing Map**

The filtered data was saved as King County node in SAS Enterprise Miner. To have an overview of the input variables and verifying any missing values, the data was processed using stat explorer utility. Any missing values in the dataset will need imputation to overcome the data irregularity problem. When we explored the data we found that we have no missing values in any of the columns and we need not have to impute any data and the need of imputation is not applicable for our dataset.

The original date field was not in a user readable format, so we created a new variable Sold\_Year which had a proper format and used it for our analysis. We then added two more columns called AGE and AGE\_RNV. We have used the formula tab in the transform variable node to implement these two columns, we found that it is usually beneficial to have the age of the house calculated for the model and the new columns will help in obtaining better performance from the models.(Refer Appendix No. 1)

For the calculation of the AGE of the house we have used the formula **(Sold\_Year) - (yr\_built),** which is of the type ‘numeric’. Here, the year at which the house was constructed is subtracted from the Date at which the house was sold to determine the age of the house, and for calculating the AGE\_RNV, we have used the formula **MIN((Sold\_Year - yr\_renovated), yr\_renovated)**, here we can see that the age of renovation is calculated by taking the minimum of the two values which are the difference between the year the house was sold and the year the house was renovated and the year that the house was renovated. This forms the new variable, ‘the age of the house since renovation’.

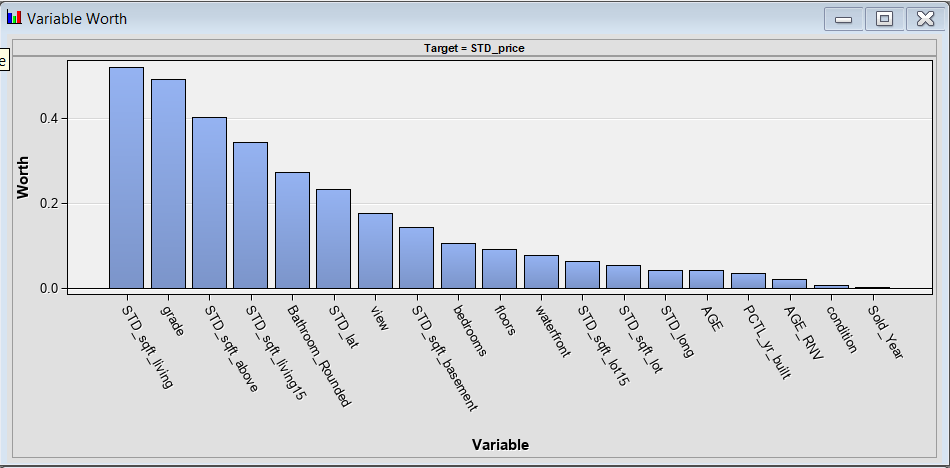
As we know, preprocessing a data will help us improve the accuracy of the model to make it more reliable, we have applied binning to reduce the effects of minor observational errors which might have crept in the dataset as we have introduced two new variables like AGE and AGE\_RNV into the dataset. This introduction is for better model understanding and reliability. We applied binning to the variables yr\_built and yr\_renovated by changing their method as optimal binning and specifying the number of bins as 4 for yr\_built and 2 for yr\_renovated. (Refer Appendix No. 1 )

We could also observe that the bathroom values are in fractions such as 1.5, 1.25 and 1.75, which we thought was impractical, Hence we rounded off the bathroom values using the formula ROUND(bathrooms) in the Transform variable 2 node. In addition, We have also standardized the inputs in the transform variable 3 node. .(Refer Appendix No. 2 )

Also, a point to note is that, as we had turned on the advanced advisor option while importing the dataset, the zipcode was rejected as it does not provide much value in the prediction of the target variable which is the price.

We then ran a stat explore node to the transform variable node 2 where we analysed the variables and the variable worth to the target variable to the price target. We also noticed the correlation of variables and their worth. We noticed that the ST\_sqft\_living, grade, STD\_sqft\_above, STD\_sqft\_living15 and Bathroom\_Rounded are the top 5 variables which will impact the target price according to the variable worth.

***Figure 1:*** *Variable worth*



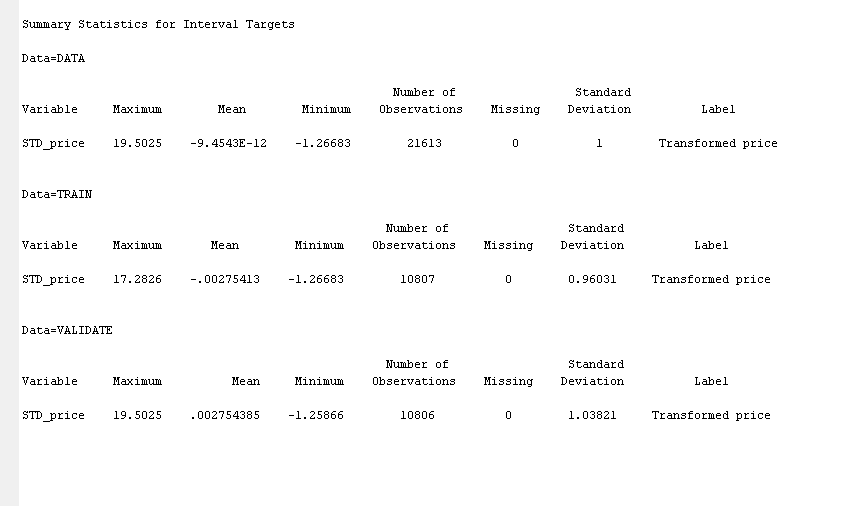
We also saw the similarities with the variable worth and the correlation worth where the top five variables are the same. (Refer. Appendix No. 3) From the correlation plot we come to know about the variable which will impact the target price.

**Data Partition:**

As we know, the dataset has to be divided to train and then the data is validated. Hence, we have split the entire dataset into two parts weighing the same percentage of division, that is, training dataset which is 50% and the validation data set, which is another 50%. There is no specification of any percentage for data, split for the test data set.

After running the node, it can be seen from the results that the total data is split into two parts, TRAIN and VALIDATE. The train part of the dataset has 10807 observations with 0 missing values and the validate part of the dataset, which has 10806 observations with 0 missing values. (Refer Appendix No. 4 )

***Figure 2:*** *Data Partition Summary statistics*



We have successfully partitioned the data and have also pre processed the dataset. Our aim is to apply various data models to compare and verify which model derives the best results for our dataset. We started our analysis by applying the decision tree to create a classification model for our dataset.

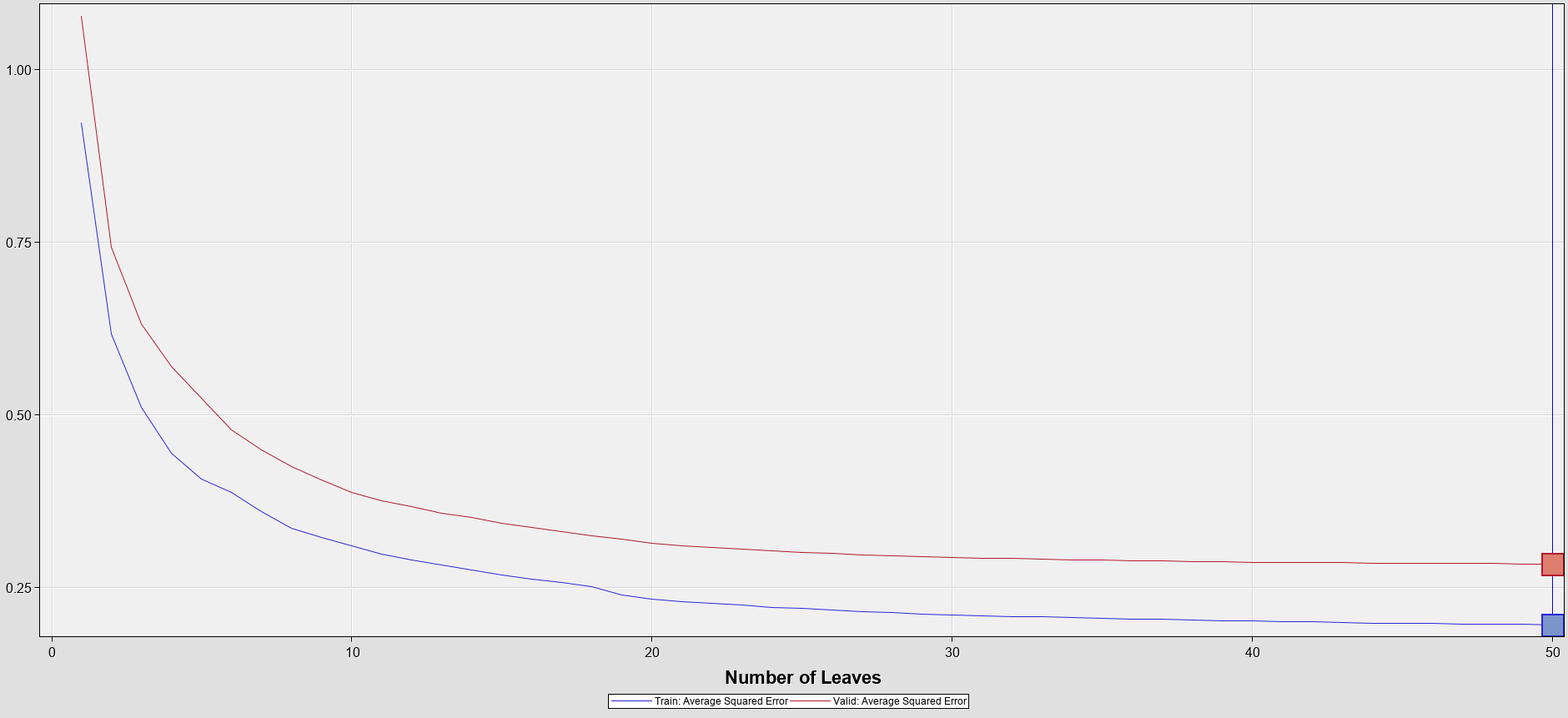
**Data modelling**

**Decision Tree**

Firstly, We have applied the decision tree, we know that it is not much suitable for the type of target variable we are trying to predict, however, we are planning to implement and see the prediction which can be used as a baseline for evaluating other models. We have set the Assessment measure to Average Square error and the maximum leaf size is set to 3.

We can see the variable importance of the decision tree from which we can determine which variables affect the price target more than others. (Refer appendix No. 5)We can see the important variables considered to be the top variables worth are the variables - grade, STD\_sqft\_living, STD\_lat, STD\_long, waterfront and age which are almost inline with the top variables we obtained during the variable worth analysis.We can note here that the added column ‘age’ is having more variable worth than the Sold\_Year of the houses. We obtained an Average squared error of 0.2846 for the validation data and a 0.1966 for the training data.

***Figure 3:*** *Average Square Error plot of Decision Tree*



From the screenshot attached above, it is evident that the decision tree is used on interval targets. Usually, decision trees are used in places where a given dataset is analyzed to arrive at the best possible decision for a successful outcome is required. Since our goal is not related to the defined goal of the decision tree model, it maynot be the best approach but it can still be used as a standard to compare other models.

**Regression**

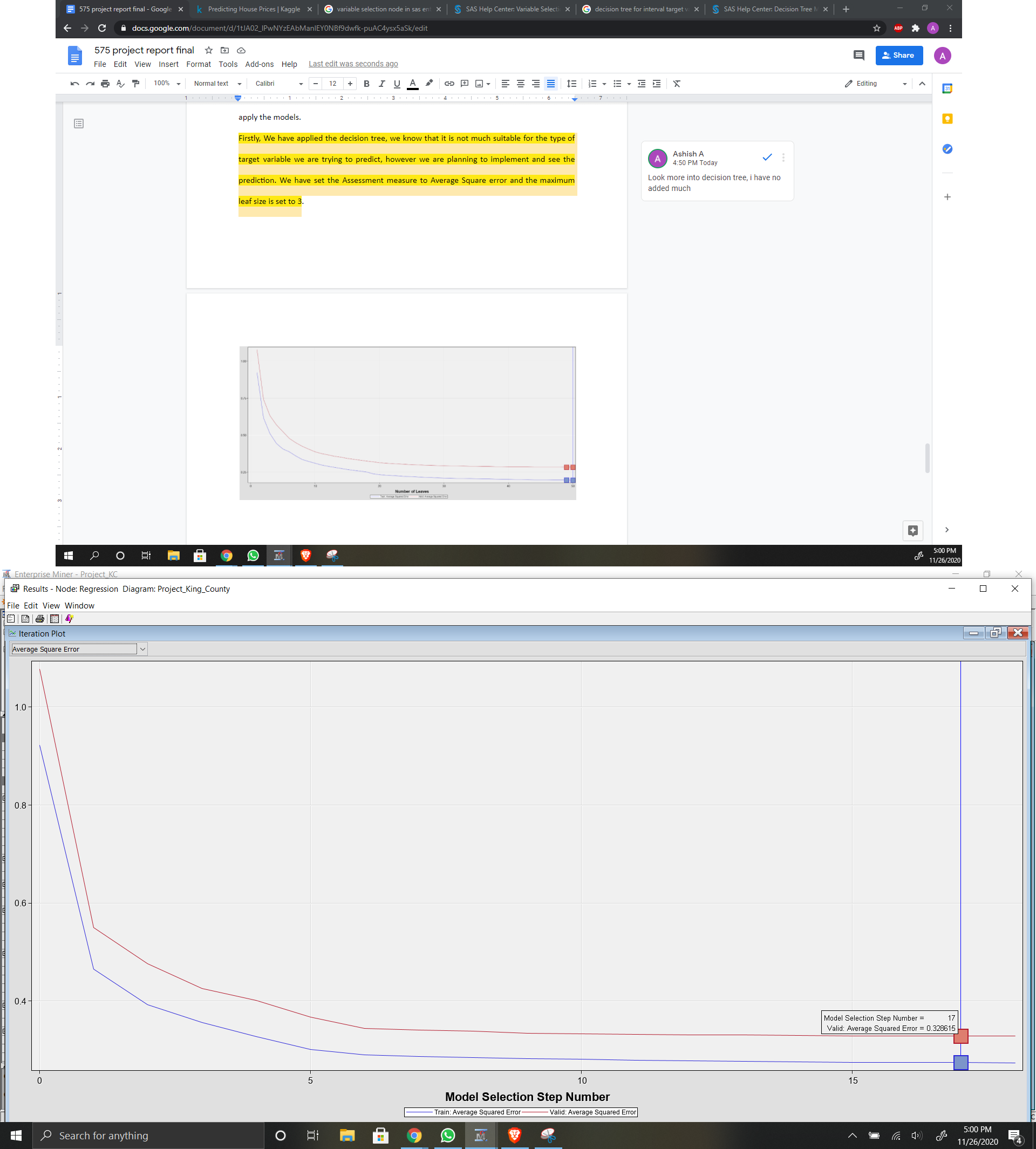
Regression model is the next model we will be utilizing after the decision tree model. In our opinion, we consider this model to be more appropriate for the use case in our current dataset as, by definition, the regression model is employed to predict the relationship between dependent and independent variables. Hence, this model would be able to predict the prices of houses based on the dataset provided to a fairly accurate degree.

In SAS Enterprise miner, we added a regression node to the diagram and connected it to the data partition node. We have used the selection criterion as “Validation Error” and the model we have used is Stepwise, which was the best selection model compared to backward and forward. We have set the selection defaults to “yes” for the regression model.

**Regression model result:**

As we can see from the Average squared error plot of the regression model, we have got the average squared error for training data as 0.2747, whereas for the validation data it is 0.3286. We can notice that as the plot is going through each step, the ASE is decreasing and at the 17th step we get the most optimum model for the regression.

***Figure 4:*** *Average Square error plot of Regression model*



When we dive deeper into understanding the variables that have affected the most in the regression model, we can see that the longitude of the place and the square feet of the house are the most impactful as we can see the F value of those being 1530.55 and 1449.56 respectively. This also makes sense as we know that the living area and the neighbourhood of the house, will have a greater impact on the price of a house compared to any other factor. We can also see that the grade of the house is also a deciding factor as it makes a significant impact with a F value of 1202.07. The other variables which make a comparatively better impact are ‘view’ and ‘waterfront’ as they have F values over 300. (Refer Appendix No. 6)

We can notice that the AGE variable that we calculated is impacting better than the yr-built with a F value of 98.76 over 44.48. This model has helped us to identify the factors that have the maximum impact on determining the price of a house in King county.

However, comparing both the models (decision tree & regression), We can conclude that the regression model has not performed better than the decision tree model, but it helped us to identify a few of the contributing factors that might impact the price of the house/property.

**Standalone Neural network**

After analyzing the results of the regression model, we proceed to implement the Neural network model. A neural network will consider all the input variables to aid us in obtaining the results. We then proceeded to implement a standalone neural network node to the data partition node and in order to have the same criteria across all the models, we proceeded to choose the selection criterion as “Average error”. The result we obtained for the Average Square Error of the training data is 0.1432 and for validation data is 0.1729 at an estimated weight of 85. When we compared this result to the results of the other two models, it is evident that the result from this model is significantly better than that of decision tree and regression models. Moreover, we know that this standalone Neural Network has considered all the input variables and the results can be further improved by selecting the input variables rather than considering all the variables. We have also set the maximum number of iterations to 50 and we can see that there are a total of only 46 iterations in the model.

When we look into the final weights of this model, we can observe that the strongest weight is for the variable condition in the H12 hidden unit. We can assume that, condition is introducing most of the positive bias for the model in predicting the price and latitude is bringing the negative bias to the model with a value of -1.32.(Refer appendix No. 7 )

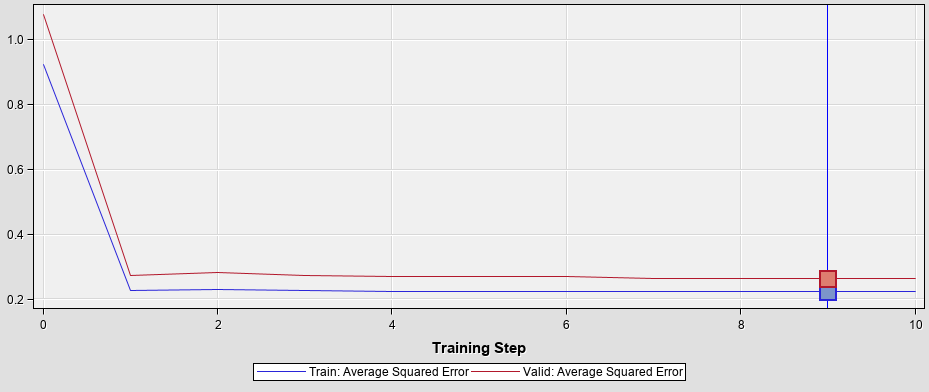
***Figure 5:*** *Average Square error plot of Standalone Neural Network model*



**AutoNeural**

After the implementation of a standalone neural network, we implemented the AutoNeural network where we have selected the criteria as “Train”, Action as “Search”, Tolerance to “low”, Direct to “No” and Normal to “No”. We have also set the number of hidden units to 2 and the total number of hidden layers is limited to 5. As we can see from the results obtained, the average square error for the training data is 0.2219 and ASE for validation data is 0.2626 at the training step of 9. Also, from the weights we can see that the highest weight and impacting variables are inline with the other models like grade, sqft\_living are the variables which are impacting the decision of the autoneural network (Refer Appendix No. 9)

***Figure 6:*** *Average Square error plot of AutoNeural model*

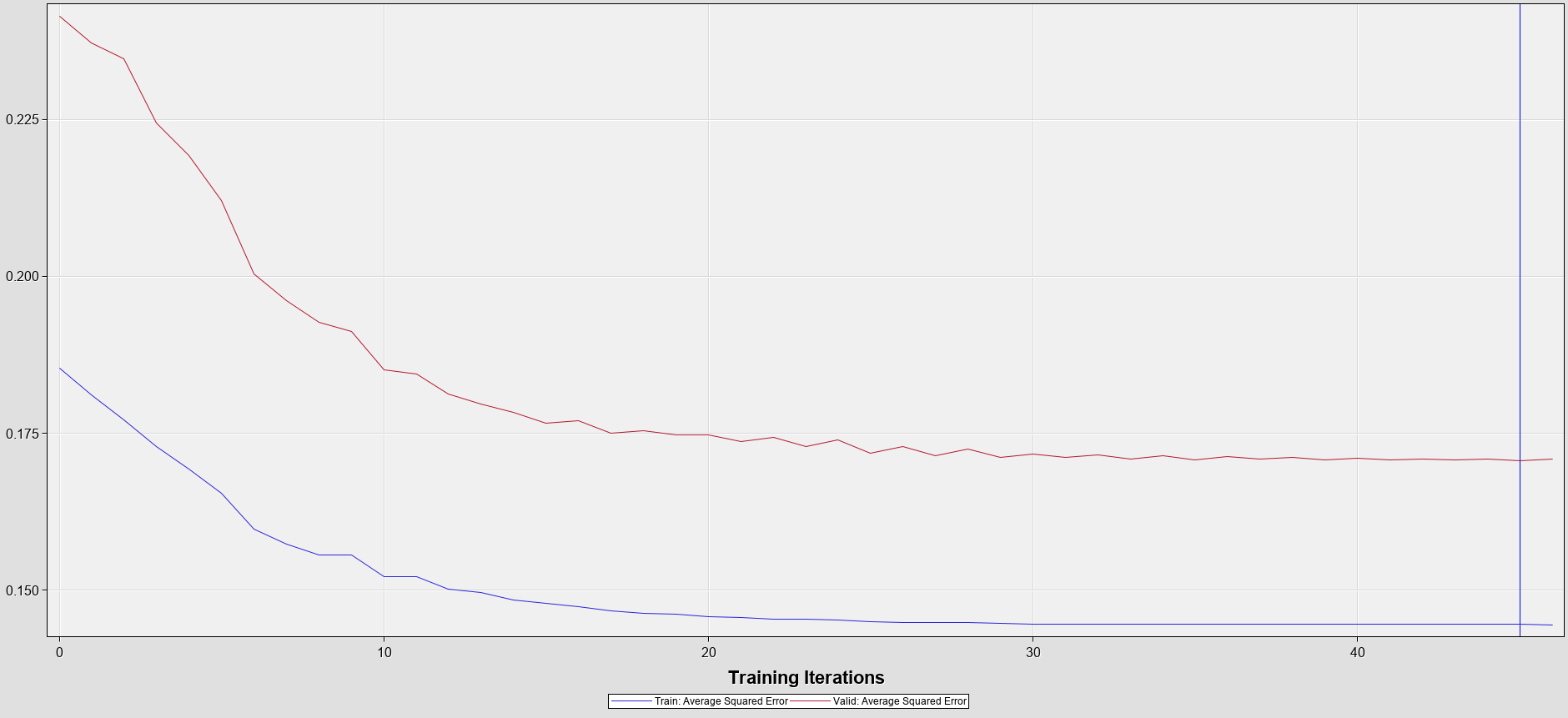


**Neural network after Regression**

As we learned in class, we applied the neural network to the regression node to minimize the input selection and let the neural network select those variables which are significant. We also changed the selection criterion to Average error, This enables us to have limited inputs to the neural network thereby helping us run the neural network with the inputs selected by the regression model.

When we ran the model, we obtained the model result for ASE of the training data = 0.1444 and ASE of the validation data = 0.1706. We can notice from the final weight model that the latitude is contributing to the model with a weight of 1.1383 at the hidden unit H11 aligning with all the other neural networks. We can visualize that this is the best performing model we have obtained.(Refer Appendix No. 10)

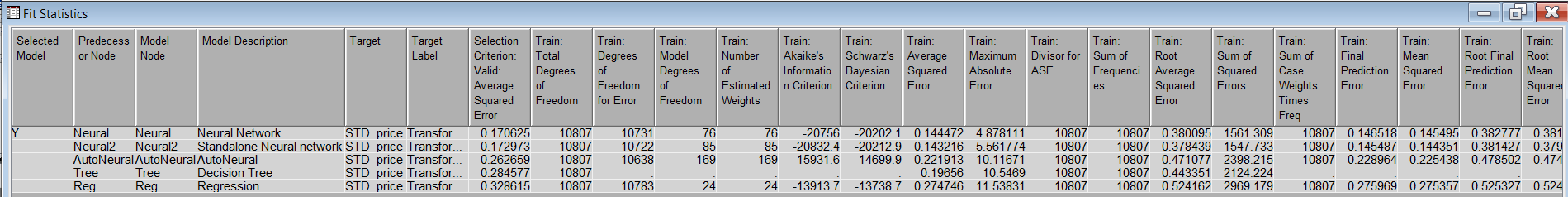
***Figure 7:*** *Average Square error plot of Neural Network after regression model*



**Model comparison node**

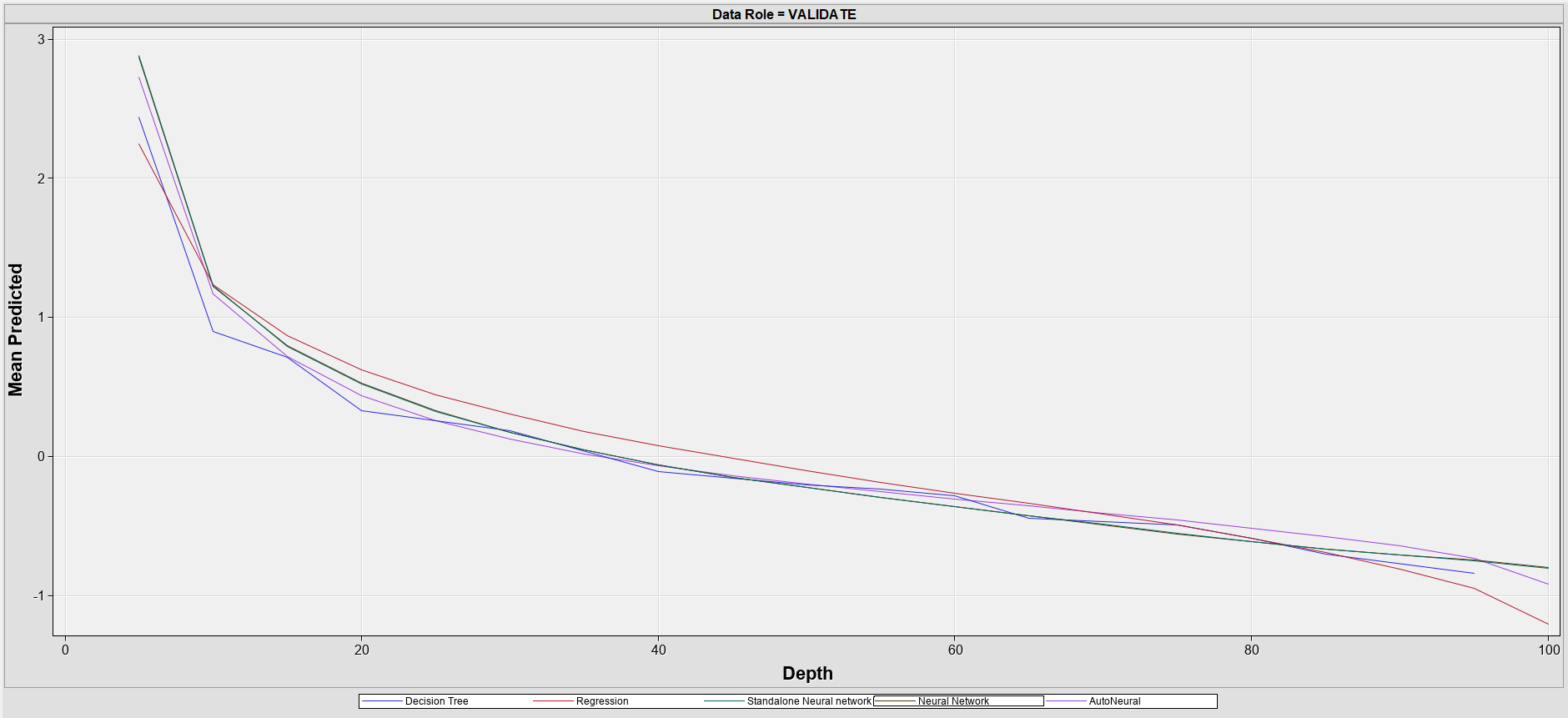
We then proceeded to compare all the models we have executed with the model comparison node. When we look at the fit statistics, it is clear that the best performing model is the neural network connected to the regression node with a validation ASE of 0.170625, We can also notice that the second best model is the standalone Neural network with an ASE of 0.172973. The regression model is the least performing model among the five models we have compared.

***Figure 8:*** *Fit Statistics of the model comparison model*



When we further dive into looking at the Score Ranking overlay of the validation data of the price we can notice that at depth of 5 the highest mean predicted os of the neural network after regression node at 2.87 and at depth 100 the same model proves to be better, however we can see in the middle depths the regression node is also performing better than other model.

***Figure 9:*** *Mean predicted plot of the validation data*



**Conclusion**

This analysis brings a lot of insight on the house sales data set of King County that was gathered between May 2014 and May 2015. The goal of this analysis was to develop a predictive analysis to determine the price of houses in the said county based on the provided data set. It takes into consideration, various attributes such as the year of construction of the house, square foot area, number of bedrooms, the neighborhood at which the house is located etc to help us determine the price of a house which has any such common factors. This could help any potential buyer/customer to determine the price of a house without any prior knowledge in the field of real estate thereby saving them from overpaying. It could also help real estate agents to accurately determine the cost of a property thereby helping them to secure good profits on their sales. In its absence, any person who is new to King County without any prior knowledge in real estate intending to buy a house in the county would have no clue where to start and would not know how to determine the budget for buying a house.

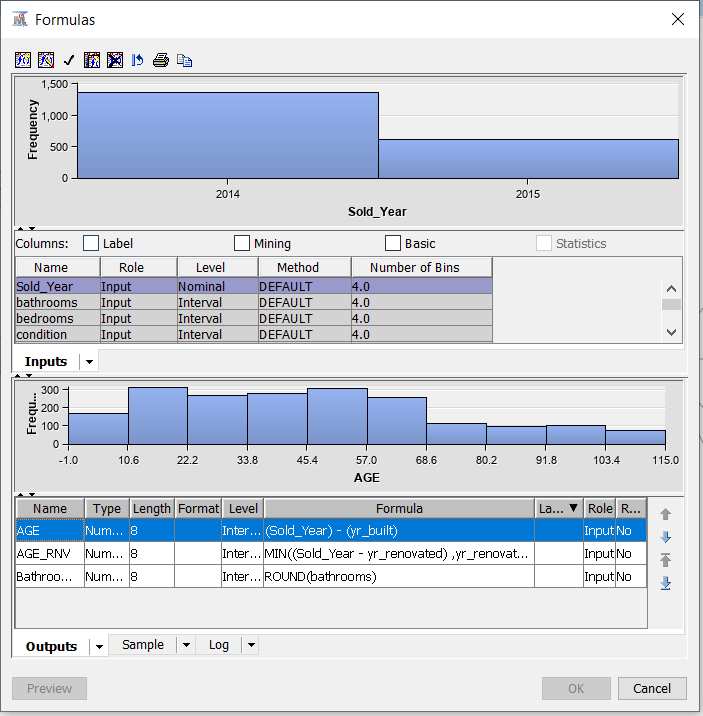
There were different models employed for analysis like decision tree model, regression model, neural networks etc. Although, each of these models are unique in their own way,

the one which is highly successful is the neural network connected to the regression node. This is because it has the least average squared error value when compared to other models in the analysis. One of the most consistent variables throughout the modelling process is the STD\_lat. This variable throws light on the most important factor while buying a house: location of the house. A property is worth investing only if the location can attract possible price growth in the future. The other variables playing a crucial part in the analysis is the STD\_sqft-living which gives us a fair idea that the square foot area of the house is also a deciding factor when buying a house. A house having a fair share of built area has a positive impact on a potential customer to make a deal. We can verify this finding by leafing through few of the online housing websites pertaining to the real estate domain. Houses with a profitable and attractive location, paired with an appreciative built area has more buyers in comparison to poorly located houses.

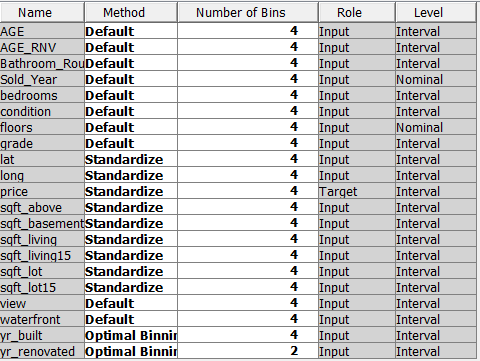
As a predictive model, it is very essential for the model to predict the deciding variables or factors to cater to a buyer’s or customer’s needs. This type of model not only helps a buyer, but also the construction companies to focus on the criterias for building a house to boost their sales. With a right predictive model in-hand, buyers as well as construction companies can have a win-win situation.

**Appendix**

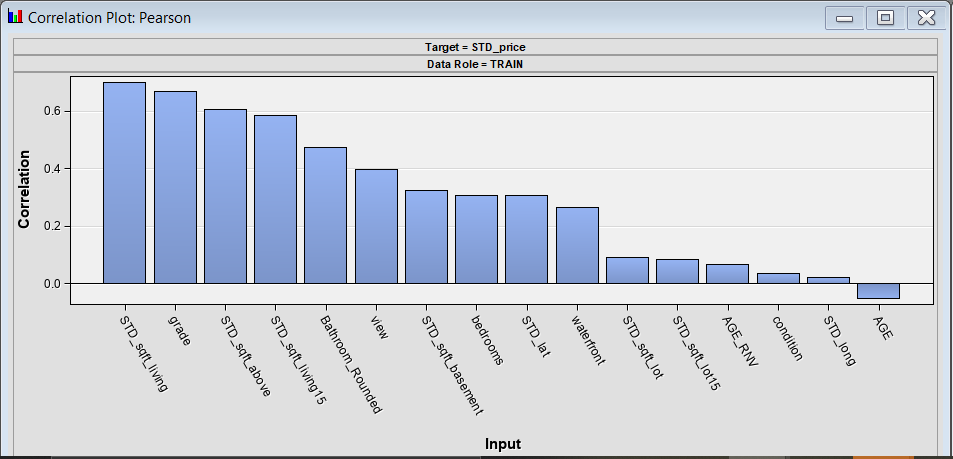
Appendix 1: Formulas Tab in transform variable



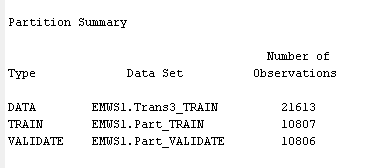
Appendix 2: Variable and method



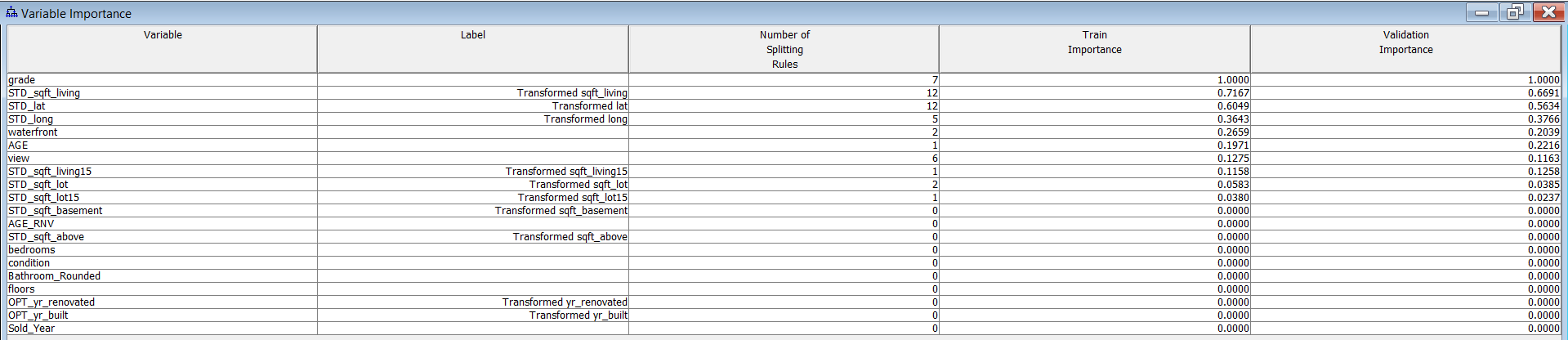
Appendix 3: Correlation plot



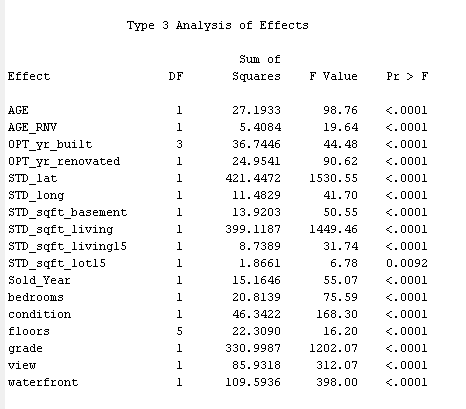
Appendix 4: Data Partition Summary



Appendix 5: Decision tree variable importance



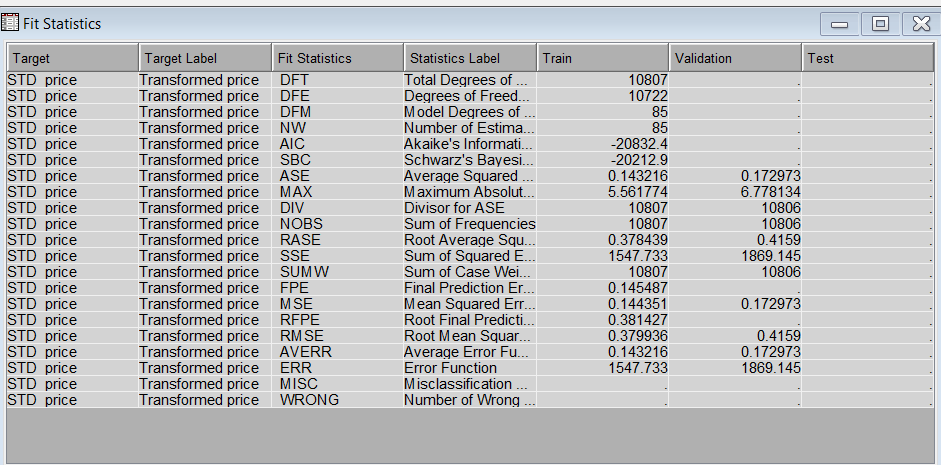
Appendix 6: Variables considered for the Regression model



Appendix 7: Final Weight of standalone Neural network



Appendix 8: Fit statistics of standalone Neural Network



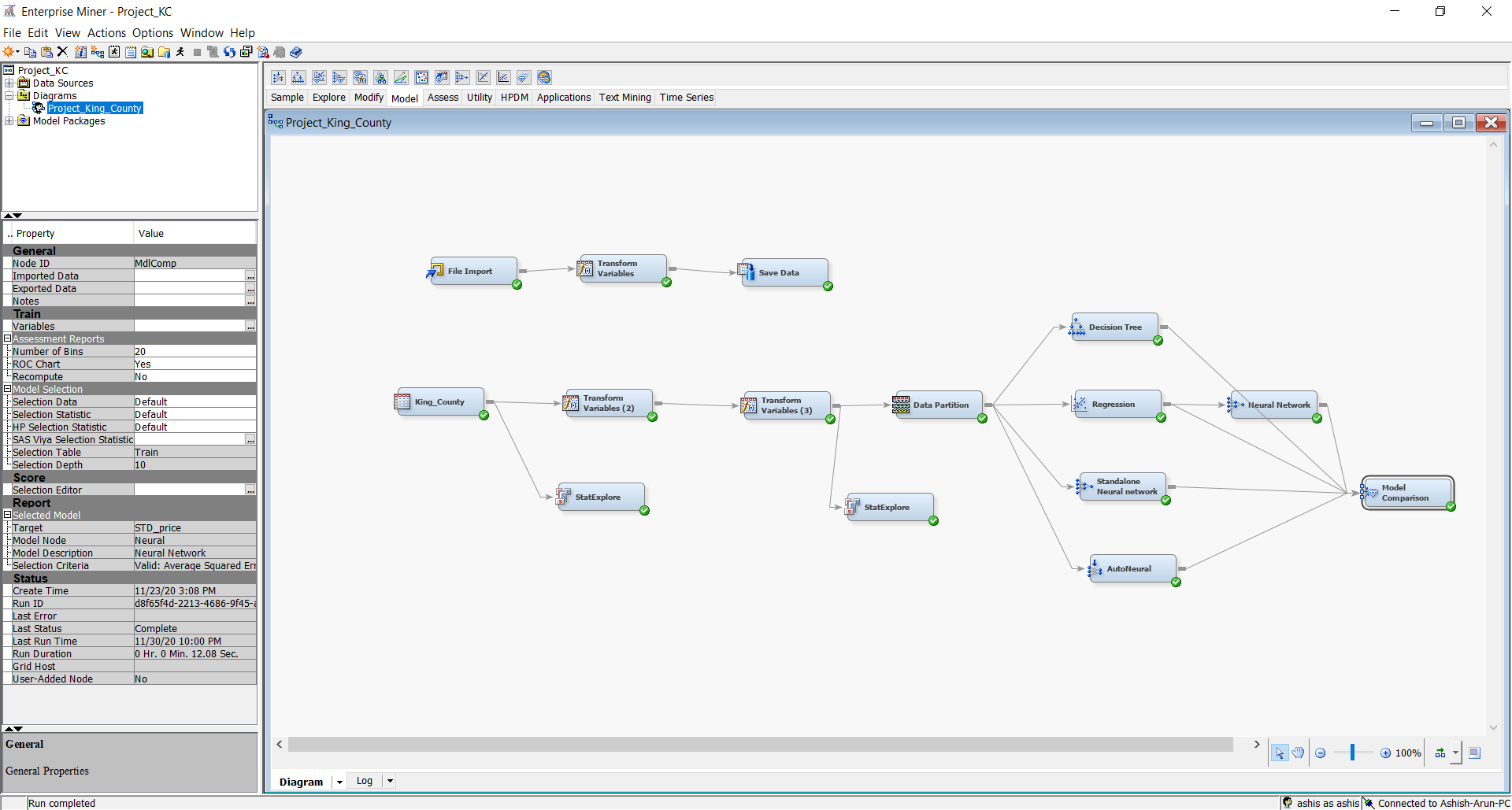
Appendix 9: Final Weight of AutoNeural network



Appendix 10: Final Weight of Neural network after regression node



Appendix 11: Final project diagram



Appendix 12: Decision Tree figure

