**Team Members:**

Aparna Unni

Caitlyn May

Kathryn Walker

Partha Ganrai

Pooja Singh

**House Price Prediction**

OPIM 5604, Fall 2016

Contents

[Executive Summary 2](#_Toc468211484)

[Strategy 3](#_Toc468211485)

[Data Dictionary 3](#_Toc468211486)

[Sampling 4](#_Toc468211487)

[Exploration & Modification 4](#_Toc468211488)

[Column Removal 5](#_Toc468211489)

[Correlation 6](#_Toc468211490)

[Data Modification 9](#_Toc468211491)

[Binning SalePrice 10](#_Toc468211492)

[Data Visualization 12](#_Toc468211493)

[Missing Value Treatment 14](#_Toc468211494)

[Outlier Analysis 16](#_Toc468211495)

[Scaled Variable 19](#_Toc468211496)

[Principle Component Analysis 20](#_Toc468211497)

[Modeling 21](#_Toc468211498)

[Cluster Analysis 21](#_Toc468211499)

[Stratified Sampling 23](#_Toc468211500)

[Predicting two category SalePrice 24](#_Toc468211501)

[Predicting three category SalePrice 29](#_Toc468211502)

[Predicting Continuous SalePrice 32](#_Toc468211503)

[Limitations 35](#_Toc468211504)

[Discussion 35](#_Toc468211505)

[Conclusion & Recommended Next Steps 36](#_Toc468211506)

[Appendix 37](#_Toc468211507)

# Executive Summary

To determine how house features add up to its price tag, the team performed data processing and applied predictive modeling techniques on a specific data set. Data was retrieved from Kaggle and it had 79 explanatory variables describing almost every aspect of residential houses in Ames, Iowa. The team thought the business value of predicting sales price with given house information would be valuable for real estate agents, house buyers, mortgage companies and possibly more business consumers as needed.

To resolve inconsistencies and eliminate redundant information, the team pre-processed the data and then used various modeling techniques to find the most accurate predictive model for predicting house sale price. It should be noted that the sample only had 1460 observations, therefore before any decisions are made to use these models at a larger scale, the team recommends more observations should be added to the sample to possibly create a more accurate model.

**Results**:

* ‘*SalePrice*’ has been established as the target variable. Besides the continuous target variable, the team has binned it to two (High, Low) and three (High, Medium, Low) categories.
* 31 variables were removed from the data set as it contained information that could be derived or were insignificant.
* Univariate, Bivariate and Multivariate outlier analysis was performed and the team decided to retain the outliers as the business value was uncertain.
* Ensemble Linear Regression determined the ‘*SalePrice*’ with an R-Squared value of 0.861.
* Ensemble Logistic Regression model determined if the ‘*SalePrice*’ is High or Low with an accuracy of 89.55%.
* Ensemble Decision Tree model determined if the ‘*SalePrice*’ is High, Medium or Low with an accuracy of 83.33%.

# Strategy

After extensive analysis on the data and data dictionary, the team decided that ‘*SalePrice*’ can be predicted as a continuous target variable or can be classified into two or three categories. ‘*SalePrice*’ has been categorized as High, Low and High, Medium, Low. The team decided to use the target variable in multiple ways to test different types of models and observe how the different target variables impacted the results of the models. Target variable is the one which receives properties of other variables. Hence, the price of the house will be predicted based on other variables.

The team followed the SEMMA approach to pre-process and create models. First, the team iteratively explored and modified the data as needed to remove inconsistencies. Next, separate models were created for both continuous and categorized target variable. Finally, as the cost of error is uncertain the models were assessed based on model accuracy, sensitivity, specificity and R-Squared value.

# 

# 

# Data Dictionary

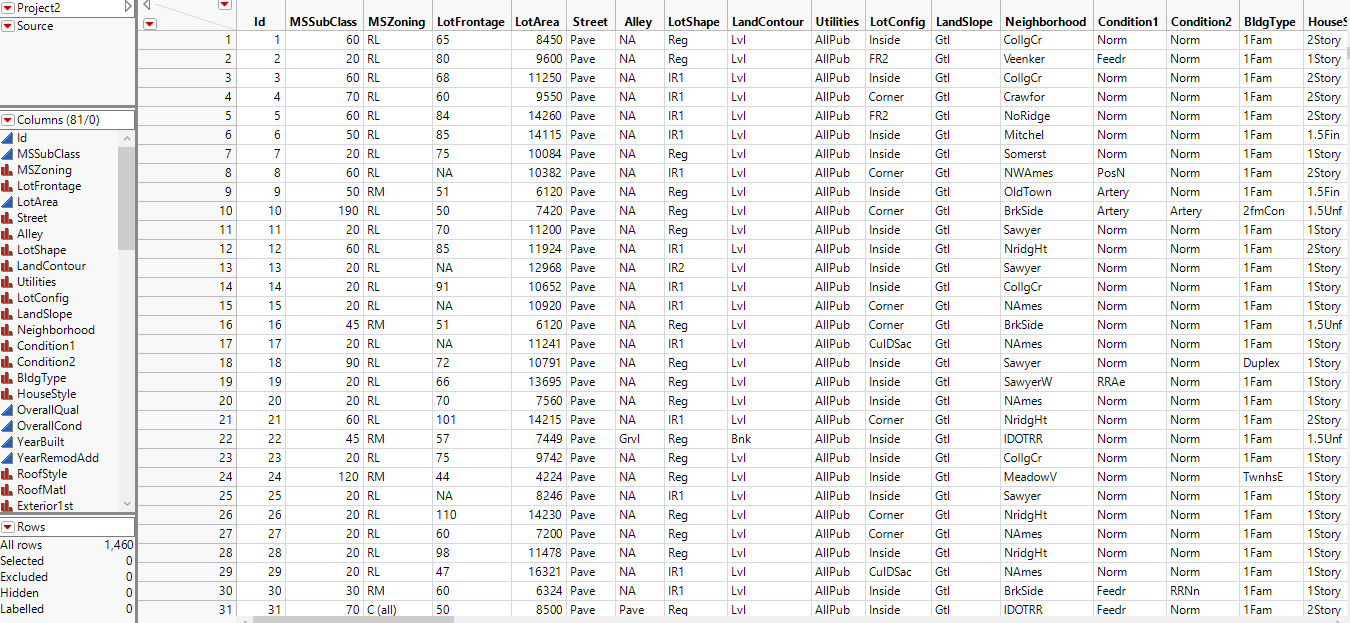
The user is recommended to read the updated data dictionary for clean direction of what each variable is before taking further action. The dictionary also includes descriptions of variable categories. A part of the data dictionary is in Table 1 and the complete dictionary can be found in [Appendix A](#_Appendix).

|  |  |  |
| --- | --- | --- |
| **Title** | **Description** | **Variables Used (Description)** |
| BldgType | Type of dwelling | 1Fam (Single-family detached) 2FmCon (Two-family conversion; originally built as one-family dwelling) Duplx (Duplex) TwnhsE (Townhouse end unit) Twnhsl(Townhouse inside unit) |
| GarageFinish | Interior finish of the garage | Fin (Finished) RFn (Rough Finished) Unf (Unfinished) NA (No Garage) |
| PavedDrive | Paved driveway | Y (Paved)  P (Partical Pavement)  N (Dirt/Gravel) |
| ExterCond | Evaluates the present condition of the material on the exterior | Ex (Excellent) Gd (Good) TA (Typical/Average) Fa (Fair) Po (Poor) |

Table 1: Data Dictionary

# Sampling

Exhibit 1 shows a screenshot of the data set. Since there were only 1460 observations in the data set, the team did not opt for random sampling. The team would focus on this sample for the exploration and modification phase.

Exhibit 1: Original Data Sample

# 

# Exploration & Modification

The exploration and modification part of this report focuses on cleaning the data of inconsistencies, as well as ensuring the data set has valuable information. The team reviewed each column in detail to determine if there was enough variance in the column to have value, if the column was defined correctly, as well as ensuring the data in the column was clear. Modification steps were taken when necessary. To focus on creating a concise data set, the team ensured that each column that would be used for modeling had abundant data and information that could not be found in another column. This approach would lead the team to success without data overload, or curse of dimensionality.

## **Column Removal**

# 

During exploration, the team reviewed each column to ensure no redundant information and enough data variance is present for proper data modeling. There were 81 explanatory variables in the original data set and to reduce the model complexity, the team removed 31 variables that can be either derived from other variables or are correlated or have very low variance. Table 2 below explains the columns we removed and the reason for their removal.

# 

|  |  |
| --- | --- |
| **Column** | **Removal Reason** |
| Street | Limited variance as only 0.4% streets were different from ‘Paved’. |
| Alley | Limited variance as only 6.16% values were different from ‘No alley’. |
| Utilities | Limited variance as only 1 utility was different from ‘AllPub’. |
| LandSlope | Limited variance as only 5.34% values were different from ‘Gtl’. |
| Condition2 | Limited variance as only 1.02% conditions were different from ‘Norm’. |
| RoofMat | Limited variance as only 1.78% roof materials were different ‘CompShg’. |
| Exterior2 | Exterior1 and Exterior2 have the same value in more than 90% cases. |
| HeatingType | Limited variance as only 2.19% heating types were different from ‘GasA’. |
| BsmtFinType2 | Limited variance as only 10% values were different from ‘Unfinished’. |
| BsmtFinType2SF | Limited variance as the area of only 10% finished basements was different from ‘0’. |
| TotalBsmntSF | Correlation of 0.93 with the sum of *FinBsmntSF* and *UnfBsmntSF*. |
| Central Air | Limited variance as only 6.5% values were different from ‘Y’. |
| LowQualityFinSF | Limited variance as 98% of the values were ‘0’. |
| Functionality | Limited variance and confusing description in data dictionary. |
| Kitchen Quality | Limited variance as more than 95% of the values fell in the same category. |
| 3SsnPorch | Limited variance as more than 98% of the values fell in the same category. |
| ScreenPorch | Limited variance as more than 90% of the values fell in the same category. |
| PoolArea | Limited variance as only 7(0.5%) houses have pools. |
| PoolQC | Limited variance as only 7(0.5%) houses have pools. |
| MiscFeature | Limited variances as only 3.6% houses have miscellaneous features. |
| MicsValue | Limited variances as only 3.6% houses have miscellaneous features. |
| GarageYrBuilt | Correlation of 0.83 between *YearBuilt* and *GarageYrBuilt*. |
| GarageCars | Correlation of 0.81 between *GarageArea* and *GarageCars*. |
| BasementCond | All categories fall under any one category of *BasementFinType1.* |
| SaleType | All categories fall under any one category of *SaleCond.* |

Table 2: Reasons for removal of columns ([Appendix B](#_Appendix_B:_Exploration))

## **Correlation**

A few variables were found to be correlated, therefore the team saw little value in keeping both columns. The team performed bivariate analysis between *YearBuilt* and *GarageYrBuilt* which informed the team that there is a correlation between these two variables (see Exhibit 2). To remove redundant information from the data set, the team removed *GarageYrBuilt*. Similarly, GarageCars has been removed as there is a correlation of 0.88 between *GarageCars* and *GarageArea* (see Exhibit 3).

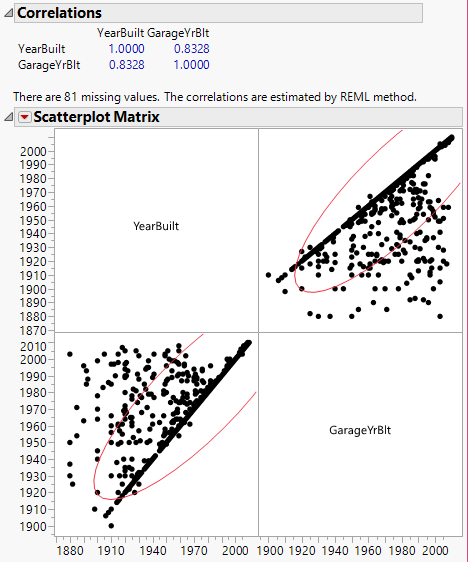
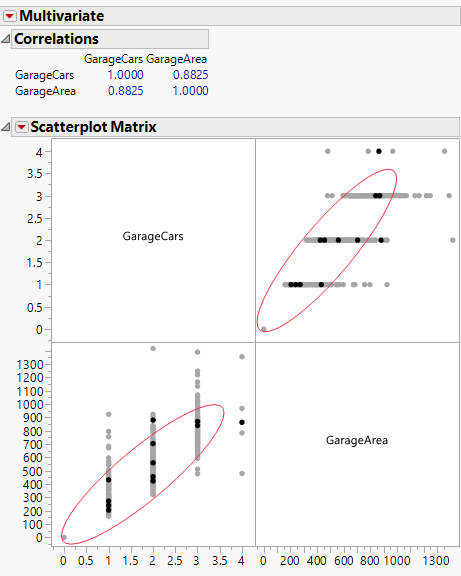
 

Exhibit 2: Correlation YearBuilt vs. GarageYrBuilt Exhibit 3: Correlation GarageCars vs. GarageArea

As *TotalBsmntSF* is highly correlated to *1stFlrSf* (Exhibit 4) or it can be derived from the sum of *FinBsmtSF* and *UnfBsmtSF* (Exhibit 5), the variable has been removed.

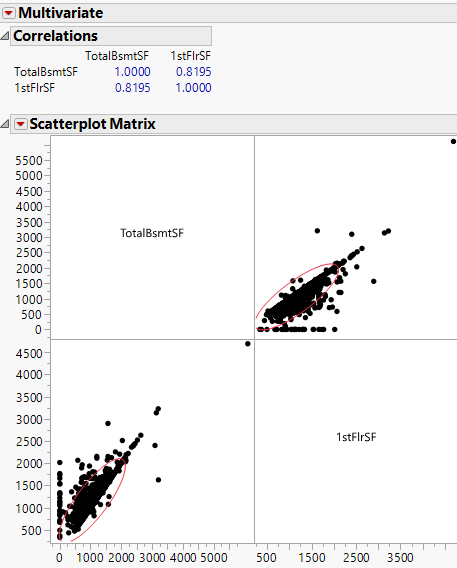
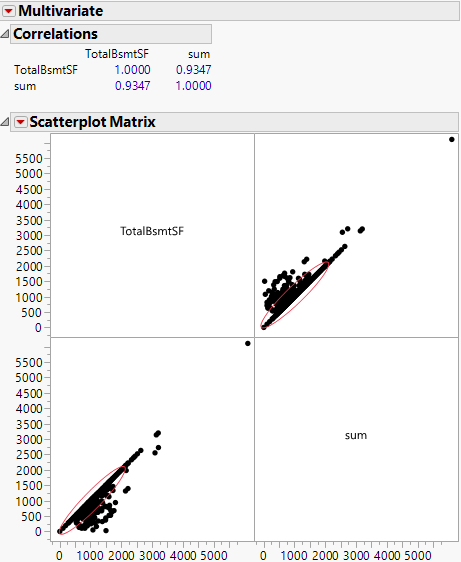
 

Exhibit 4: Correlation TotalBsmntSF vs. 1stFlrSF Exhibit 5: Correlation TotalBSF vs. sum of Fin and Unf Bsmt

The variable *GrLivArea* has been removed as there is high correlation between *GrLivArea* and sum of *1stFlrSf* and *2ndFlrSF* (See Exhibit 6).

Exhibit 7 displays a contingency table plotted between *Fireplaces* and *FireplaceQuality.* As almost allcategories of *FireplaceQuality* fall under any one category in *Fireplaces*, *FireplaceQuality* has been removed.

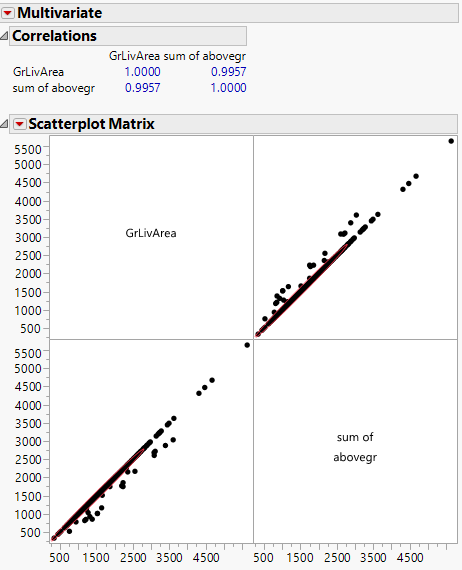
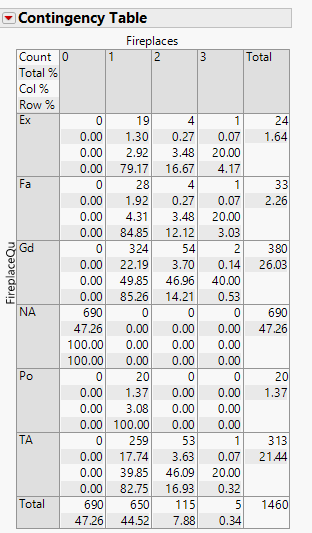
 

Exhibit 6: Correlation GrLivArea vs. sum Exhibit 7:Contingency table Fireplaces vs. FireplaceQu

Exhibit 8 displays a contingency table plotted between *SaleType* and *SaleCondition* which shows almost all categories of *SaleCondition* fall under any one category in *SaleType*. So, the team removed *SaleType* as it can be derived from *SaleCondition*. Similarly, *BsmntCond* has been removed as all categories fall under any category of *BasementFinType1* (see Exhibit 9).

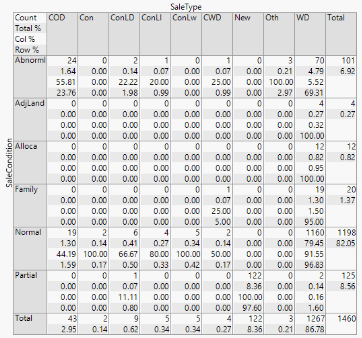
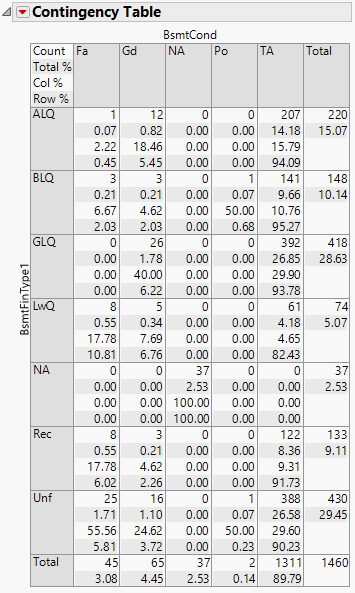
 

Exhibit 8: Contingency table SaleCond vs. SaleType Exhibit 9: Contingency BsmntFinType1 vs. BsmntCond

# 

# 

## **Data Modification**

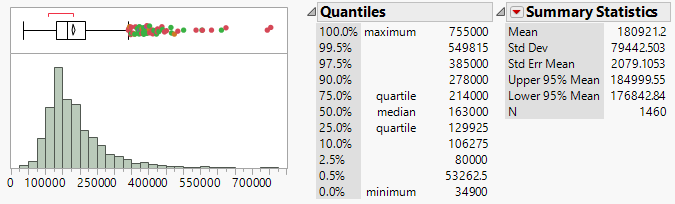
There were a few columns that needed modification using recode or a data type change. The team recognized these changes by reviewing each column’s property, the data in the column and the data dictionary definition. The team processed the changes mentioned below to clean the data to match the data dictionary definition. Two Columns (*TotalFullBath* and *TotalHalfBath*) were added to combine information from four columns. See Table 3 for better understanding of all changes the team made to the original data.

|  |  |
| --- | --- |
| **Column** | **Modification** |
| MMSubClass | Changed to Nominal |
| MS Zoning | Changed to Nominal. Changed ‘C (all)’ to ‘C’ as per data dictionary definition. |
| LotFrontage | Changed Blanks to 0. |
| OverallQual | Changed to ordinal as the data dictionary has mentioned the ranking of the categories. |
| OverCondition | Changed to ordinal as the data dictionary has mentioned the ranking of the categories. |
| MasVnrType | Changed N/A to None. N/A is not a valid data dictionary value |
| MasVnrArea | Changed Blanks to 0. |
| ExteriorQual | Changed to ordinal as the data dictionary has mentioned the ranking of the categories. |
| ExteriorCont | Changed to ordinal as the data dictionary has mentioned the ranking of the categories. |
| BsmtQual | Changed to ordinal as the data dictionary has mentioned the ranking of the categories. |
| BsmtCont | Changed to ordinal as the data dictionary has mentioned the ranking of the categories. |
| Electrical | One record had NA, so changed it to mode value of ‘SBrkr’. |
| TotalFullBath | Created a new column from *BsmtFullBath* + *FullBath.* Will be used as nominal or continuous as per the model. |
| TotalHalfBath | Created a new column from *HalfBath* + *BsmtHalfBath.* Will be used as nominal or continuous as per the model. |
| BedroonAbdGr | This column will be used as nominal for decision trees and continuous for other types of models that require continuous variables. |
| KitchenQual | Changed to Ordinal |
| TotalRmsAbvGrd | This column will be used as nominal for decision trees and continuous for other types of models that require continuous variables. |
| Fireplaces | Changed to Nominal |
| FireplaceQual | Changed to Ordinal |
| MoSold | Changed to Nominal |

Table 3: Data variable modifications

## **Binning *SalePrice***

The target variable *SalePrice* is a continuous variable. See exhibit 10 for the distribution of *SalePrice.* To implement different types of models, the team has binned *SalePrice* in two different ways. Binning refers to dividing continuous variables into groups to create a nominal variable. The team chose to bin *SalePrice* to have the option of creating a model for both continuous and nominal target variable. The result is the creation of two variables: *SalePrice2Cat* and *SalePrice3Cat*.

Exhibit 10: *SalePrice* Distribution

By creating the new variable *SalePrice2Cat*, the team has created a nominal target variable for *SalePrice* that is binary, or only has 2 values. The team used the median value of $163,000 to split the house price into two categories (See Exhibit 11). The team also created a variable *SalePrice3cat* to classify *‘SalePrice’* intothree categories. The team divided the distribution of *SalePrice* into approximate thirds to decide the range of each category (see Exhibit 12). Thus, the team has three different columns that could be used as target variables depending on the type of model. Each target variable’s range is defined in Table 4.

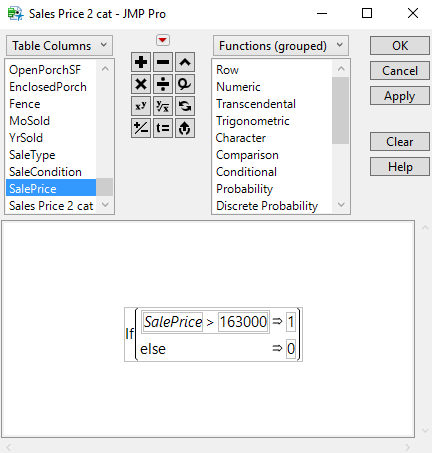
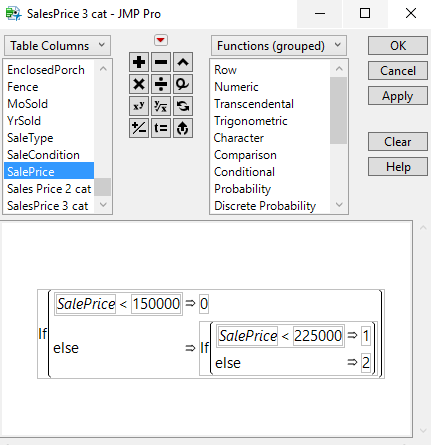
 

Exhibit 11: Formula for *SalePrice2Cat*  Exhibit 12: Formula for *SalePrice3Cat*

|  |  |
| --- | --- |
| **Target Variable** | **Description** |
| SalePrice | Continuous Variable. Exact sale price of the house. |
| SalePrice2Cat | Nominal Variable. Values are defined below:  Low: 0 - Sale Price = $0 - $162,999  High: 1 - Sale Price = $163,000 - ∞ |
| SalePrice3Cat | Nominal Variable. Values are defined below:  Low: 0 - Sale Price = $0 - $149,999  Medium: 1 - Sale Price = $150,000 - $224,999  High: 2 - Sale Price - $225,000- ∞ |

Table 4: Target Variable Definitions

## **Data Visualization**

# 

The team used visualization techniques to explore patterns and gain useful insights of variables with respect to the target variable. Table 5 consists of the graphs and insights that the team deduced.

|  |  |  |
| --- | --- | --- |
| **Graph** | **Insights** | |
| **LotShape vs. SalePrice** | There is not much variation in the *SalePrice* with respect to IR1(Slightly irregular) and Regular *LotShape* over the years sold. In IR2 and IR3, a fluctuation can be seen in the *SalePrice*.  Average *SalePrice* of houses with irregular (IR1, IR2, IR3) *LotShape* is more than that of Regular. | |
| **Neighborhood vs. SalePrice** | The average *SalePrice* is highest when there is no ridge in the neighborhood and it is least when there is a meadow. | |
| **Age of the house vs. SalePrice** | The average *SalePrice* is highest when the age of house is between 0 and 5 years and its least when it’s more than 60 years old. The average decreases with age of house. | |
| **HouseStyle vs. SalePrice** | Average *SalePrice* increase with the number of story with an exception of 1 story building.  Finished houses cost more than unfinished houses. | |
| **GarageType vs SalePrice** | | Houses with a built-in garage have the highest average *SalePrice* and houses with no garages have the least average *SalePrice*. |
| **BsmntExpr vs. SalePrice** | | Houses with a good basement exposure have the highest average *SalePrice* and houses with no basements have the least average *SalePrice*. |
| **TotalRmsAbvGrd vs. SalePrice** | | Average *SalePrice* of the houses increases with the number of rooms and exhibits a decreasing pattern beyond 11. Houses with 11 rooms above grade have the highest average *SalePrice*. |
| **Fireplace vs SalePrice** | | Average *SalePrice* of the houses increases with the number of fireplaces. |

Table 5: Data visualization of variables with the target variable

## **Missing Value Treatment**

In this data set, several variables contained missing observations which needed to be further evaluated before any modeling. To further identify these missing observations, a missing data pattern analysis was completed as seen below in Exhibit 13.

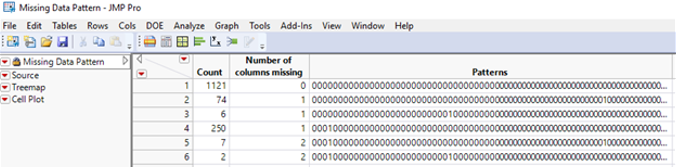


Exhibit 13: Initial Missing data pattern

In the first example, as shown in Exhibit 14, the team substituted the missing values in *LotFrontage* with ‘0’. This was done as a higher cost of error would be associated with replacing the missing observations with the mean or mode due to its potential effect on housing price. For example, if the linear feet of street in the front of the property (*LotFrontage*) was replaced with the mean of the data, say 70.0 linear feet when the house only had 21.0 linear feet the housing price may reflect to be significantly higher than what it is worth.

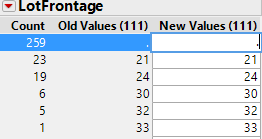
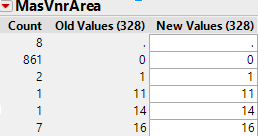
 

Exhibit 14: Missing value treatment for *LotFrontage* Exhibit 15: Missing value treatment for *MasVnrArea*

Similarly, in the second example the missing value ‘.’ in *MasVnrArea* were replaced with ‘0’. In this example the missing values were carefully identified by the team and changed to not only the mode but to the least predicted cost of error for the models accuracy.

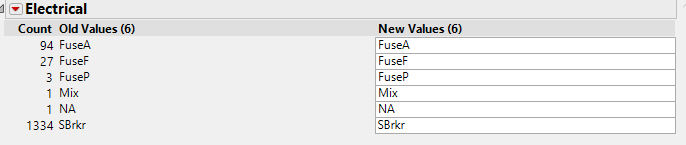


Exhibit 16: Missing value treatment for *Electrical*

In the third example, as shown in Exhibit 16, the team chose to replace the ‘NA’ value with the mode of electrical type titled *SBrkr* representing “circuit breaker”. The one missing value was replaced with mode to ensure the standard deviation, mean and median of the data would not be greatly altered with the replacement.

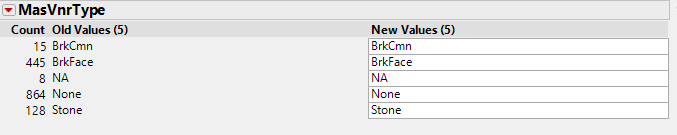


Exhibit 17: Missing value treatment for *MasVnrType*

In the fourth example, as shown in Exhibit 17, the team chose to replace the eight ‘NA’ observations to the variable titled ‘None’ which also was the column’s mode. The 864 values that had a variable of ‘None’ simply meant that the houses in the data set did not have Masonry Veneer Type.

Once all the missing value treatment was complete the missing data pattern analysis was completed for a second time. Exhibit 18 seen below depicts that the data set no longer contains any missing observations once this process was complete.

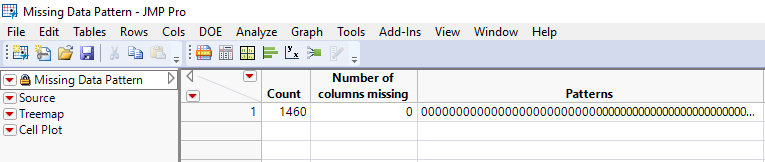
****

Exhibit 18: Missing data pattern for modified data set

## **Outlier Analysis**

Outlier analysis within the data was a major task in the exploration and modification process of the team's data set. To truly understand the outliers within the data set several outlier detection methods were used to visually inspect the values that did not conform to the expected pattern of the column. As the team does not know the entirety of the business scenario, the decision was made that all the outlier variables should remain included due to their potential importance in the data set.

Univariate

The team first completed its univariate outlier detection by using distribution graphs which gave insight into the data distribution. Exhibit 19 provides an overview of the distribution graphs for a sample of the variables. The team observed that significant amount of values fell outside the whiskers of the box plot. The entirety of the distribution graphs for all variables can be seen in [Appendix D.1](#_Appendix_D.1:_Univariate:). For extremely skewed variables, continuous fit was applied to transform the variable (see Exhibit 20).

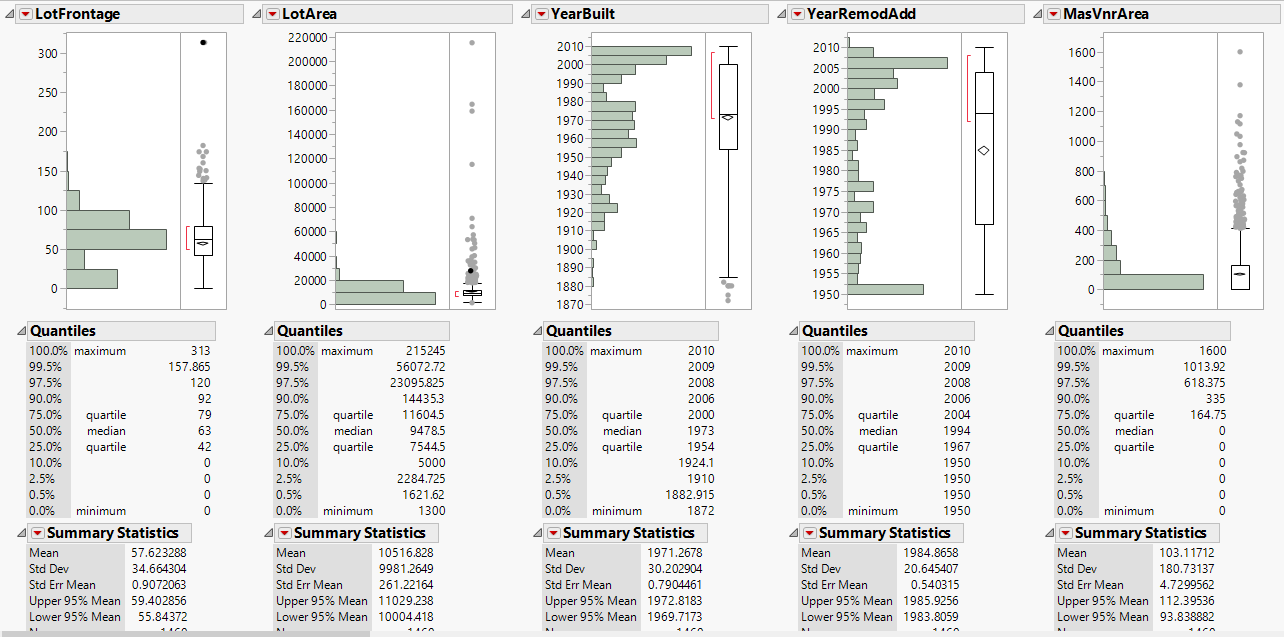


Exhibit 19: Distribution graph for univariate outlier analysis

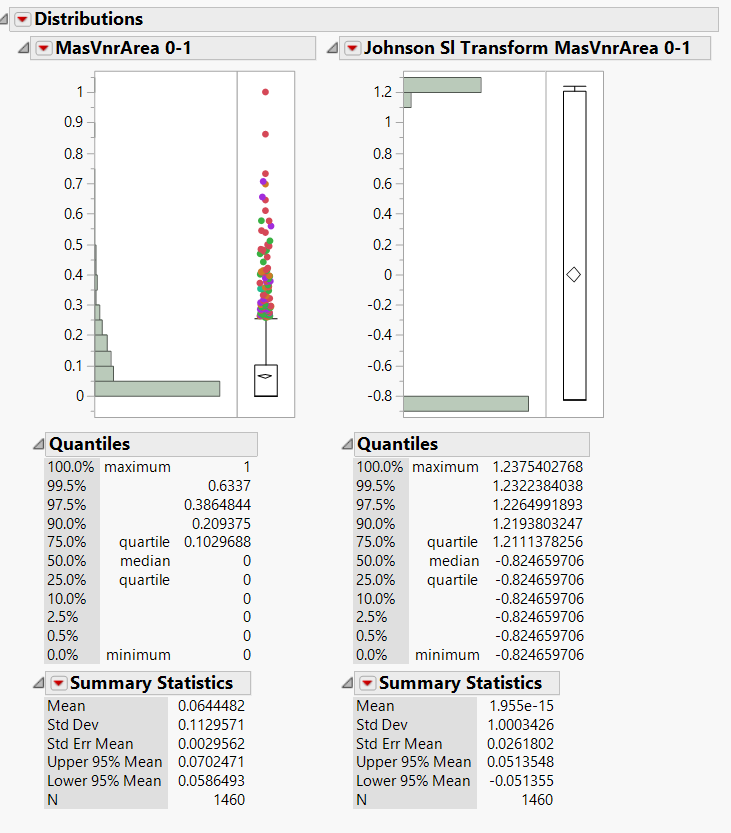


Exhibit 20: Extremely skewed MasVnrArea before and after transformation

Bivariate

To conduct a bivariate outlier analysis, the team computed correlation between *1stFlrSF* and *2ndFlrSF*. The narrowness of the ellipse represents the correlation between two variables. The values lying outside the ellipse is usually considered outliers. In the below example, as shown in Exhibit 21, we can see that the variables are not correlated and a lot of values will be considered as outliers. Similarly, a scatterplot matrix was plotted for all continuous variables and a full overview including the correlation review of the matrix can be found in [Appendix D.2](#_Appendix_D.2:_Bivariate:).

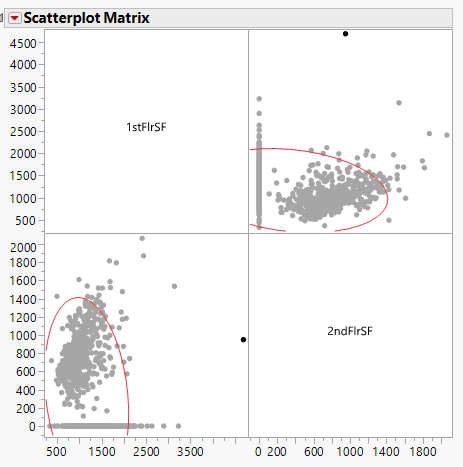


Exhibit 21: Scatterplot matrix of 1stFlrSF and 2ndFlrSF

Multivariate

Lastly, the team used the robust estimation method for outlier detection called Mahalanobis Distance which can be seen in Exhibit 22. This method identifies the multivariate observations that are inconsistent with the correlation structure of the data set. As seen below in Exhibit 22 and 23, this data set had few outliers that fell outside the immediate clustering of variables.

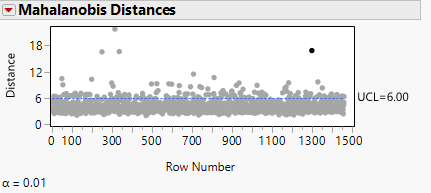


Exhibit 22: Mahalanobis distances plotted for the continuous variables with an Alpha level 0.01

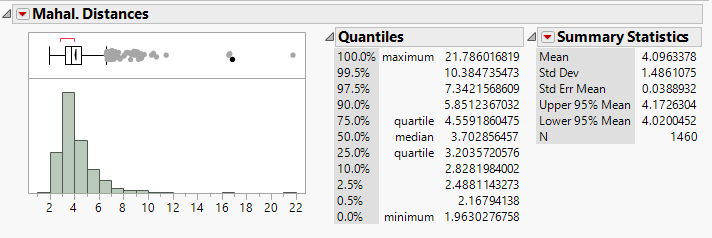


Exhibit 23: Distribution graph of saved Mahalanobis distance

As mentioned in the introduction to outlier analysis, the team chose not to remove any outliers in this data set due to the potential information that can be found by further evaluation and investigation of the outliers.

## **Scaled Variable**

In the final dataset, the user will find that all continuous variables have been scaled. These columns can be identified in the header as ‘*Column Name 0-1’.* The MinMax formula (Xi-Min(X)/(Max(X)-Min(X)) was applied to each continuous column to standardize the records to the scale 0-1. The

advantage of scaling each continuous variable is, when finding relationships between two variables, any variable with larger scale will exert more influence on the relation when compared to other variables. Since there are no variables in the dataset that should be given more weight, all continuous variables have been brought down to the same scale. See Exhibit 24 for an example of a new scaled column that was created from the MinMax formula. The scaled variables are recommended for modeling since they have the same variance as of the original column, just shown between the 0-1.

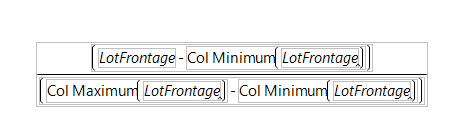


Exhibit 24: *LotFrontage* scaling formula

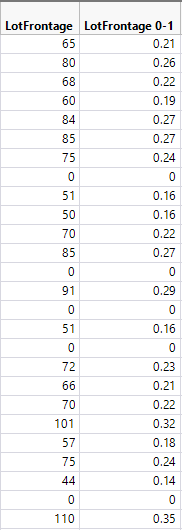
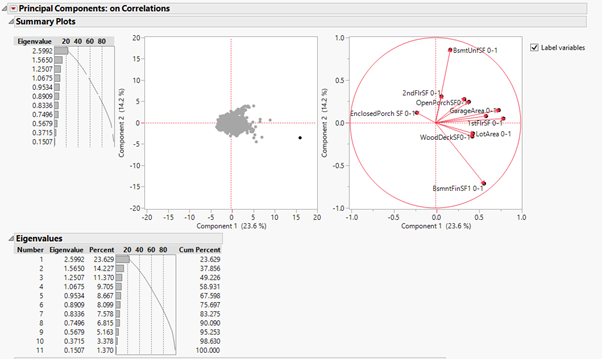


Exhibit 25: *LotFrontage* before and after scaling

## **Principle Component Analysis**

The team performed a Principle Component Analysis to reduce the dimensionality, while retaining maximum variance of the continuous data information. Once the analysis was completed, the team decided to keep a 95% variance of the data using nine of the eleven variables (See Exhibit 26). Due to the minimal number of columns that would be reduced in the data set and as the acceptable variance for this business decision is unknown, the team did not use it in modeling.



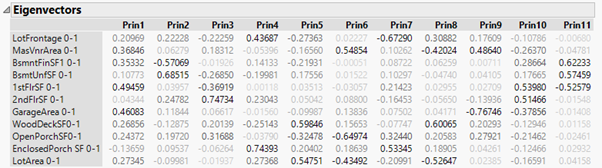


Exhibit 26: Principle Component Analysis outcomes on data set

# 

# Modeling

After pre-processing the data, the team went through each modeling technique with the identified target variables. The team compared and analysed the results of all implemented models to identify the most accurate model.

## **Cluster Analysis**

Clustering is the process of grouping observations into clusters so that elements within a cluster look very similar while those in different clusters look diverse. Different algorithms and methods can be used to determine the similarity within clusters. On performing hierarchical clustering using continuous explanatory variables in JMP, 20 distinct clusters were reported. The team set the number of clusters to 6 to absorb the information better. Out of the 6 clusters, 3 of the clusters accounted for about 75% of the information (see Exhibit 27). Exhibit 28 shows the information that was provided by 6 clusters when hierarchical method was used with nominal variables.

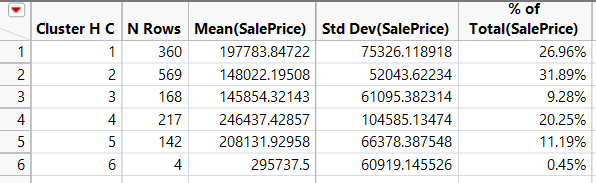


Exhibit 27(a): Summary Table of Clusters (hierarchical continuous) with *SalePrice*

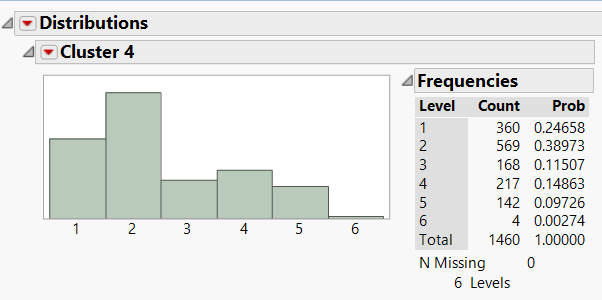


Exhibit 27(b): Distribution graph of Clusters

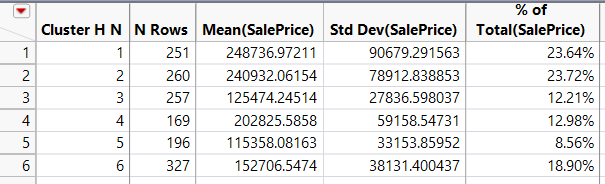


Exhibit 28: Summary table of Clusters (hierarchical nominal) with *SalePrice*

The mean *SalePrice* for each cluster formed by K-Means clustering, can be seen in Exhibit 29. The team used clustering analysis as a primary modeling tool to learn the data. The team also learned that, with so many explanatory variables, clustering analysis is difficult to explain without increasing modeling sophistication. Clustering may be more useful if there is a handful of variables that have high explanatory power on the target variable. For now, it serves as visualization on how the data can be grouped. See [Appendix E](#_Appendix_E:_Clustering) for more details on clustering.

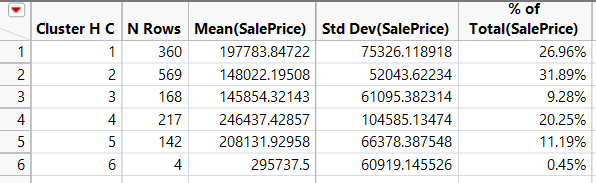


Exhibit 29 (a): Cluster Analysis K-means: Nominal Variables

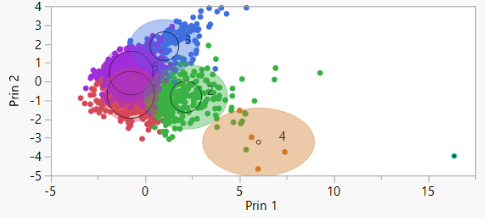


Exhibit 29(b): Biplot of K-Means clusters

# 

# 

## **Stratified Sampling**

The add-in ‘Stratified Split Balanced’ was used to partition the data. As the data set is small, it has been split into training and validation. The sample was stratified on the Nominal Target Variable *SalePrice2Cat* to create 2 balanced samples. The training sample has 70% of the observations, and validation has 30%. Each sample would have 50% observations where *SalePrice2Cat* = 0, and 50% observations where *SalePrice2Cat* = 1 (See definitions in Data Dictionary). Exhibit 30 shows the distribution in the number of observations in each sample.

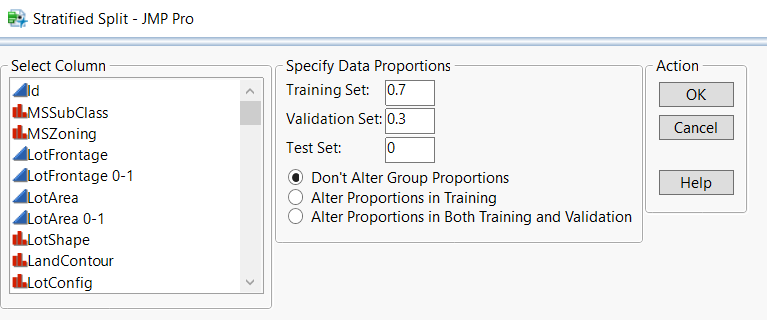


Exhibit 30 (a): Specifying data proportions in the add-in

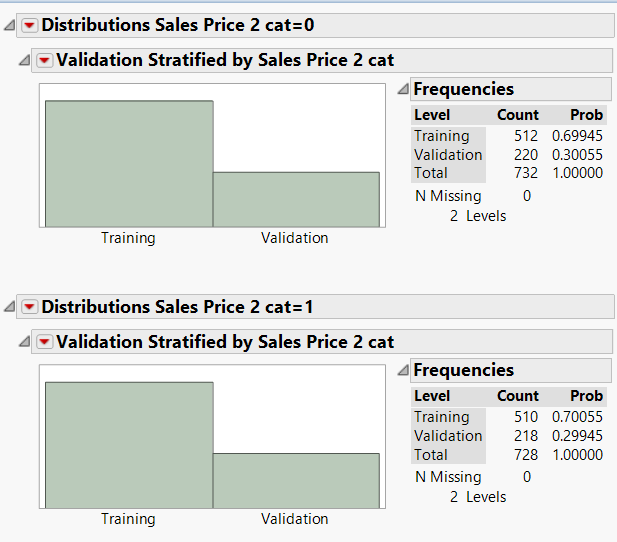


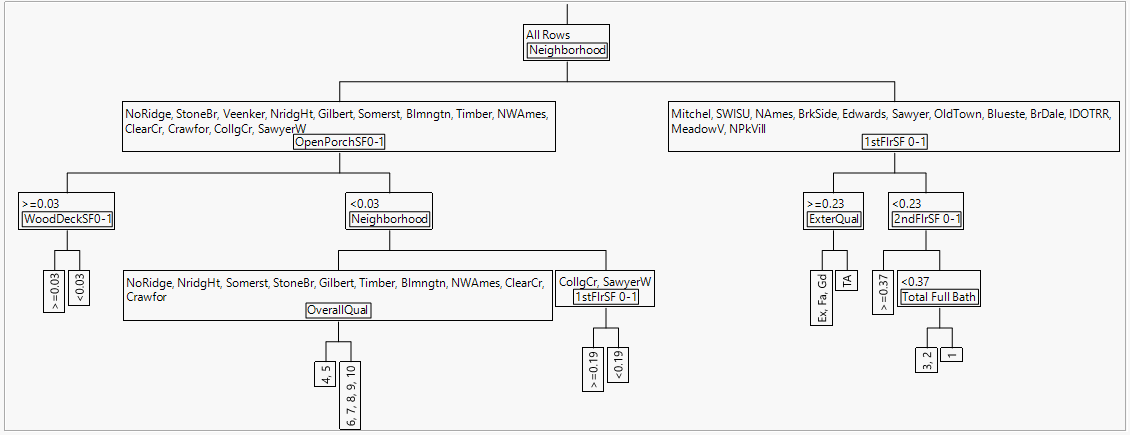
Exhibit 30(b): Distribution graphs of *SalePrice2cat* 0 and 1

## **Predicting two category *SalePrice***

SalePrice has been binned into two categories, namely ‘High’ (1) and ‘Low’ (0). As the business scenario is uncertain, the cutoff value used for assessment metrics is the default value of 0.5.

Decision Tree

Decision tree is a classification algorithm that repeatedly splits observations into branches and creates a tree of predictions. This modelling technique is usually the easiest to explain to business partners. For predicting *SalePrice2Cat* using Decision Tree, all continuous and nominal explanatory variables were used. The tree (Exhibit 31) split 10 times, with *Neighborhood* having the highest information gain. An example of branch expansion is “If overall Quality is 4 or 5 and the Neighborhood is ‘No ridge, StoneBr,..Crawfor’ and OpenPorchSF is less than 0.03, then *SalePrice2Cat* is High. The accuracy for this model is 87.43% (see Table 6). See [Appendix F.1.1](#_Appendix_F.1.1:_Decision) for more detailed information.

Exhibit 31: Small Tree View for Decision Tree predicting *SalePrice2Cat*

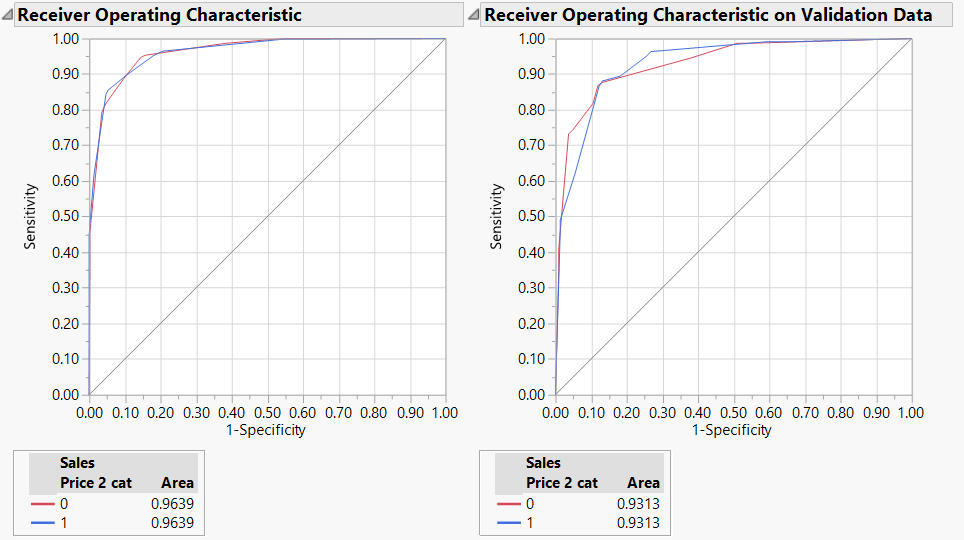


Exhibit 32: ROC curve for Decision Tree predicting *SalePriceCat2*

Discriminant Analysis

Linear Discriminant Analysis is a classification method which uses the approach of Eigen vectors. The model is limited as only continuous explanatory variables can be used. From the ROC Curve (Exhibit 33), the team observed that the model performs equally well in predicting true negatives and true positives, i.e, the Area Under Curve (AUC) is equal for both the categories. The accuracy for this model is 87.67%. (see Table 6) See [Appendix F 1.2](#_Appendix_F.1.2:_Linear) for more detailed information.

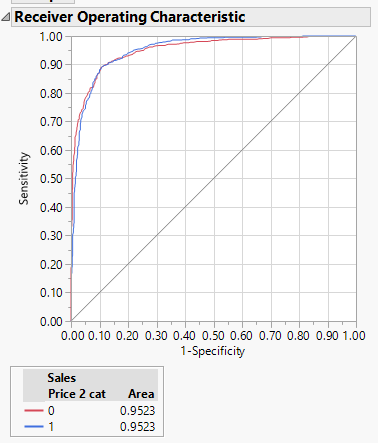


Exhibit 33: ROC curve for Linear Discriminant Analysis for predicting SalePrice2Cat

Logistic Regression

Logistic regression is a regression model used for predicting binary dependent variable. The team created a Logistic Regression model to predict *SalePrice2Cat* using continuous explanatory variables. As seen in Exhibit 35, the variable with a low P value have high statistical significance. The accuracy for this model is 88.12% (see Table 6). See [Appendix F 1.3](#_Appendix_F.1.2:_Logistic) for more detailed information.

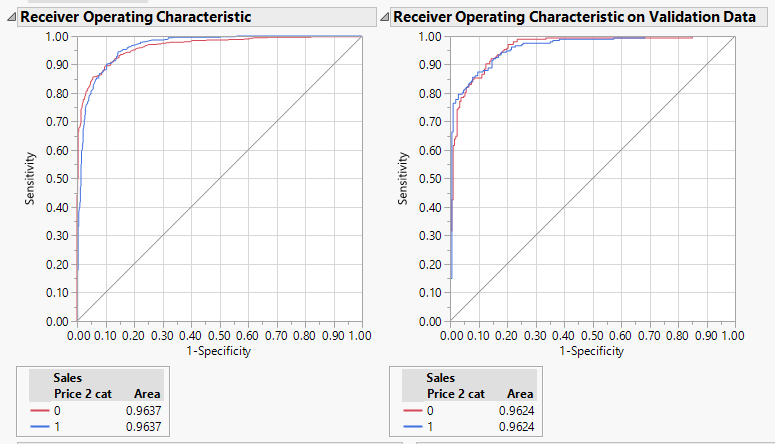


Exhibit 34: ROC curve for Logistic Regression for predicting *SalePrice2Cat*

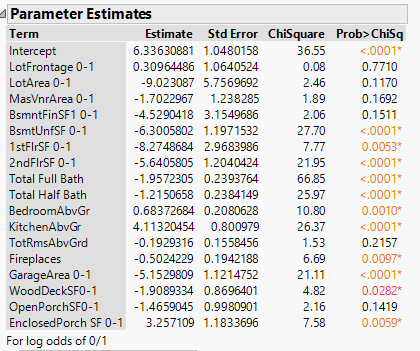


Exhibit 35: Parameter Estimates for Logistic Regression

Neural Networks

Neural networks is the model that is referred to as being biologically inspired or the model that most represents the brain due to its ability to learn. This analytical technique can determine the function that it should perform based on the inputs from the sample that is provided and offer outputs that correspond to the inputs even if the model is not familiar with it. The team used this model two ways to predict *SalePrice2Cat*. The first model (Exhibit 36) is complex, with 3 and 2 levels of Tangential, Linear and Gaussian in the first layer and second layer respectively. The second model is simpler, it has only 1 layer of linear type. Both continuous and nominal variables have been used as explanatory variables. The accuracy for this model is 89.95% (see Table 6). See [Appendix F1.4](#_Appendix_F.1.4:_Neural) for detailed information.

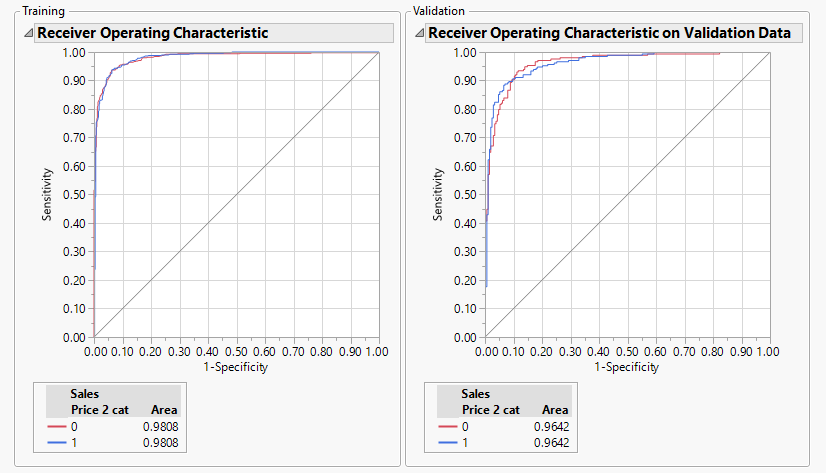


Exhibit 36: ROC curve for complex neural networks for predicting *SalePrice2Cat*

Ensemble

Ensemble modeling is the process of running two or more related but different analytical models and then combining the results into a single prediction to improve the accuracy of predictive analytics. As seen below in Table 6, there is an overview of all models and ensemble models that were used in the data set.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Predicted | | | |  |  |  |
| Model | Actual | Training | | Validation | | Accuracy (%) | Sensitivity (%) | Specificity (%) |
| Sales Price | 0 | 1 | 0 | 1 |
| Decision Tree | 0 | 485 | 27 | 191 | 29 | 87.43 | 88.07 | 86.81 |
| 1 | 74 | 436 | 26 | 192 |
|  | | | | | | | | |
| Linear Discriminant Analysis | 0 | 453 | 59 | 189 | 31 | 87.67 | 89.44 | 85.90 |
| 1 | 51 | 459 | 23 | 195 |
|  | | | | | | | | |
| Logistic Regression | 0 | 459 | 53 | 195 | 25 | 88.12 | 87.61 | 88.63 |
| 1 | 51 | 459 | 27 | 191 |
|  | | | | | | | | |
| Neural Network | 0 | 480 | 32 | 195 | 25 | 89.95 | 91.28 | 88.63 |
| 1 | 36 | 474 | 19 | 199 |
|  | | | | | | | | |
| Neural Network Single layer | 0 | 487 | 25 | 192 | 28 | 89.72 | 92.20 | 87.27 |
| 1 | 26 | 484 | 17 | 201 |
|  | | | | | | | | |
| Ensemble Decision Tree | 0 | 487 | 25 | 192 | 28 | 89.95 | 92.66 | 87.27 |
| 1 | 21 | 489 | 16 | 202 |
|  | | | | | | | | |
| Ensemble LDA | 0 | 484 | 28 | 197 | 23 | 90.86 | 92.20 | 89.54 |
| 1 | 33 | 477 | 17 | 201 |
|  | | | | | | | | |
| Ensemble Logistic Regression | 0 | 486 | 26 | 197 | 23 | 91.10 | 92.66 | 89.55 |
| 1 | 23 | 487 | 16 | 202 |
|  | | | | | | | | |
| Ensemble Neural Network | 0 | 482 | 30 | 196 | 24 | 90.87 | 92.66 | 89.09 |
| 1 | 27 | 483 | 16 | 202 |

Table 6: Comparison of all models used in predicting SalePrice2cat

True positives: Price of house is high and has been predicted as high

False positives: Price of house is low and has been predicted as high

True negatives: Price of house is low and has been predicted as low

False negatives: Price of house is high and has been predicted as low

Accuracy: (True Positives+True Negatives)/Sum of all observations

Sensitivity: (True positives)/(True positive + False Negatives)

Specificity: (True negatives)/(True negatives+ False Positives)

## **Predicting three category *SalePrice***

*SalePrice* has been binned into three categories, namely ‘High’ (2), ‘Medium’(1) and ‘Low’ (0).

Decision Tree

For predicting *SalePrice3Cat* using Decision Tree, all continuous and nominal explanatory variables were used. The tree (Exhibit 37) split 20 times, with *Neighborhood* having the highest information gain. From the ROC Curve (Exhibit 38), the team observed that the model performs best in predicting category ‘High’, i.e, the Area Under Curve (AUC) for ‘High’ is greater than the other two categories. The accuracy for this model is 77.85% (see Table 7). See [Appendix F.2.1](#_Appendix_F.1.1:_Decision) for more detailed information.

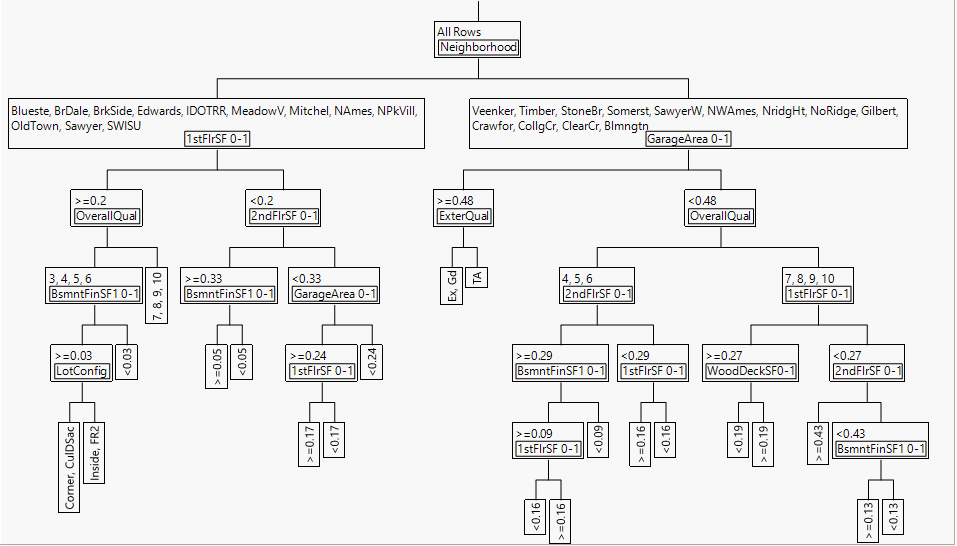


Exhibit 37: Decision Tree Results *SalePrice3Cat*

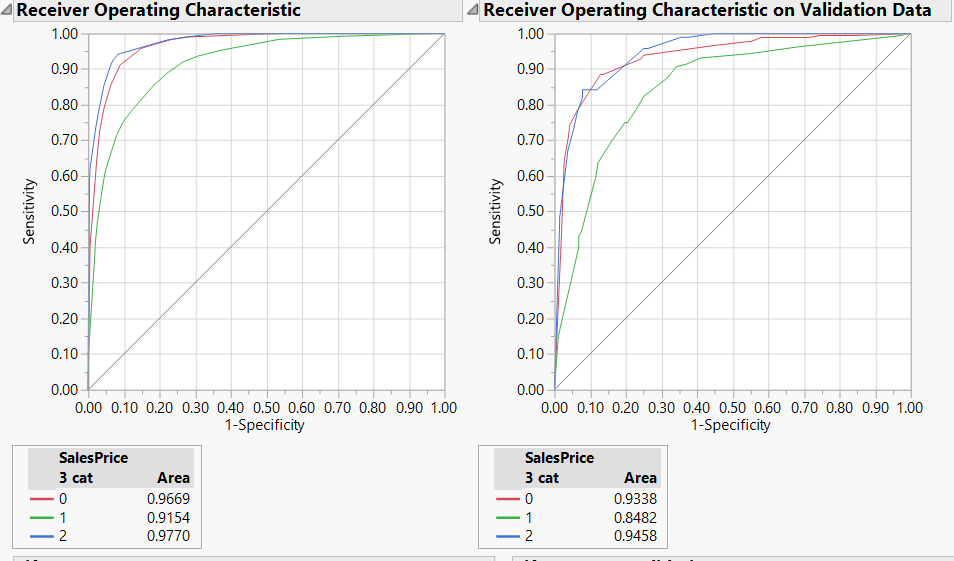


Exhibit 38: Decision Tree ROC curves *SalePrice3C**at*

Discriminant analysis:

Continuous explanatory variables were used, for Discriminant Analysis model to predict the 3-category sale price. From the ROC Curve (Exhibit 39), the team observed that the model performs best in predicting category ‘High’, i.e, the Area Under Curve (AUC) for ‘High’ is greater than the other two categories. The accuracy for this model is 79% (see Table 7). See [Appendix F2.2](#_Appendix_F.2.2:_Linear) for more detailed information.

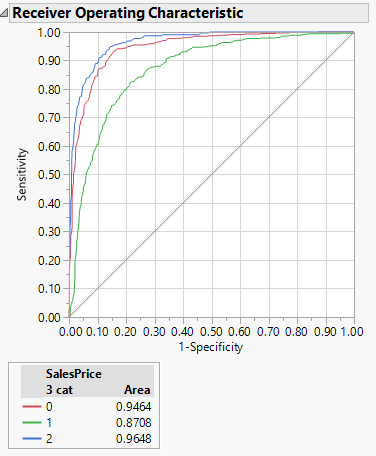


Exhibit 39: Discriminant analysis-ROC curves *SalePrice3Cat*

Neural Networks:

The team used two layered neural networks model to predict *SalePrice3Cat.* The first layer and second layer had 3 and 2 levels of Tangential, Linear and Gaussian respectively. Both continuous and nominal variables have been used as explanatory variables. Please see Exhibit 40 for the ROC curves corresponding to each target output. The accuracy for this model is 83.11% (see Table 7). See [Appendix F2.3](#_Appendix_F.2.3:_Neural) for more detailed information.

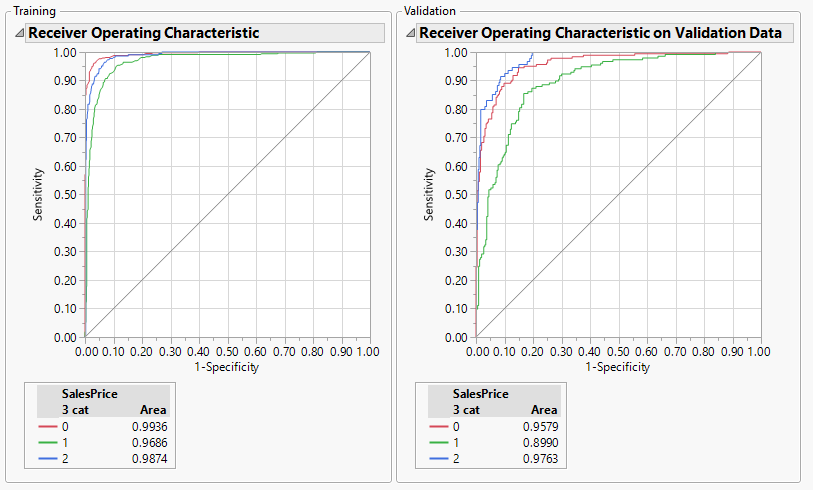


Exhibit 40: Neural Networks-ROC curves *SalePrice3Cat*

Ensemble Models:

As mentioned earlier, Ensemble models use the output from different prediction models and generate the results into one single prediction. As seen below in Table 7, there is an overview of all models and ensemble models that were used to predict the 3-category sales price.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Predicted | | | | | |  |
| Model | Actual | Training | | | Validation | | | Accuracy (%) |
| SalePrice | 0 | 1 | 2 | 0 | 1 | 2 |
| Decision Tree | 0 | 394 | 38 | 0 | 151 | 31 | 1 | 77.85 |
| 1 | 52 | 279 | 35 | 23 | 116 | 21 |
| 2 | 1 | 32 | 191 | 0 | 21 | 74 |
|  | | | | | | | | |
| Linear Discriminant Analysis | 0 | 367 | 61 | 4 | 147 | 35 | 1 | 79.00 |
| 1 | 52 | 277 | 37 | 22 | 123 | 15 |
| 2 | 1 | 41 | 182 | 2 | 17 | 76 |
|  | | | | | | | | |
| Neural Networks | 0 | 418 | 14 | 0 | 156 | 26 | 1 | 83.11 |
| 1 | 23 | 316 | 27 | 20 | 129 | 11 |
| 2 | 0 | 21 | 203 | 0 | 16 | 79 |
|  | | | | | | | | |
| Ensemble Decision Tree | 0 | 421 | 11 | 0 | 157 | 25 | 1 | 83.33 |
| 1 | 26 | 320 | 20 | 21 | 132 | 7 |
| 2 | 0 | 25 | 199 | 0 | 19 | 76 |
|  | | | | | | | | |
| Ensemble LDA | 0 | 417 | 15 | 0 | 156 | 26 | 1 | 82.88 |
| 1 | 21 | 322 | 23 | 20 | 127 | 13 |
| 2 | 0 | 15 | 209 | 0 | 15 | 80 |
|  | | | | | | | | |
| Ensemble Neural Networks | 0 | 411 | 21 | 0 | 153 | 29 | 1 | 83.11 |
| 1 | 14 | 335 | 17 | 17 | 131 | 12 |
| 2 | 0 | 22 | 202 | 0 | 15 | 80 |
|  | | | | | | | | |

Table 7: Comparison of all models used in predicting SalePrice3cat

Accuracy: (True 0s+True 1s+ True 2s)/Sum of all observations

## **Predicting Continuous *SalePrice***

Decision Tree:

When creating a model with *SalePrice* as a Continuous Target variable, the decision tree had 23 splits. Overall Quality had the highest information gain in this model, but when predicting a nominal *SalePrice* Overall Quality had less significance. The R-Squared value of this model is around .75, so it can explain about 75% of the variability of *SalePrice*. See exhibit 41 for the actual vs. predicted graph. Please refer the [Appendix F.3.1](#_Appendix_F.3.1:_Decision) for more detailed information.

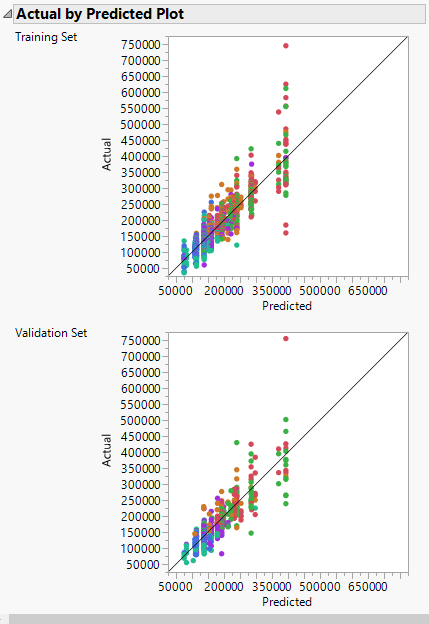
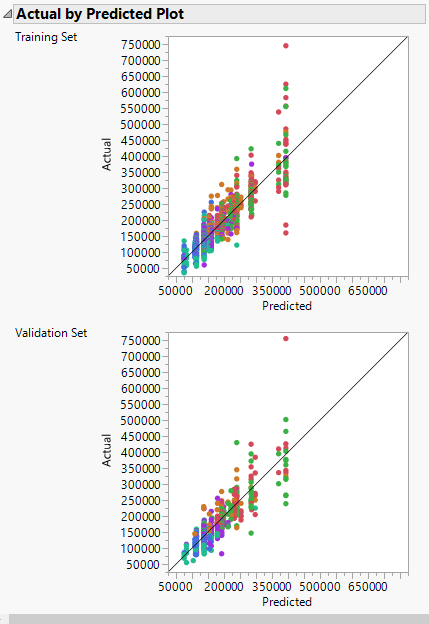


Exhibit 41: Actual vs predicted in Decision Tree

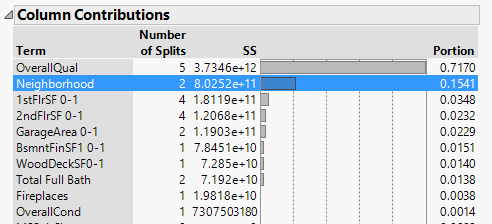


Exhibit 42: Column contribution in Decision Tree

Linear Regression-Least squares:

Least square is a statistical technique for finding the best fitting curve using the data points, simultaneously minimizing the sum of squares of residuals. Upon performing Linear regression using least squares method with *SalePrice* as target variable, a strong model with an R-Squared value of 0.82 is obtained. It can be observed from Exhibit 43 that the *OverallQuality* has the highest weightage in the model. Please refer the [Appendix F.3.2](#_Appendix_F.3.2:_Linear) for more detailed information.

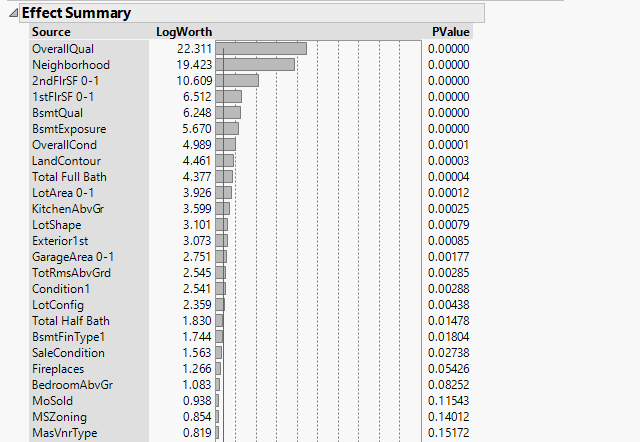


Exhibit 43: Effect summary for Least squares Linear Regression model

Linear Regression-Stepwise

Stepwise Linear Regression is a semi-automated process of building a model by successively adding or removing variables based solely on the *t*-statistics of their estimated coefficients. The best R-Squared value (Exhibit 44) for this model is around .87.



Exhibit 44: Model coefficients for Stepwise Linear Regression

Neural networks:

The team used two layered neural networks model to predict *SalePrice.* The first layer and second layer had 3 and 2 levels of Tangential, Linear and Gaussian respectively. Both continuous and nominal variables have been used as explanatory variables. Please see Exhibit 45 where Actual vs. Predicted values in both training and validation datasets have been plotted. It is evident from Exhibit 45 that the model performed equally well in training and validation data sets. The R-Squared value for this model is 0.86. Please refer [Appendix F3.4](#_Appendix_F.3.3:_Neural) for more detailed information.

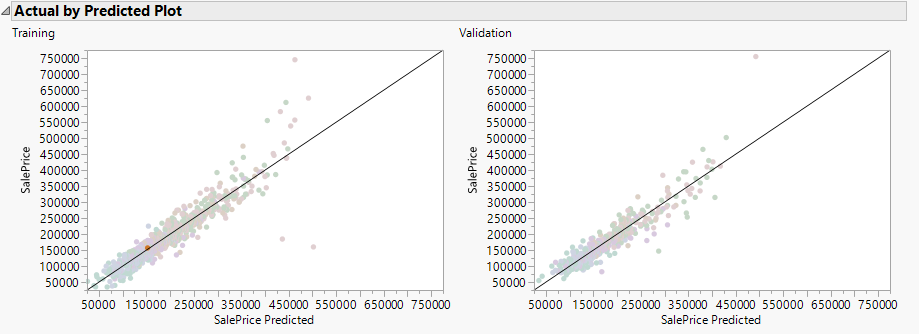


Exhibit 45: Actual vs. Predicted in Neural Networks

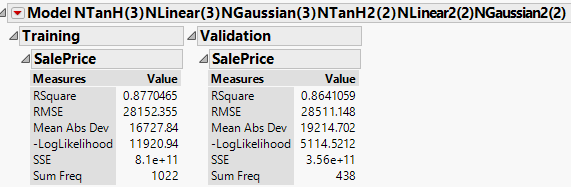


Exhibit 45: Model coefficients for Neural Networks

Ensemble Models:

Using the prediction output of the above models, the team performed Ensemble Decision tree analysis, Ensemble Linear regression and Ensemble Neural network models to predict the *SalePrice*. Table 8 is an overview of all models and ensemble models that were used to predict the continuous *SalePrice*.

|  |  |  |
| --- | --- | --- |
| Model | R-Squared value | |
| Training | Validation |
| Decision Tree | 0.791 | 0.734 |
| Linear Regression-Least Squares | 0.897 | 0.825 |
| Linear Regression- Stepwise | 0.875 | 0.855 |
| Neural Network | 0.877 | 0.864 |
| Ensemble Decision Tree | 0.918 | 0.856 |
| Ensemble Linear Regression | 0.899 | 0.861 |
| Ensemble Neural network | 0.896 | 0.870 |

Table 8: Model Comparison using R-Squared value

# Limitations

It should be noted in this report that there were several limitations that may provide to the accuracy of the modeling process including the number of observations, any housing marketing rules and regulations that were not known, and the accuracy of the data given. The team was able to develop the models with the available observations but additional insight to the data would only make the model more accurate and create better results.

# Discussion

After completing the entirety of the SEMMA process, the team could determine the most accurate model for each type of Target Variable. For the *SalePrice* continuous variable, the team has discovered that the Ensemble Linear Regression model is the most accurate, with an R-Squared value of .861. For *SalePrice2Cat*, the team found that the Ensemble Logistic Regression is the most accurate, with an accuracy of 91.10%. For *SalePrice3Cat*, the team learnt that the Ensemble Decision Tree Model is the most accurate, with an accuracy of 83.33%. Neural Network models had a slightly higher accuracy than other models, but as the cost of implementing these models are very high, the team decided to consider the next best model.

Our recommendation to a user of this model would be to use all three models to narrow down how much a house may cost. A sample use case would be when a real estate agent received a request to sell a house, they could enter in all the explanatory values into the prediction formula of all three models and obtain a project price of the house. This can assist the real estate agent in determining a price range of the house, before going to see the house. Specialized real estate agents may not want to accept business within a certain price range, and these models can automatically determine those results.

Users should be cautioned on the accuracy of the models. A full cost analysis should be completed to determine the cost of failure for the approximately ~15% incorrect values the model predicts. In the current state, the team recommends the models be used as informational, and no decision making should be done until a user confirms the results.

# Conclusion & Recommended Next Steps

The team believes that depending on which target variable is to be used, either Ensemble Linear Regression, Ensemble Logistic Regression or Ensemble Decision Tree model, will be the most viable resource for this specific data set in providing home appraisers and realtors the closest estimate for the housing market prices. During this process, the team learned the importance of number of observations. The chosen dataset only had 1460 observations which could be a limiting factor in our model accuracy. Also, the team learned more about the curse of dimensionality. The data set chosen originally had 81 variables, but while looking further the team realized 31 of them would not be useful in modelling, so they could be removed. With this reflection, the team understands the value of predictive modeling, but more importantly the value of data quality.

The team has a few recommended next steps highlighted below.

1. Collect more observations to add integrity to the model through continuous improvement
2. Users should decide the cost of inaccuracy and an alpha value to decide if the current model will suffice
3. Use the models in their current state as value add information, however users should add their expertise to any model outputs to conclude to a final decision
4. Depending on how the user intends to use the model, adjust the explanatory value

Overall, the team has completed the initial steps of creating a predictive model for *SalePrice*. Depending on the end user’s needs, this can be analysed further and additional steps can be taken.

# 

# Appendix

Appendix A: Data Dictionary

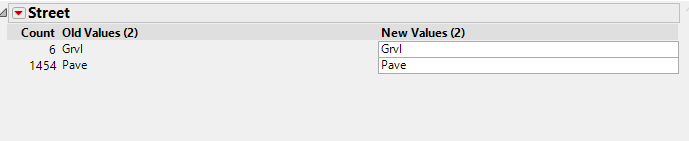
|  |  |  |
| --- | --- | --- |
| **Title** | **Description** | **Variables Used (Description)** |
| MSSubClass | Identifies the type of dwelling involved in the sale. | 20 (1-story 1946 & newer all styles) 30 (1-story 1945 & older) 40 (1-story w/finished attic all ages) 45 (1-1/2 story - unfinished all ages) 50 (1-1/2 story finished all ages) 60 (2-story 1946 & newer) 70 (2-story 1945 & older) 75 (2-1/2 story all ages) 80 (split or multi-level) 85 (split foyer) 90 (duplex - all styles and ages) 120 (1-story planned unit development - 1946 & newer) 150 (1-1/2 story planned unit development - all ages) 160 (2-story planned unit development - 1946 & newer) 180 (planned unit development - multilevel - including split lev/foyer) 190 (2 family conversion - all styles and ages) |
| MSZoning | Identifies the general zoning classification of the sale. | A (Agriculture)  C (Commercial) FV (Floating Village Residential)  I (Industrial)  RH (Residential High Density)  RL (Residential Low Density)  RP (Residential Low Density Park)  RM (Residential Medium Density) |
| LotFrontage | Linear feet of street connected to property |  |
| LotArea | Lot size in square feet |  |
| LotShape | General shape of property | Reg (Regular) IR1 (Slightly irregular) IR2 (Moderately irregular) IR3 (Irregular) |
| LandContour | Flatness of the property | Lvl (Near Flat/Level) Bnk (Banked- Quick and significant rise from street grade to building) HLS (Hillside- Significant slope from side to side) Low (Depression) |
| LotConfig | Lot configuration | Inside (Inside lot) Corner (Corner lot) CulDSac (Cul-de-sac) FR2 (Frontage on 2 sides of property) FR3 (Frontage on 3 sides of property) |
| LandSlope | Slope of property | Gtl (Gentle slope) Mod (Moderate slope) Sev (Severe slope) |
| Neighborhood | Physical locations within Ames city limits | Blmngtn (Bloomington Heights) Blueste (Bluestem) BrDale (Briardale) BrkSide (Brookside) ClearCr (Clear Creek) CollgCr (College Creek) Crawfor (Crawford) Edwards (Edwards) Gilbert (Gilbert) IDOTRR (Iowa DOT and Rail Road) MeadowV (Meadow Village) Mitchel (Mitchell) Names (North Ames) NoRidge (Northridge) NPkVill (Northpark Villa) NridgHt (Northridge Heights) NWAmes (Northwest Ames) OldTown (Old Town) SWISU (South & West of Iowa State University) Sawyer (Sawyer) SawyerW (Sawyer West) Somerst (Somerset) StoneBr (Stone Brook) Timber (Timberland) Veenker (Veenker) |
| Condition1 | Proximity to various conditions | Artery (Adjacent to arterial street) Feedr (Adjacent to feeder street)  Norm (Normal) RRNn (Within 200' of North-South Railroad) RRAn (Adjacent to North-South Railroad) PosN (Near positive off-site feature--park, greenbelt, etc.) PosA (Adjacent to positive off-site feature) RRNe (Within 200' of East-West Railroad) RRAe (Adjacent to East-West Railroad) |
| BldgType | Type of dwelling | 1Fam (Single-family detached) 2FmCon (Two-family conversion; originally built as one-family dwelling) Duplx (Duplex) TwnhsE (Townhouse end unit) Twnhsl(Townhouse inside unit) |
| HouseStyle | Style of dwelling | 1Story (One story) 1.5Fin (One and a half story: 2nd level finished) 1.5Unf (One and a hald sotry: 2nd leven unfinished) 2Story (Two story) 2.5Fin (Two and one-half story: 2nd level finished) 2.5Unf (Two and one-half story: 2nd level unfinished) Sfoyer (Split foyer) SLvl (Split level) |
| OverallQual | Rates the overall material and finish of the house | 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) |
| OverallCond | Rates the overall condition of the house | 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) |
| YearBuilt | Original construction date |  |
| YearRemodAdd | Remodel date (same as construction date if no remodeling or additions) |  |
| RoofStyle | Type of roof | Flat (Flat) Gable (Gable) Gambrel (Gabrel- Barn) Hip (Hip) Mansard (Mansard) Shed (Shed) |
| Exterior1st | Exterior covering on house | AsbShng (Asbestos Shingles) AsphShn (Asphalt Shingles) BrkComm (Brick Common) BrkFace (Brick Face) CBlock (Cinder Block) CemntBd (Cement Board) HdBoard (Hard Board) ImStucc (Imitation Stucco) MetalSd (Metal Siding) Other (Other) Plywood (Plywood) PreCast (PreCast) Stone (Stone) Stucco (Stucco) VinylSd (Vinyl Siding) Wd Sdng (Wood Siding) WdShing (Wood Shingles) |
| MasVnrType | Masonry veneer type | BrkCmn (Brick Common) BrkFace (Brick Face) Cblock (Cinder Block) None (None) Stone (Stone) |
| MasVnrArea | Masonry veneer area in square feet |  |
| ExterQual | Evaluates the quality of the material on the exterior | Ex (Excellent) Gd (Good) TA (Typical/Average) Fa (Fair) Po (Poor) |
| ExterCond | Evaluates the present condition of the material on the exterior | Ex (Excellent) Gd (Good) TA (Typical/Average) Fa (Fair) Po (Poor) |
| Foundation | Type of foundation | BrkTil (Brick & Tile) Cblock (Cinder Block) Pconc (Poured Contrete) Slab (Slab) Stone (Stone) Wood (Wood) |
| BsmtQual | Evaluates the height of the basement | Ex (Excellent- 100+ inches) Gd (Good- 90-99 inches) TA (Typical- 80-89 inches) Fa (Fair- 70-79 inches ) Po (Poor- <70 inches) NA (No basement) |
| BsmtCond | Evaluates the general condition of the basement | Ex (Excellent) Gd (Good) TA (Typical- Slight dampness allowed) Fa (Fair- Dampness or some cracking or settling) Po (Poor- Severe cracking, settling, or wetness) NA (No basement) |
| BsmtExposure | Refers to walkout or garden level walls | Gd (Good Exposure) Av (Average Exposure- split levels or foyers typically score average or above) Mn (Minimum Exposure) No (No Exposure) NA (No Basement) |
| BsmtFinType1 | Rating of basement finished area | GLQ (Good Living Quarters) ALQ (Average Living Quarters) BLQ (Below Average Living Quarters Rec (Average Rec Room) LwQ (Low Quality) Unf (Unfinished) NA (No Basement) |
| BsmtFinSF1 | Type 1 finished square feet |  |
| BsmtUnfSF | Unfinished square feet of basement area |  |
| TotalBsmtSF | Total square feet of basement area |  |
| Electrical | Electrical system | SBrkr (Standard Curcuit Breakers and Romex) FuseA (Fuse Box over 60 AMP and all Romex Wiring- Average) FuseF (60 AMP Fuse Box and mostly Romex wiring- Fair) FuseP (60 AMP Fuse Box and mostly knob and tube wiring- Poor) Mix (Mixed) |
| 1stFlrSF | First Floor square feet |  |
| 2ndFlrSF | Second floor square feet |  |
| GrLivArea | Above grade (ground) living area square feet |  |
| BsmtFullBath | Basement full bathrooms |  |
| BsmtHalfBath | Basement half bathrooms |  |
| FullBath | Full bathrooms above grade |  |
| HalfBath | Half baths above grade |  |
| Bedroom | Bedrooms above grade (does NOT include basement bedrooms) |  |
| Kitchen | Kitchens above grade |  |
| KitchenQual | Kitchen quality | Ex (Excellent) Gd (Good) TA (Typical/Average) Fa (Fair) Po (Poor) |
| TotRmsAbvGrd | Total rooms above grade (does not include bathrooms) |  |
| Fireplaces | Number of fireplaces |  |
| FireplaceQu | Fireplace quality | Ex (Excellent- exceptional masonary fireplace) Gd (Good- masonry fireplace in main level) TA (Average- prefabricated fireplace in main living area or masonry fireplace in basement) Fa (Fair- prefabricated fireplace in basement) Po (Poor- Ben Franklin stove) NA (No fireplace) |
| GarageType | Garage location | 2Types (More than one type of garage) Attchd (Attached to home) Basment (Basement Garage) BuiltIn (Built-In Garage part of house- typically has room above garage) CarPort (Car Port) Detchd (Detached from home) NA (No Garage) |
| GarageYrBlt | Year garage was built |  |
| GarageFinish | Interior finish of the garage | Fin (Finished) RFn (Rough Finished) Unf (Unfinished) NA (No Garage) |
| GarageCars | Size of garage in car capacity |  |
| GarageArea | Size of garage in square feet |  |
| GarageCond | Garage condition | Ex (Excellent) Gd (Good) TA (Typical/Average) Fa (Fair) Po (Poor) NA (No Garage) |
| PavedDrive | Paved driveway | Y (Paved)  P (Partical Pavement)  N (Dirt/Gravel) |
| WoodDeckSF | Wood deck area in square feet |  |
| OpenPorchSF | Open porch area in square feet |  |
| EnclosedPorch | Enclosed porch area in square feet |  |
| Fence | Fence quality | GdPrv (Good Privacy) MnPrv (Minimum Privacy) GdWo (Good Wood) MnWw (Minimum Wood/Wire) NA (No Fence) |
| MoSold | Month Sold (MM) |  |
| YrSold | Year Sold (YYYY) |  |
| SaleType | Type of sale | WD (Warranty Deed – Conventional) CWD (Warranty Deed – Cash) VWD (Warranty Deed - VA Loan)  New (Home just constructed and sold) COD (Court Officer Deed/Estate) Con (Contract 15% Down payment regular terms) ConLw (Contract Low Down payment and low interest) ConLI (Contract Low Interest) ConLD (Contract Low Down) Oth (Other) |
| SaleCondition | Condition of sale | Normal (Normal Sale) Abnorml (Abnormal Sale- trade, foreclosure, short sale) AdjLand (Adjoining land purchase) Alloca (Allocation- two linked properties with separate deeds, typically condo with garage unit) Family (Sale between family members) Partial (Home was not completed when last assessed- associated with new homes) |

**Exploration & Modification**

Appendix B: Column Removal

Appendix B.1 Street

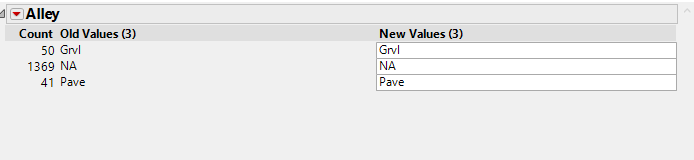
**Original:**



**Action**: Removed

Appendix B.2 Alley

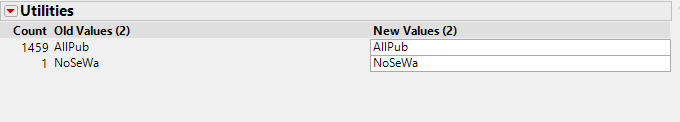
**Original:**



**Action**: Removed

Appendix B.3 *Utilities*

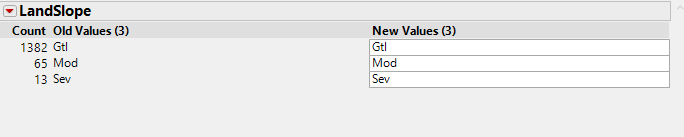
**Original:**



**Action**: Removed

Appendix B.4 *LandSlope*

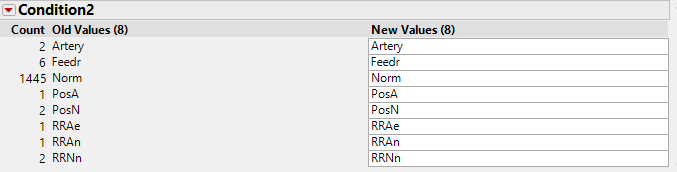
**Original:**



**Action**: Removed

Appendix B.5 *Condition2*

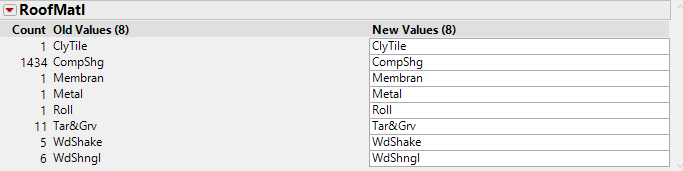
**Original:**



**Action**: Removed

Appendix B.6 *RoofMatl*

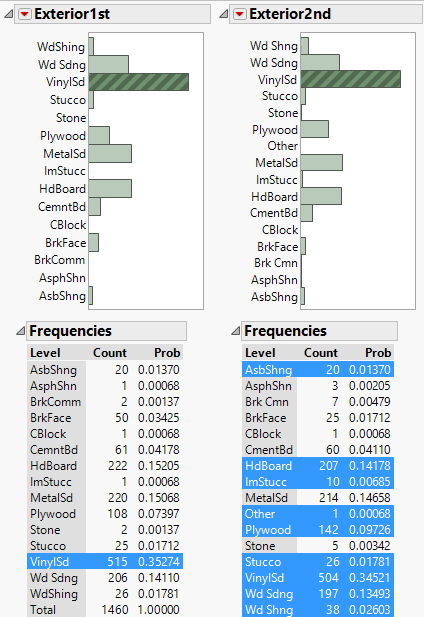
**Original:**



**Action**: Removed

Appendix B.7 *Exterior2*

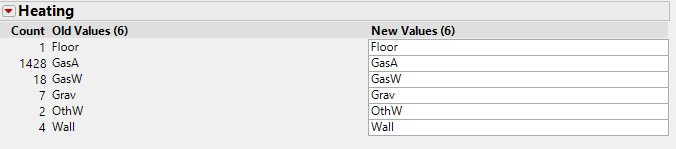
**Original:**



**Action**: Removed

Appendix B.8 *Heating Type*

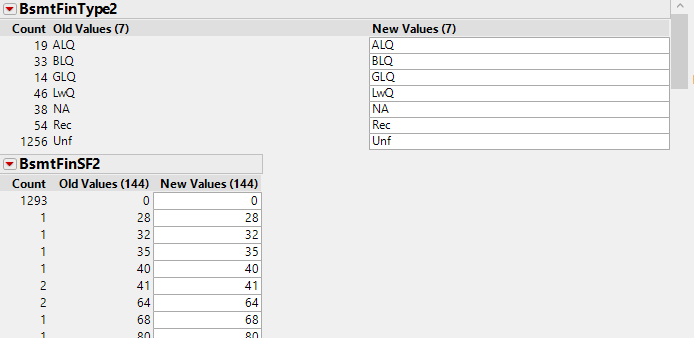
**Original:**



**Action**: Removed

Appendix B.9 *BsmrtFinType2* & *BsmntSF2*

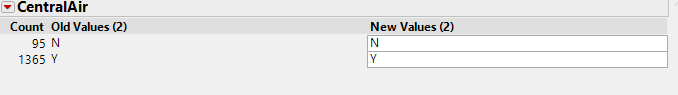
**Original:**



**Action**: Removed both BsmrtFinType2 & BsmntSF2

Appendix B.10 *CentralAir*

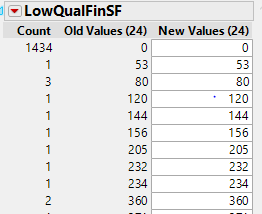
**Original:**



**Action**: Removed

Appendix B.11 *LowQualFinSF*

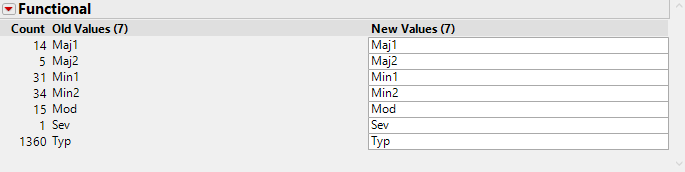
**Original:**



**Action**: Removed

Appendix B.12 *Functional*

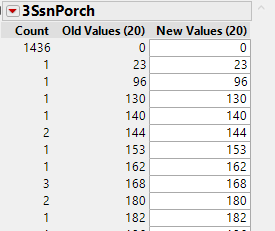
**Original:**



**Action**: Removed

Appendix B.13 *3SsnPorch*

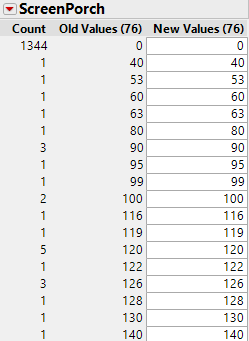
**Original:**



**Action**: Removed

Appendix B.13 *ScreenPorch*

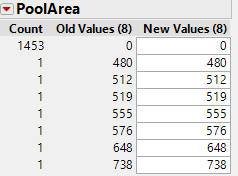
**Original:**



**Action**: Removed

Appendix B.13 *PoolArea*

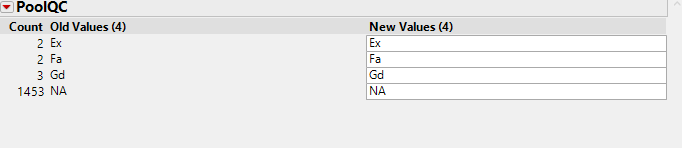
**Original:**



**Action**: Removed

Appendix B.14 *PoolQC*

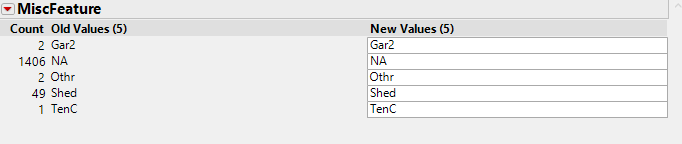
**Original:**



**Action**: Removed

Appendix B.15 *MiscFeature*

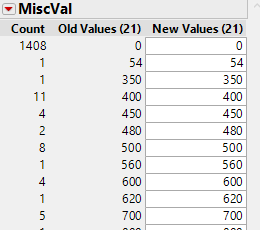
**Original:**



**Action**: Removed

Appendix B.16 *MiscValue*

**Original:**

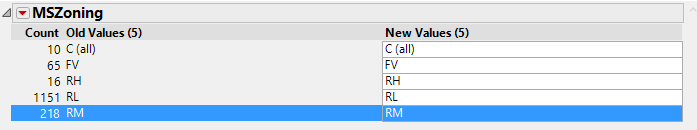


**Action**: Removed

Appendix C: Data Modification

Appendix C.1 *MSZoning*

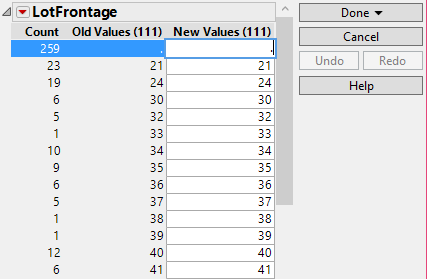
**Original:**



**Action**: Change C(all) to C

Appendix C.2 *LotFrontage*

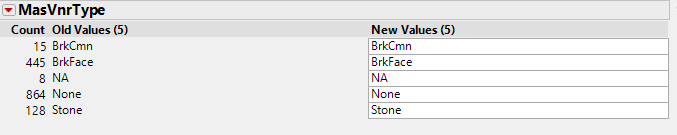
**Original:**



**Action**: Changing . to 0

Appendix C.3 *MasVnrType*

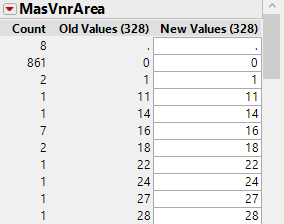
**Original:**



**Action**: Change NA to none

Appendix C.4 *MasVnrArea*

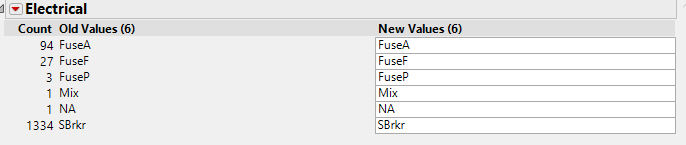
**Original:**



**Action**: Change ‘.’ To 0

Appendix C.5 *Electrical*

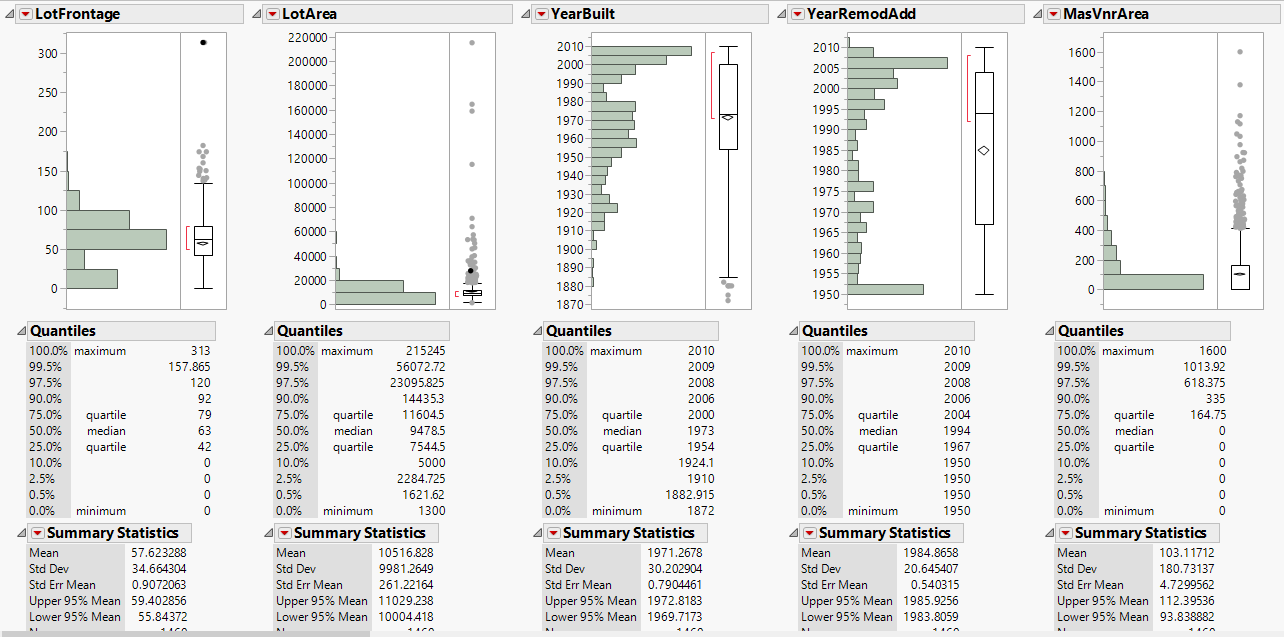
**Original:**

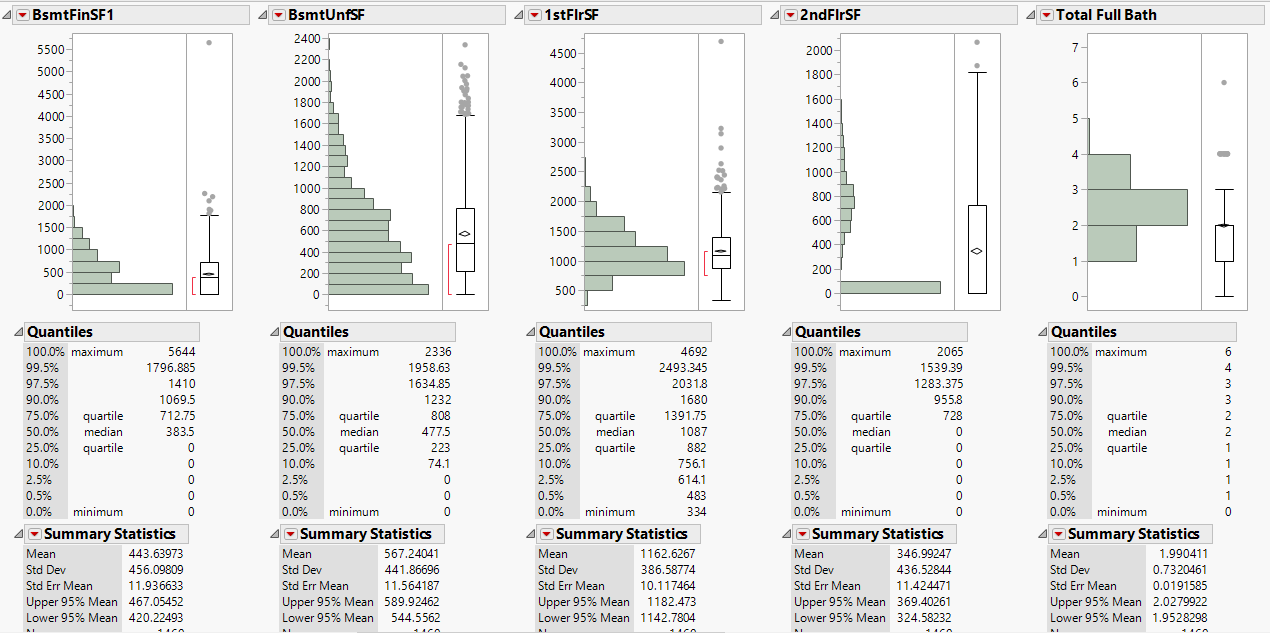


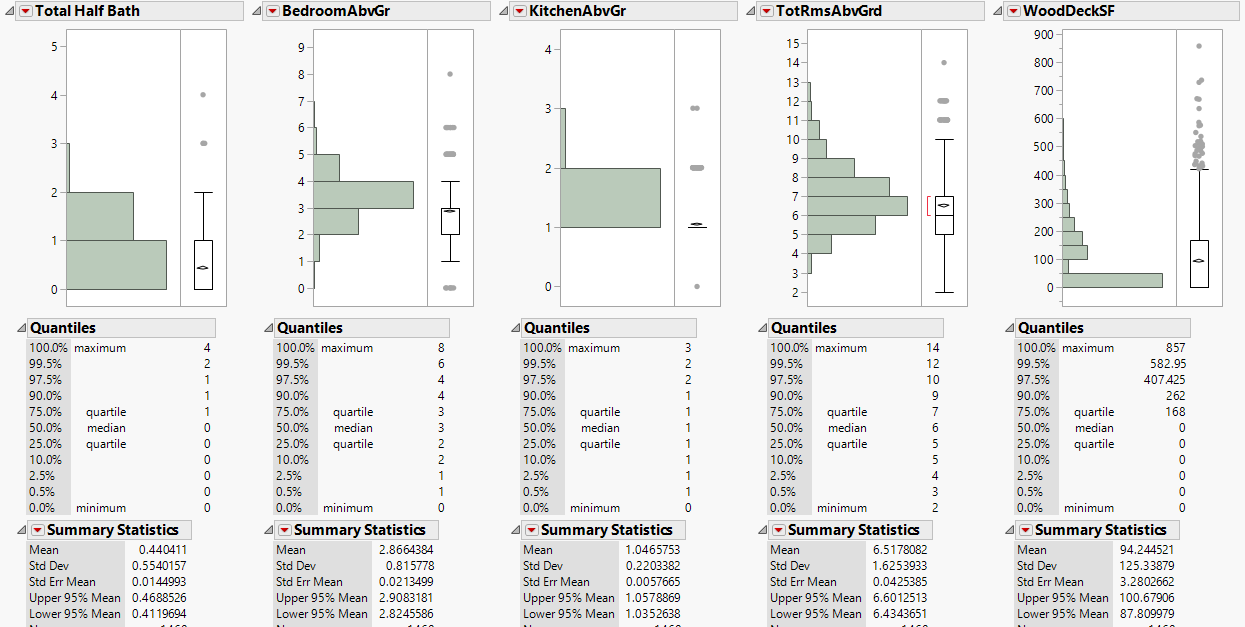
**Action**: Replace NA with mode ‘SBrkr’

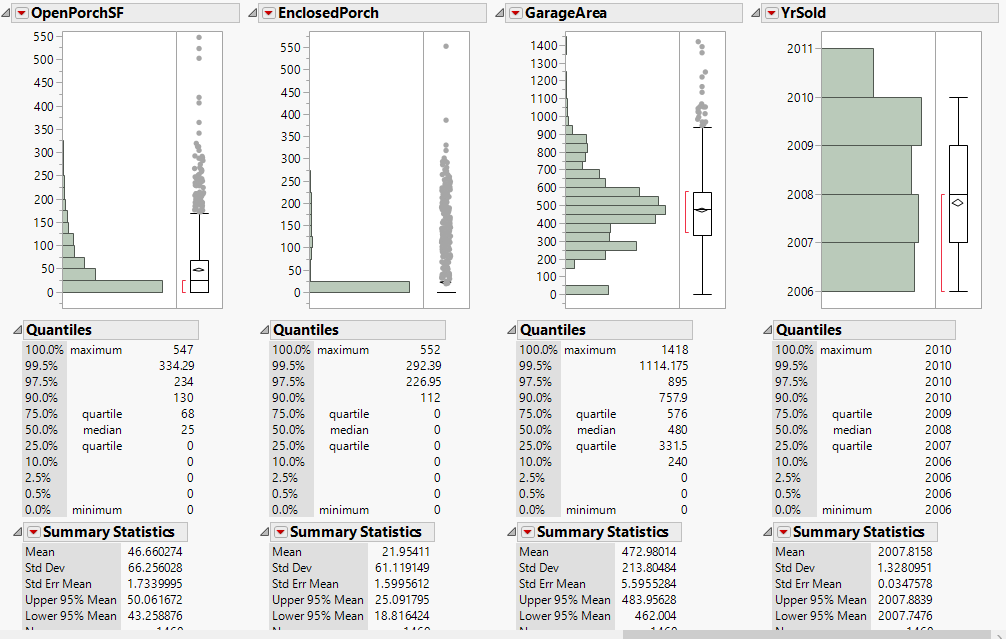
Appendix D: OUTLIER ANALYSIS

Appendix D.1: Univariate:

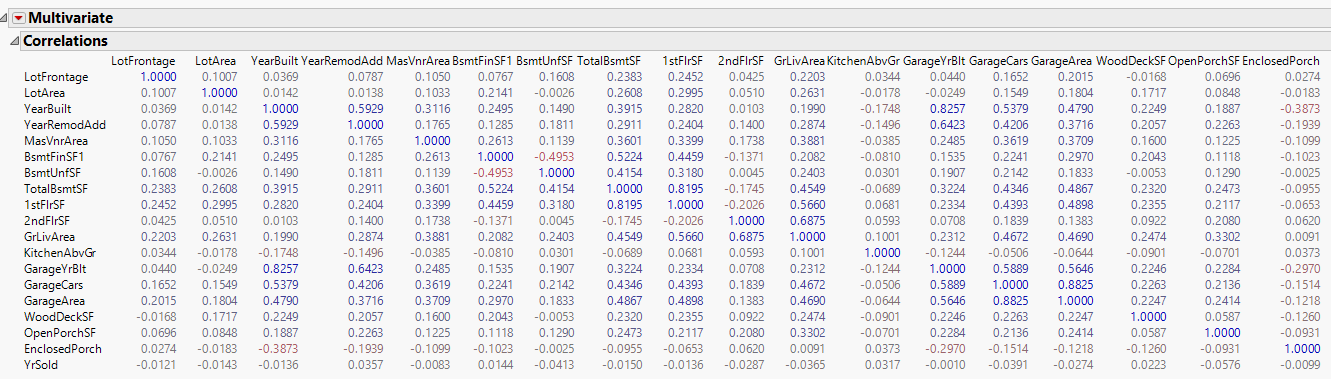


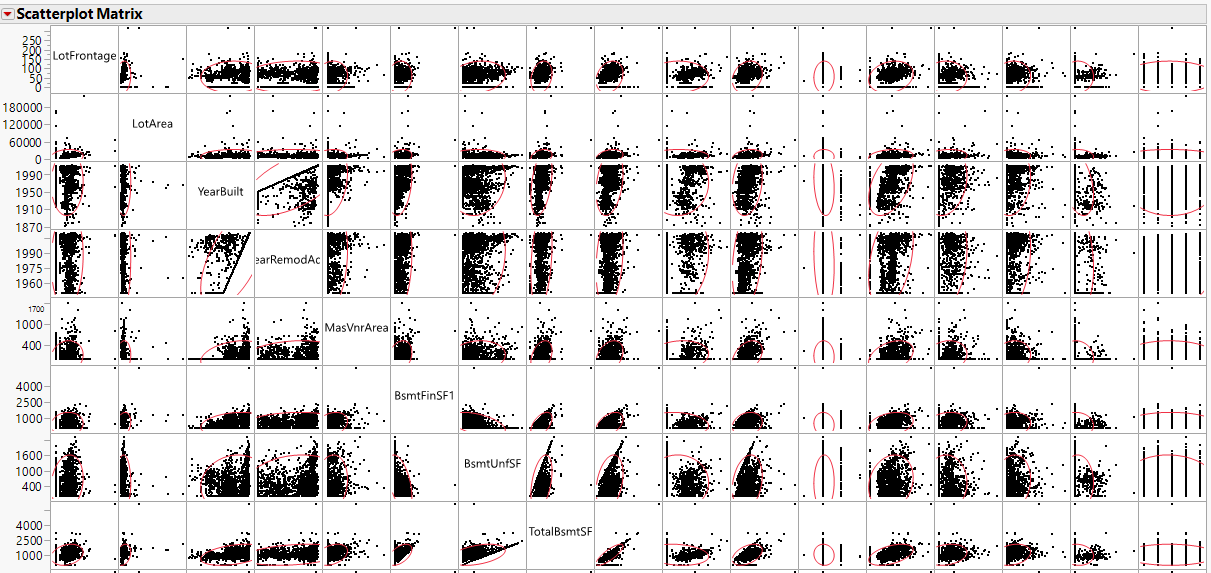






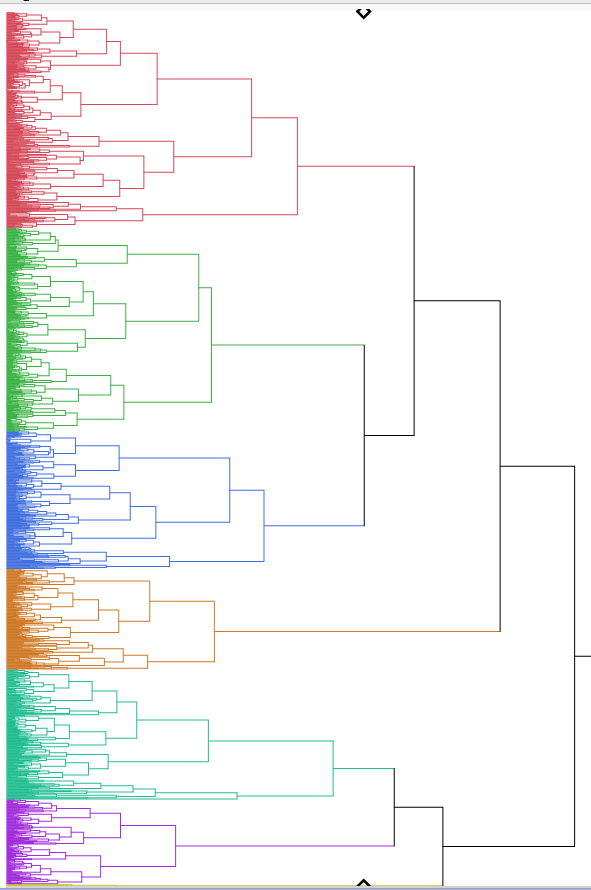
Appendix D.2: Bivariate:

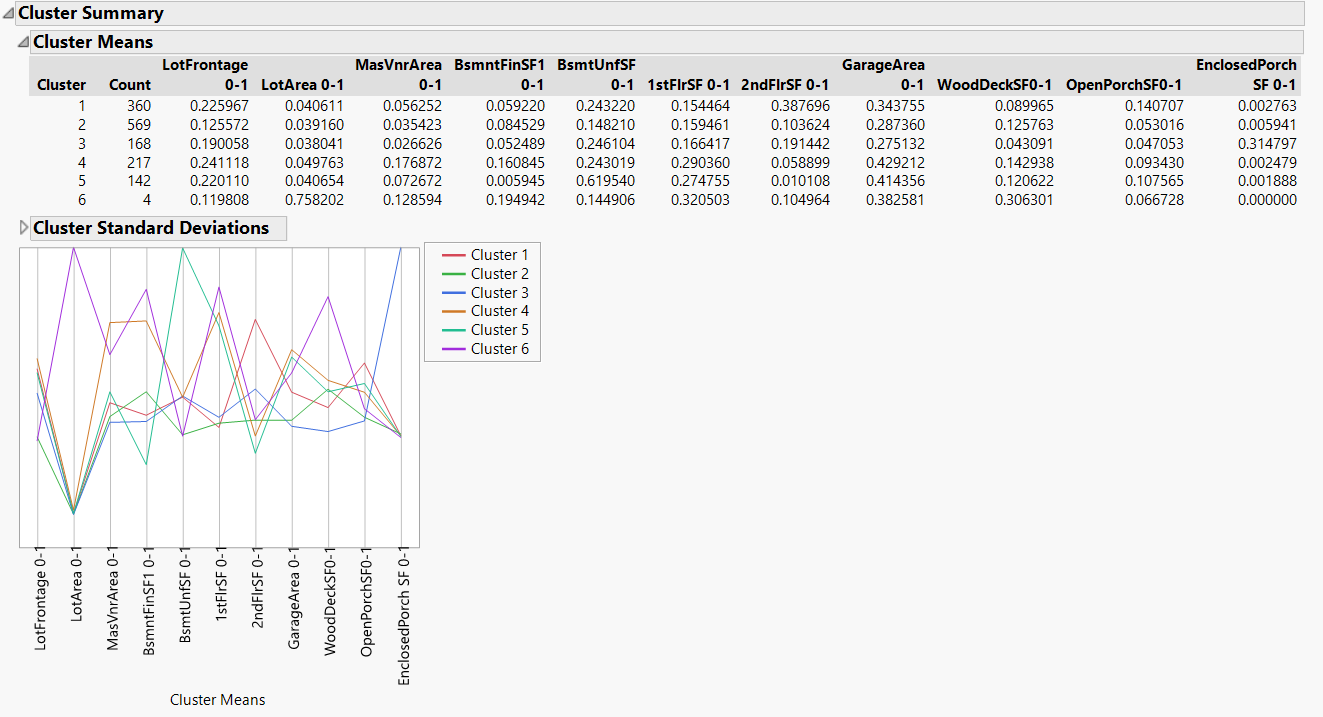




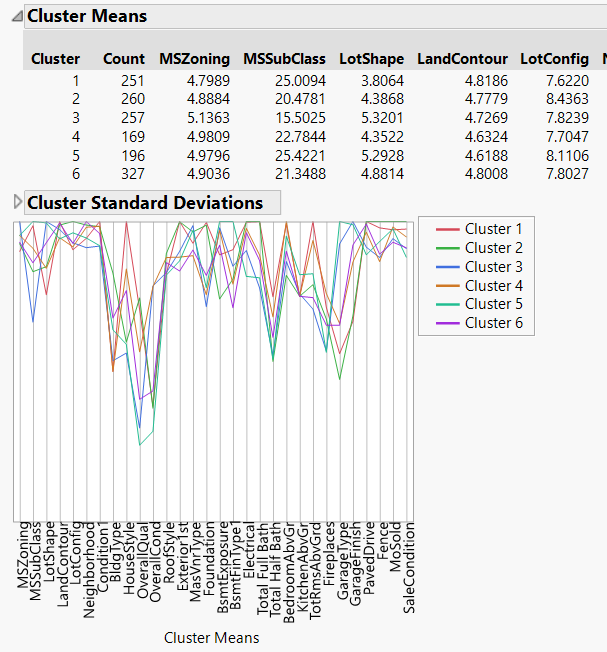
Appendix E: Clustering

Appendix E.1:Hierarchical

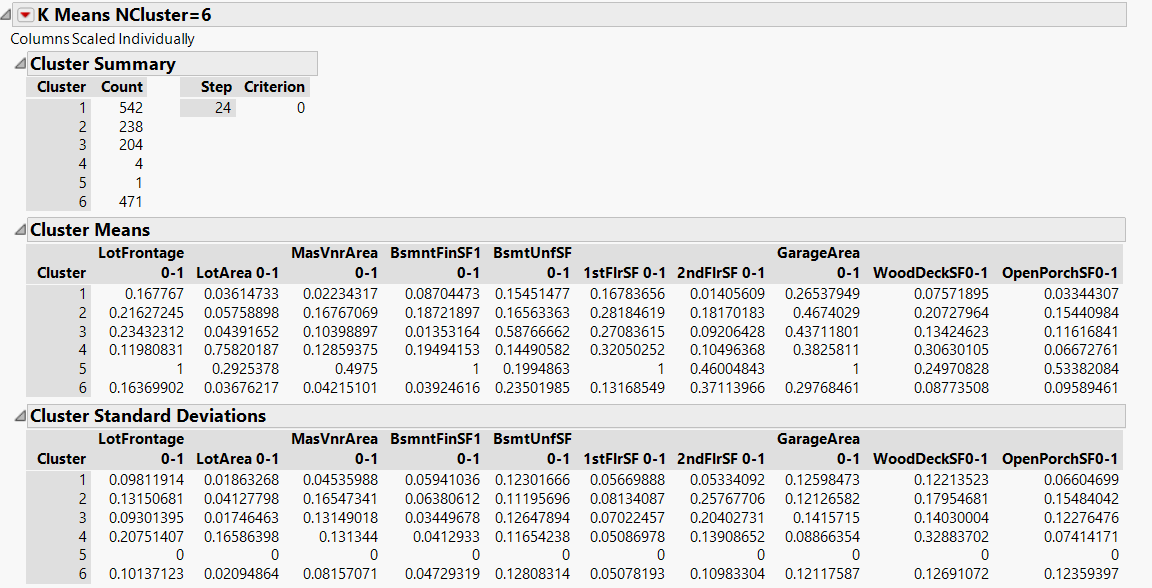


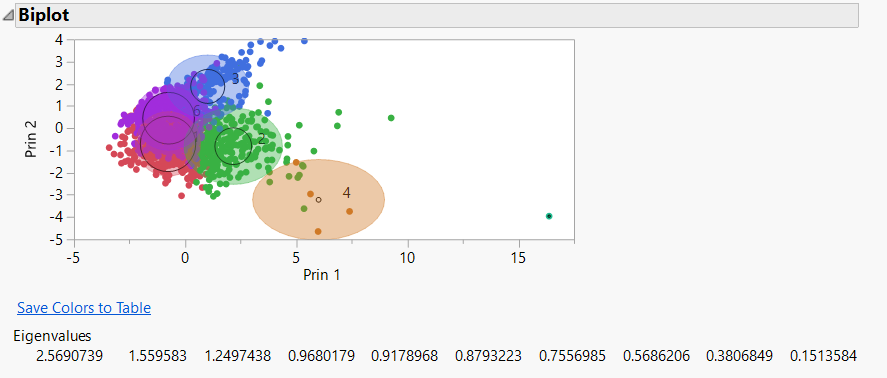


Appendix E.2:Nominal:



Appendix E.3:K-Means:



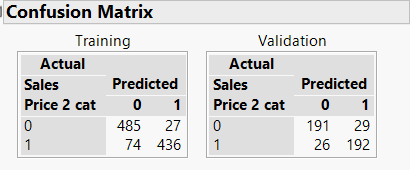


Appendix F: Modeling

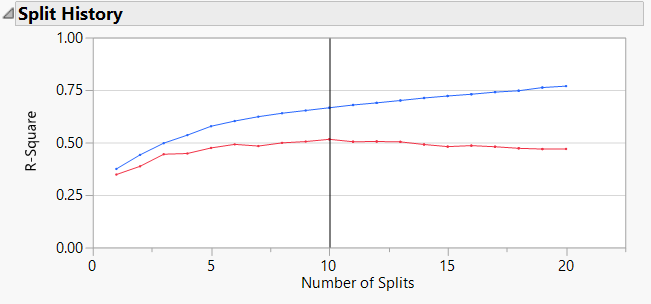
Predicting SalePrice2Cat

Appendix F.1.1: Decision Tree

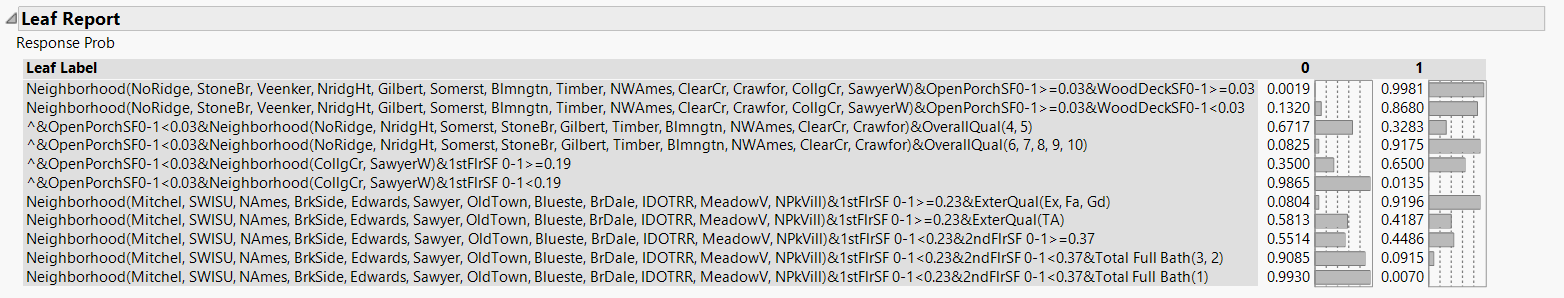
Confusion Matrix:



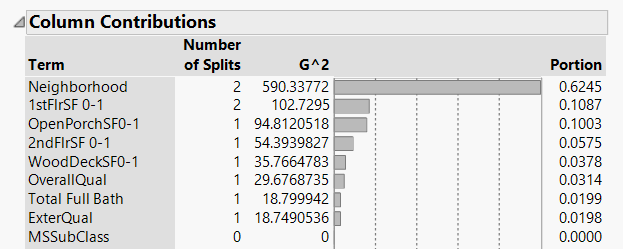
Split History:



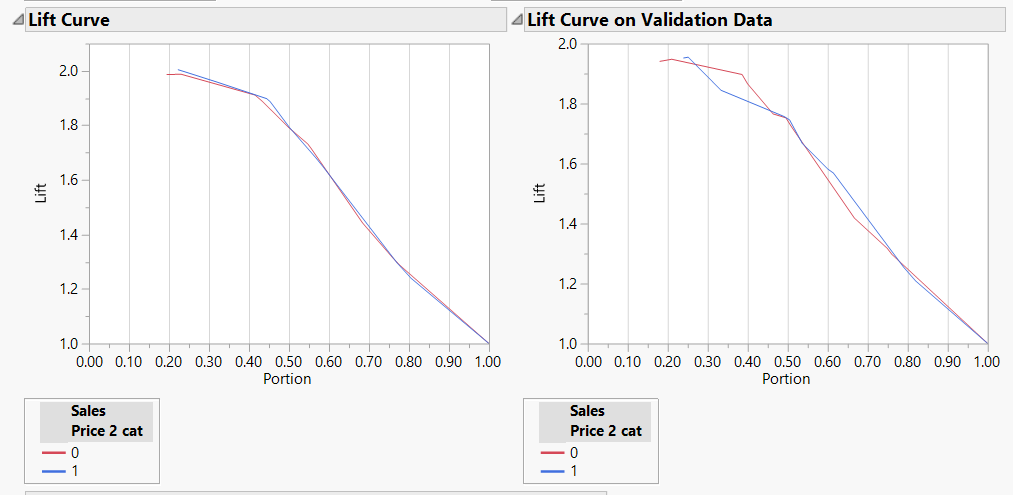
Leaf report:



Column Contributions:

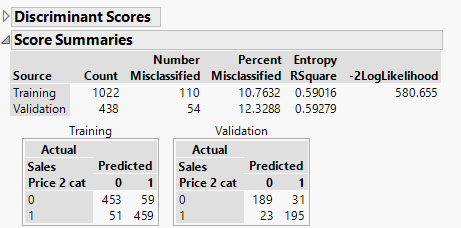


Lift curve:



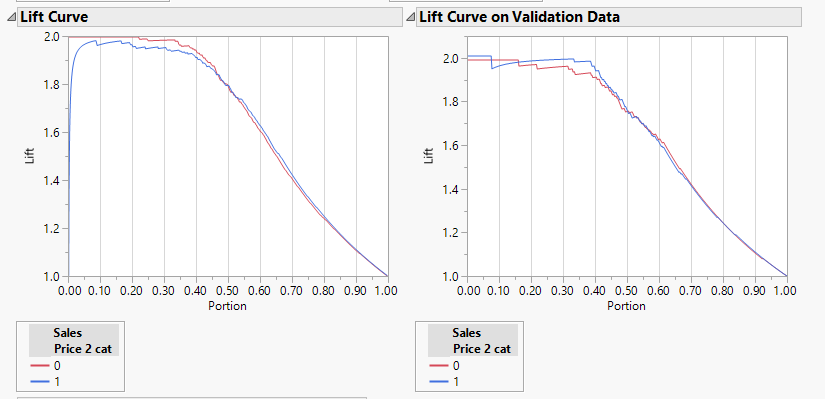
Appendix F.1.2: Linear Discriminant Analysis:

Confusion Matrix:

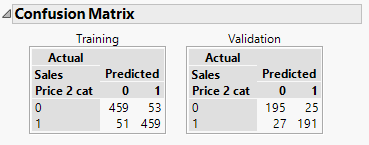


Appendix F.1.3: Logistic Regression:

Lift Curve:

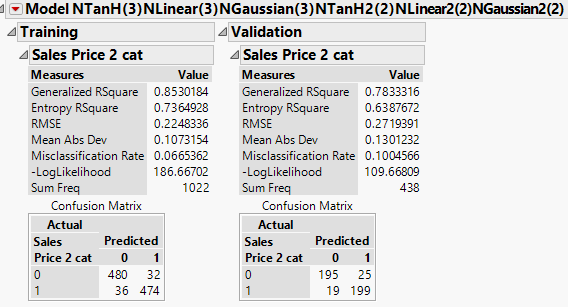


Confusion Matrix:



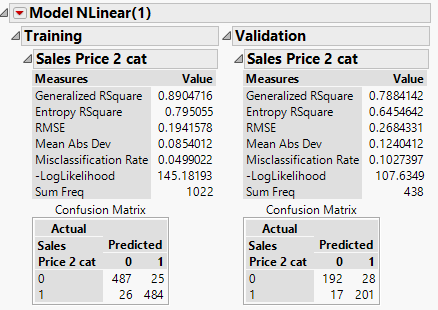
Appendix F.1.4: Neural Networks

Confusion Matrix:



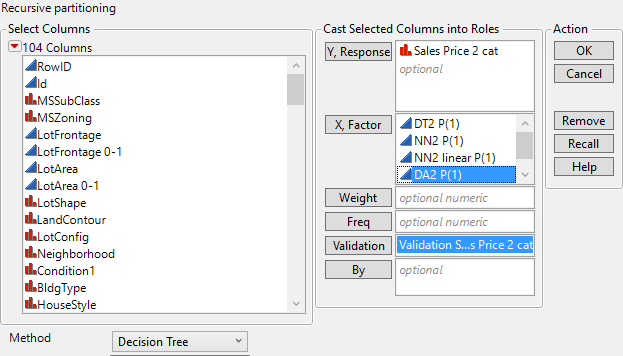
**Neural Networks Linear:**

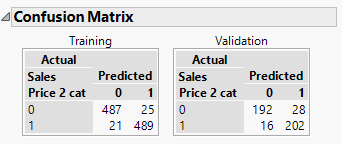
Confusion matrix:



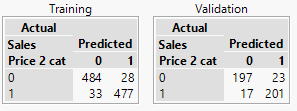
Appendix F.1.5: Ensembles

Appendix F.1.5.1: Ensembles: Decision Tree

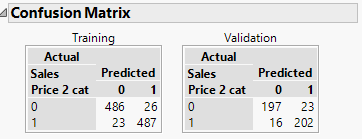




Appendix F.1.5.2: Ensembles: Linear Discriminant Analysis:



Appendix F.1.5.3: Ensembles: Logistic Regression:



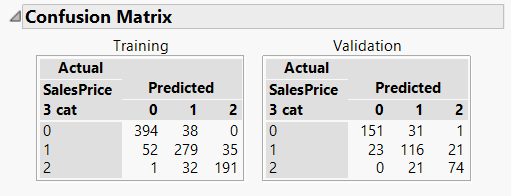
Appendix F.1.5.3: Ensembles: Neural Networks:



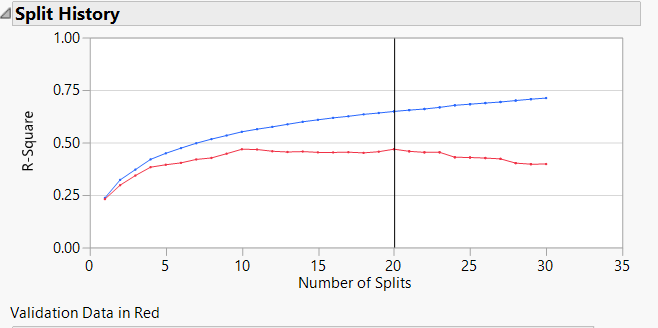
**Predicting *SalePrice3Cat***

Appendix F.2.1: Decision Tree:

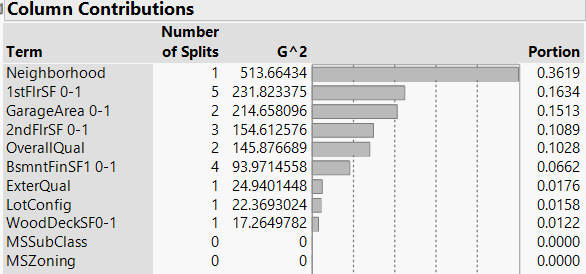
Confusion Matrix:



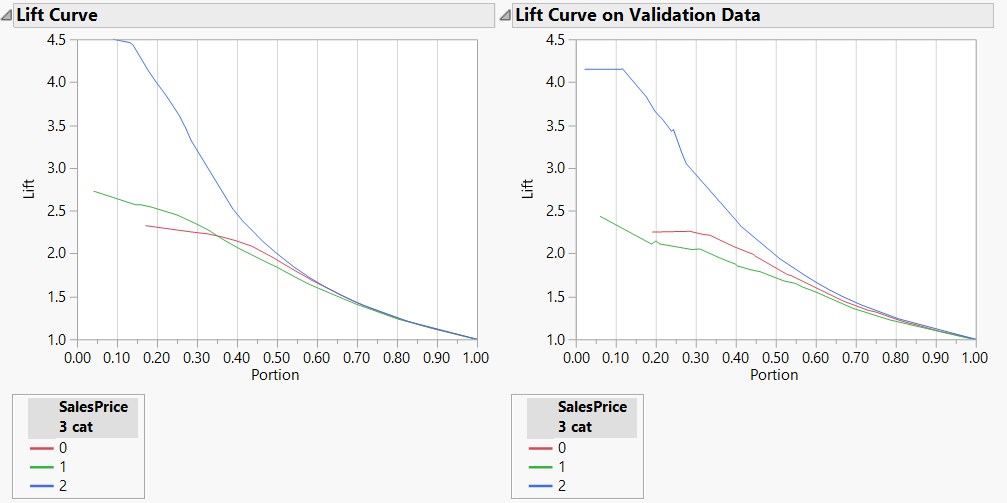
Split History:



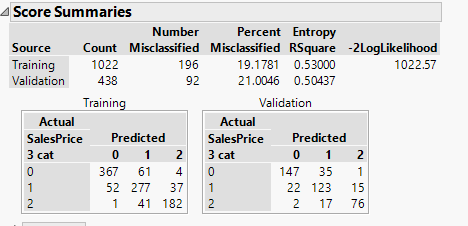
Column Contributions:



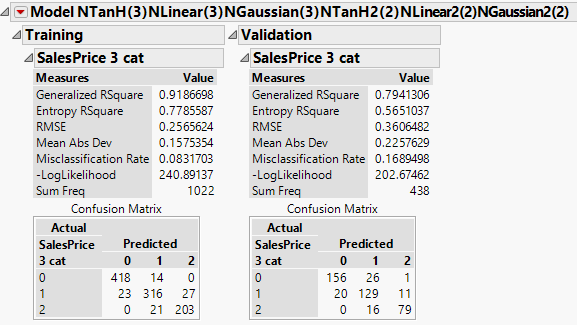
Lift Curve:



Appendix F.2.2: Linear Discriminant Analysis:

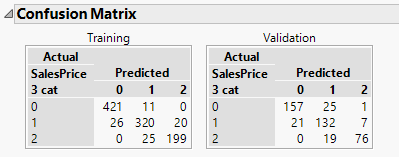


Appendix F.2.3: Neural Networks:

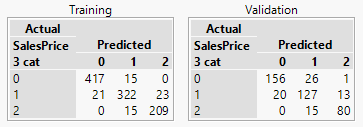


Appendix F.2.4: Ensemble

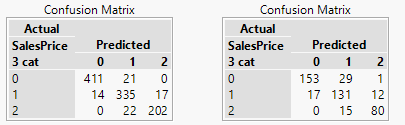
Appendix F.2.4.1: Ensemble Decision Tree:



Appendix F.2.4.2: Ensemble Linear Discriminant Analysis:

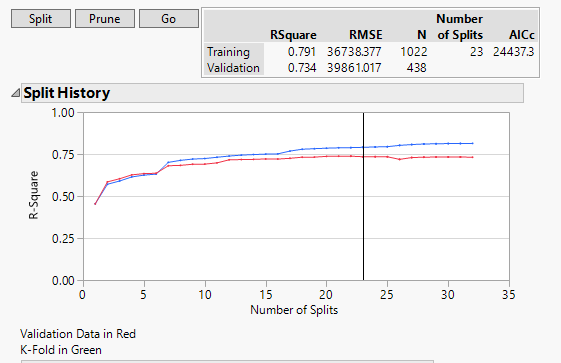


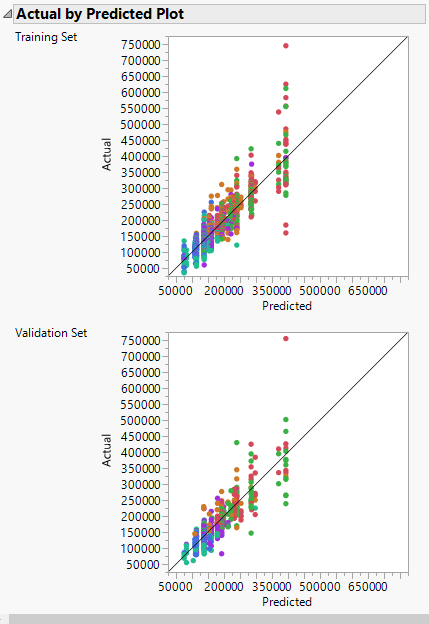
Appendix F.2.4.3: Ensemble Neural Networks:

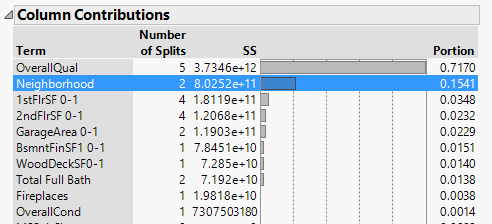


**Predicting *SalePrice* Continuous:**

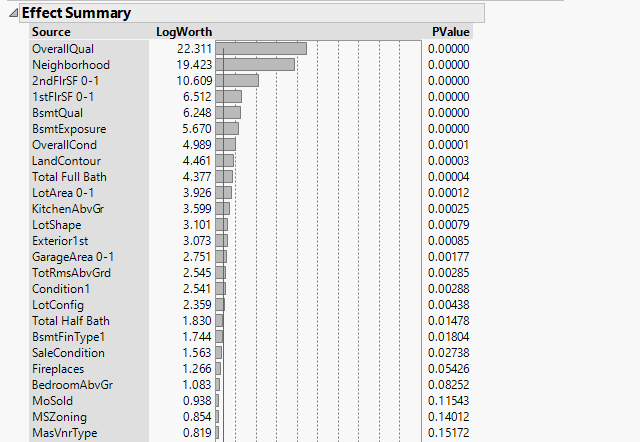
Appendix F.3.1: Decision Tree:

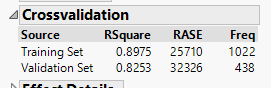






Appendix F.3.2: Linear Regression Least square

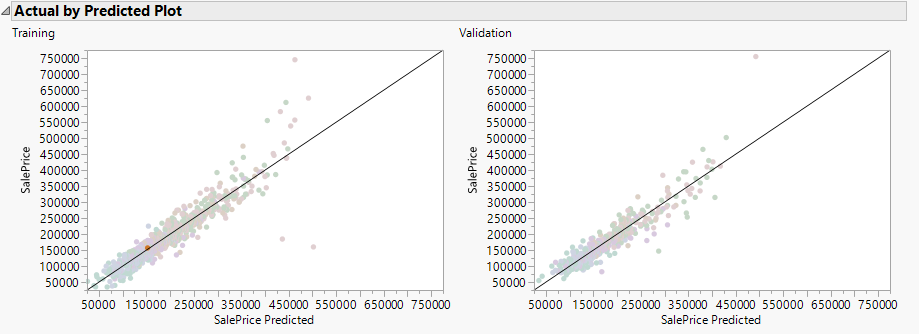
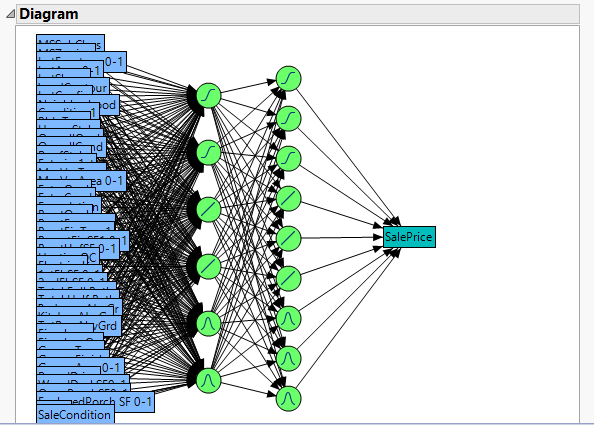


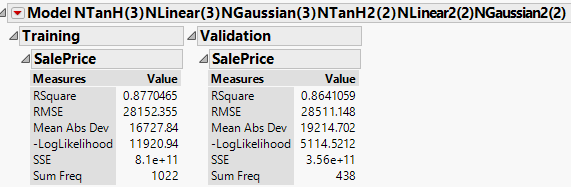


Appendix F.3.2.1: Linear Regression Stepwise



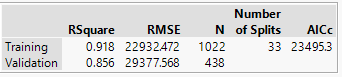
Appendix F.3.3: Neural Networks



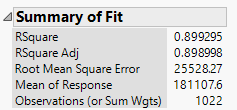


Appendix F.3.4: Ensemble

Appendix F.3.4.1: Ensemble Decision Tree



Appendix F.3.4.2: Ensemble Linear Regression Least Squares



Appendix F.3.4.2: Ensemble Neural Networks

