**Predicting Sales Price of Homes Using Multiple Linear Regression**

Yao Yao, Ian Kinskey, & Robert Gill

MSDS 6372 Experimental Statistics II

Section 403

Project 1

June 11, 2017

**Introduction**

When predicting the selling price of a home, intuition would suggest traditionally considered factors such as exterior appearance, square footage, number of bedrooms and bathrooms, and number of floors to be the best explanatory variables. However, many other features of a home may also be utilized to improve price prediction.

**Problem Statement**

Develop a regression model to predict final selling price of homes based on explanatory variables in the Ames Housing dataset.

**Constraints and Limitations**

The analysis was performed on Ames housing data that were collected from the Ames Assessor’s Office for individual residential properties sold in Ames, IA from 2006 to 2010. Because these data are observational in nature, no causal inferences can be made about the relationship between the explanatory variables and the response variable. These data constitute a census for the time period and city observed, and thus the scope of inference is limited to this data set alone. No claims based on this analysis should be made for other cities or time periods. Finally, it is possible that explanatory variables not included in these data may be important predictors of a home’s final selling price.

**Data Set Description**

The Ames Housing data compiled by Dean De Cock are utilized. These data describe the sale of individual residential properties in Ames, Iowa from 2006 to 2010. This data set contains 2,919 observations comprised of seventy-nine potential explanatory variables. The data set is equally partitioned into a training set and test set. The training set contains an additional variable SalePrice, which is the response variable for this analysis.

**Exploratory Data Analysis**

|  |  |
| --- | --- |
| **Figure 1**. Basic descriptive statistics of response variable SalePrice |  |
| **Figure 2**. Quantiles of response variable SalePrice |  |
| **Figure 3**. Histogram of response variable SalePrice |  |

The mean of the response variable, SalePrice, is $180,921 (Figure 1). The distribution of SalePrice is right skewed as evidenced by both the median, $163,000 (Figure 2), and the histogram of SalePrice (Figure 3). In practical terms, the spread of the data is quite large, as evidenced by the range of $720,100 (Figure 2) , and the standard deviation of $79,443 (Figure 2).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1**. Percentage of data populated and count of imputed observations by variable order by population percentage prior to imputation (train and test sets merged) | | | | | | | | | | | | | | |
| **Variable** | **% Populated** | **Type** | **Missing Imputed** |  | **Variable** | **% Populated** | **Type** | **Missing Imputed** |  | **Variable** | **% Populated** | **Type** | | **Missing Imputed** |
| BldgType | 100 | Categorical | 0 |  | TotRmsAbvGrd | 100 | Continuous | 0 |  | GarageFinish | 94.55 | Categorical | | 159 |
| CentralAir | 100 | Categorical | 0 |  | X1stFlrSF | 100 | Continuous | 0 |  | GarageQual | 94.55 | Categorical | | 159 |
| Condition1 | 100 | Categorical | 0 |  | YearBuilt | 100 | Continuous | 0 |  | GarageYrBlt | 94.55 | Continuous | | 159 |
| Condition2 | 100 | Categorical | 0 |  | YearRemodAdd | 100 | Continuous | 0 |  | BsmtUnfSF | 91.71 | Continuous | | 242 |
| ExterCond | 100 | Categorical | 0 |  | YrSold | 100 | Continuous | 0 |  | LotFrontage | 83.35 | Continuous | | 486 |
| ExterQual | 100 | Categorical | 0 |  | Electrical | 99.97 | Categorical | 1 |  | BsmtFinSF1 | 68.14 | Continuous | | 930 |
| Foundation | 100 | Categorical | 0 |  | Exterior1st | 99.97 | Categorical | 1 |  | OpenPorchSF | 55.53 | Continuous | | 1298 |
| GrLivArea | 100 | Continuous | 0 |  | Exterior2nd | 99.97 | Categorical | 1 |  | FireplaceQu | 51.35 | Categorical | | 1420 |
| Heating | 100 | Categorical | 0 |  | KitchenQual | 99.97 | Categorical | 1 |  | Fireplaces | 51.35 | Continuous | | 1420 |
| HeatingQC | 100 | Categorical | 0 |  | SaleType | 99.97 | Categorical | 1 |  | WoodDeckSF | 47.82 | Continuous | | 1523 |
| HouseStyle | 100 | Categorical | 0 |  | Functional | 99.93 | Categorical | 2 |  | X2ndFlrSF | 42.86 | Continuous | | 1668 |
| LandContour | 100 | Categorical | 0 |  | Utilities | 99.93 | Categorical | 2 |  | BsmtFullBath | 41.52 | Continuous | | 1707 |
| LandSlope | 100 | Categorical | 0 |  | KitchenAbvGr | 99.9 | Continuous | 3 |  | MasVnrArea | 39.67 | Continuous | | 1761 |
| LotArea | 100 | Continuous | 0 |  | MSZoning | 99.86 | Categorical | 4 |  | HalfBath | 37.17 | Continuous | | 1834 |
| LotConfig | 100 | Categorical | 0 |  | BedroomAbvGr | 99.73 | Continuous | 8 |  | Fence | 19.56 | Categorical | | 2348 |
| LotShape | 100 | Categorical | 0 |  | FullBath | 99.59 | Continuous | 12 |  | EnclosedPorch | 15.72 | Continuous | | 2460 |
| MoSold | 100 | Continuous | 0 |  | MasVnrType | 99.18 | Categorical | 24 |  | BsmtFinSF2 | 11.89 | Continuous | | 2572 |
| MSSubClass | 100 | Continuous | 0 |  | BsmtFinType1 | 97.29 | Categorical | 79 |  | ScreenPorch | 8.77 | Continuous | | 2663 |
| Neighborhood | 100 | Categorical | 0 |  | TotalBsmtSF | 97.29 | Continuous | 79 |  | Alley | 6.78 | Categorical | | 2721 |
| OverallCond | 100 | Continuous | 0 |  | BsmtFinType2 | 97.26 | Categorical | 80 |  | BsmtHalfBath | 6 | Continuous | | 2744 |
| OverallQual | 100 | Continuous | 0 |  | BsmtQual | 97.23 | Categorical | 81 |  | MiscFeature | 3.6 | Categorical | | 2814 |
| PavedDrive | 100 | Categorical | 0 |  | BsmtCond | 97.19 | Categorical | 82 |  | MiscVal | 3.53 | Continuous | | 2816 |
| RoofMatl | 100 | Categorical | 0 |  | BsmtExposure | 97.19 | Categorical | 82 |  | LowQualFinSF | 1.37 | Continuous | | 2879 |
| RoofStyle | 100 | Categorical | 0 |  | GarageType | 94.62 | Categorical | 157 |  | X3SsnPorch | 1.27 | Continuous | | 2882 |
| SaleCondition | 100 | Categorical | 0 |  | GarageArea | 94.59 | Continuous | 158 |  | PoolArea | 0.45 | Continuous | | 2906 |
| SalePrice | 100 | Continuous | 0 |  | GarageCars | 94.59 | Continuous | 158 |  | PoolQC | 0.34 | Categorical | | 2909 |
| Street | 100 | Categorical | 0 |  | GarageCond | 94.55 | Categorical | 159 |  |  | | |  | |  |  |

Not all observations have complete data. Table 1 shows the population percentage for each variable. Because not all observations have values for all variables, imputation is necessary.

**Imputation by Bootstrap**

To accommodate for all the missing values, bootstrap was used to generate values based on available data. For continuous variables, bootstrap generated values from around the mean and standard deviation while preserving skewness of the existing sample size. For categorical variables, some NA values are considered its own category, based on the data description, where a change in parameter name completed the data column. Otherwise, we converted the categories into numerical classes and used bootstrap to generate integer categories based on available frequency of the existing sample size.

Advanced methods of multi-relational imputation, where missing values are generated based on co-linearity between columns, could be considered, but this recursive process of creating missing values could extend to more bias towards the final multi-linear model due to false correlations instead of a random cloud in the diagnostic plots. Based on MSE, advanced imputation techniques by generating missing values based on high correlations between columns lead to lesser predictive models in comparison to simple imputation based on frequency of single-column data, where all missing values are generated using R (<http://gradworks.umi.com/36/34/3634075.html>).

False positive bias could be generated from imputed columns where a majority of the data is missing and generated from the minority of the sample available (Table 1). Therefore, it is strategic to eliminate false trends by eliminating columns based on original insufficient data prior to the imputation by not having those specific columns in the selection process for the multi-linear modeling. Arbitrarily, we chose 50% and 75% original data as the cut-off point for using auto-selection on columns that were completed by imputation as well as the standard full table with all columns completed by imputation.

**Log Transformation**

For multiple linear regression and high dimensional modeling, log transformations are necessary to reduce the skewness of continuous data when plotted against each other. Although normalization of the data is not an assumption made in multi-linear regression, better trend lines could be made from a random cloud with data points that occupies the full plot rather than a cluster bias with outliers and high leverage points. Overall, relationships between all continuous data show adequate distributions to fit a better model (Figure 4).

Some continuous data look categorical when plotted against each other such as overall quality and overall condition, while overall quality and lot area looks right skewed (Figure 4a). For there to be less discrepancy on the skewness of one category over another when plotted against another category, it is strategic to log all continuous data so that overall skewness among different categories are lowered.

|  |  |
| --- | --- |
| Before log transformation | After log transformation |
|  |  |
|  |  |

**Figure 4a**: Matrix plots of continuous variables before and after log transformation comparison

Some categories are co-linear with each other, such as year built and year remodeled (Figure 4a) because it is realistically impossible to remodel a house prior to the house being built. Basement finish type and first floor square feet have a "fanning out" data trend that could be easily be solved with log-log plots (Figure 4a).

|  |  |
| --- | --- |
|  |  |
|  |  |

**Figure 4b**: Matrix plots of continuous variables before and after log transformation comparison

The number of full baths in relation to half baths is numerical although it could be numerical (Figure 4b). Due to the relationship of full baths increasing or decreasing based on the other variable or overall house quality, we left them as a numerical value (Figure 4c).

|  |  |
| --- | --- |
|  |  |
|  |  |

**Figure 4c**: Matrix plots of continuous variables before and after log transformation comparison

For miscellaneous categories such as pool area versus misc val that originally had a very small data sample to begin with, we can see that the imputation of small data sets leads to unsuccessful data clusters or lines that have a "L" shape that are not correctable even after log transformation (Figure 4d). Due to the inability to create a random cloud in the matrix plots for continuous data, it is strategic to have a cut-off point for using auto-selection on columns where the majority of the missing values were completed by imputation.

|  |  |
| --- | --- |
|  |  |

**Figure 4d**: Matrix plots of continuous variables before and after log transformation comparison

**Results Based on Auto Selection**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 2**. Model selection methods and results by multiple selection metrics (Kaggle score, CV Press, and adjusted R^2) by multiple imputation percentage cutoffs  IMP50 – Variables for which 50% or more of values were populated were made available as candidates for inclusion to model  IMP75 - Variables for which 75% or more of values were populated were made available as candidates for inclusion to model  TOTAL – All variables were made available as candidates for inclusion to model, regardless of the number of imputed observations | | | | | | | | | |
| **Selection Method** | **External Kaggle Score** | | | **Internal CV Press** | | | **Internal Adjusted R^2** | | |
|  | **TOTAL** | **IMP50** | **IMP75** | **TOTAL** | **IMP50** | **IMP75** | **TOTAL** | **IMP50** | **IMP75** |
| **Backward** | 0.11520 | 0.11222 | 0.11254 | 46.22 | 47.26 | 47.16 | 0.929 | 0.926 | 0.925 |
| **Backward w/ Cross Validation** | 0.10572 | 0.10645 | 0.10763 | 22.74 | 23.15 | 23.56 | 0.930 | 0.922 | 0.922 |
| **Forward** | 0.11822 | 0.11646 | 0.11939 | 53.35 | 52.74 | 52.96 | 0.917 | 0.916 | 0.916 |
| **Forward w/ Cross Validation** | 0.11373 | 0.11503 | 0.11493 | 25.47 | 25.72 | 25.80 | 0.922 | 0.915 | 0.913 |
| **Stepwise** | 0.12109 | 0.11945 | 0.12070 | 46.22 | 47.26 | 47.16 | 0.917 | 0.916 | 0.916 |
| **Stepwise w/ Cross Validation** | 0.11373 | 0.11503 | 0.11493 | 23.25 | 24.07 | 24.36 | 0.913 | 0.911 | 0.909 |
| **LARS** | 0.15755 | 0.15656 | 0.15143 | 46.42 | 47.15 | 65.74 | 0.844 | 0.846 | 0.829 |
| **LARS w/ Cross Validation** | 0.12977 | 0.13059 | 0.16696 | 28.18 | 28.68 | 28.04 | 0.893 | 0.892 | 0.854 |
| **LASSO** | 0.15755 | 0.15686 | 0.15143 | 46.42 | 47.15 | 65.74 | 0.844 | 0.846 | 0.829 |
| **LASSO w/ Cross Validation** | 0.12977 | 0.13059 | 0.16696 | 28.18 | 28.68 | 28.04 | 0.893 | 0.892 | 0.854 |

Classes of categorical and continuous numerical values are eliminated using GLMSELECT with forward selection, backward elimination, stepwise selection, LASSO, and LARS methods via SAS, where the stopping point is when AIC is minimized. For cross validation splits, we used 20% CV, 30% test, and 50% train to develop the model where AIC is minimized. These cross validation splits performed well, but it may be possible to further optimize model selection by varying these splits. Other potential gains may be possible by employing different selection criteria, such as AICC, SBC, adjusted R^2, or average square error.

Imputation thresholds were also compared (Table 2). For “IMP75”, 60 explanatory variables were made available for selection. For “IMP50”, 64 variables explanatory variables were made available for selection. For “TOTAL”, all explanatory variables were made available for selection.

Methods employing cross validation tended to perform better as the model were able to self-check within its training data before making predictions against the test data set, thus avoiding over fitting. Backward elimination with cross validation splits resulted in the best model across all selection criterion and imputation thresholds. This selection method using all imputed data resulted in a Kaggle score (root mean square logarithmic error) of 0.10572, an internal CV PRESS of 22.74, and an adjusted R^2 of 0.93.

**Analysis of Best Auto Selection Model: Backwards with CV**

**Assumptions: Idealism vs Predictive Application**

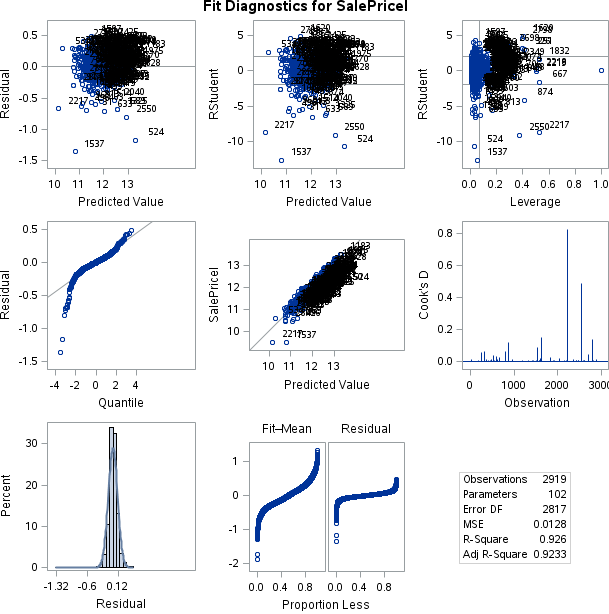
Normality: The intentions of creating a predictive model is based on accuracy of the mean square error of the test set sales price on kaggle. Therefore, the skewness of the QQ-plot and histogram are not vital issues when evaluating the best model (Figure 5).

Constant Variance: The residual plot and the studentized residual plot show random cloud about the zero line, providing evidence for constant variance. Logging all continuous variables helped with better constant variance with fewer outliers (Figure 5).

Linearity: In the Predicted vs. Actual Values plot, the observations create a thick linear line of good predictability (Figure 5).

Independence: Although some variables explored have colinearity, the auto-selection process removes variables that are irrelevant to the prediction model. Due to practical applications of predictability, there could be high VIFs and co-linear variables in the model (Figure 5).

For high Cook's D, high leverage points, and outliers, the application of better predictions could be made when all the data points in the model are present rather than removed (Figure 5).



**Figure 5**: Diagnostic plots of the best auto selection model

Even though the diagnostic predictive model is less than ideal, the application of real world prediction makes some assumptions and restrictions moot when obtaining better prediction values towards a lower Kaggle score.

**Fit Criteria for Backwards Selection with CV**

Backwards selection with CV of the complete table only removed 3 variables: Exterior1st, SaleType, and HouseStyle (Figure 6), which means 12 variables that had 25% or less original data prior to imputation were included: Fence, EnclosedPorch, BsmtFinSF2, ScreenPorch, Alley, BsmtHalfBath, MiscFeature, MiscVal, LowQualFinSF, X3SsnPorch, PoolArea, and PoolQC (Table 1).

What was more surpising as we later learn is that of the categorical variables that had 25% or less original data, MiscFeature 2, MiscFeature 3, and MiscFeature 4 had p-values that were significant to the data (Table 3). Even though the end goal is to get the best prediction value based on kaggle score, the imputation of missing values that had 25% or less original data worked too well to have the data be significant towards the final prediction model.

For the best model based on AIC: AICC, SBC, adjusted R^2, and CV press were at their optimal level while Validation ASE was optimal since the beginning. The adjusted R^2 was 0.93 with a CV press of 25.47 (Figure 6).

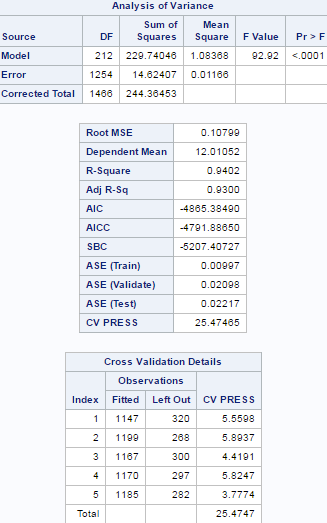
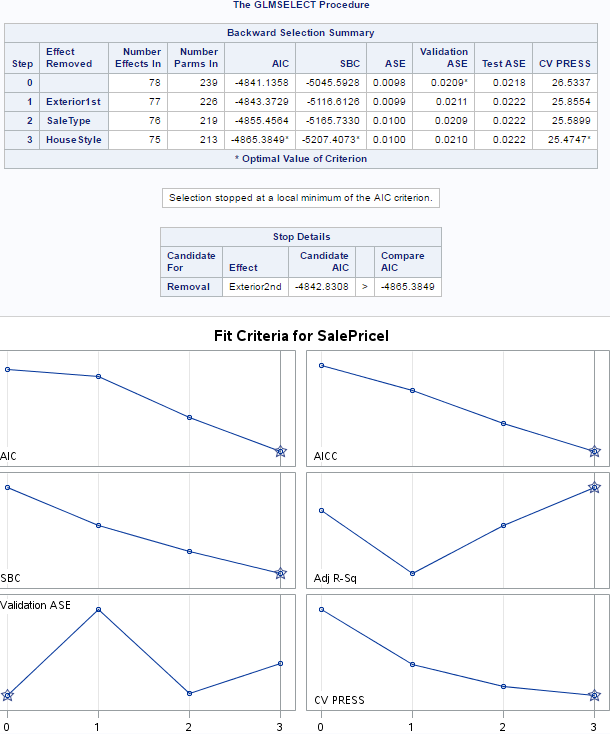


Figure 6: Details of backwards selection with cross validation

|  |
| --- |
| **Table 3**: Kaggle results of custom model based on best backwards selection with CV |

|  |  |
| --- | --- |
| **Diagnostic Conversion** | **Model from** |
| **Best GLMSELECT** | **BACKWARDCV** |
| **Sig P-val only (vif ≈ 1)** | 0.11890 |
| **CV only (1 train:1 CV)** | 0.11373 |
| **REG Conversion** | 0.11120 |
| **GLM Conversion** | 0.10677 |

When creating dummy variables to convert the best PROC GLMSELECT backwards selection model into PROC REG, the resulting model had a worse Kaggle score from 0.10572 to 0.11120 (Table 2 and 3). Since 1's and 0's were used for categorical values, logits were used for PROC REG to differentiate the black and white numerical values. Unlike PROC GLM that accommodates for grey-scale categorical parameters within a variable, the black and white distinction of 0's and 1's lost some predictability after the model conversion.

Upon exploring increase of a different CV split for the best backwards selection, the model's accuracy also went down. When we truncated the model to only include significant categorical parameters within a variable instead of all categorical parameters within a model, the prediction lost accuracy in the data (Table 4). Even the GLMSELECT to GLM conversion lost relevant predictability because 282 observations are left out when fitted in the CV prediction model (Figure 6).

Therefore, it is only strategic to include the full model in its original PROC GLMSELECT format to create the best model for better accuracy regardless of imputed data from an original data sample of 25% or less.

|  |
| --- |
| **Table 4**: The parameter estimates of the backwards selection model with cross validation in PROC GLM conversion. The highlighted values are the significant terms |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | |
| **Parameter** | **Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | -33.768815 | 4.39349891 | -7.69 | <.0001 |
| **MSZoning 3** | 0.01516702 | 0.02343851 | 0.65 | 0.5176 |
| **MSZoning 5** | 0.0229041 | 0.0257293 | 0.89 | 0.3734 |
| **MSZoning 6** | 0.01291615 | 0.01145898 | 1.13 | 0.2598 |
| **Street Grvl** | -0.1348395 | 0.03644049 | -3.7 | 0.0002 |
| **Alley 1** | 0.00246791 | 0.00436598 | 0.57 | 0.5719 |
| **LotShape IR1** | -0.0068843 | 0.00531136 | -1.3 | 0.195 |
| **LotShape IR2** | 0.0038619 | 0.01457937 | 0.26 | 0.7911 |
| **LotShape IR3** | -0.0043579 | 0.03029488 | -0.14 | 0.8856 |
| **LandContour Bnk** | -0.0307155 | 0.01197958 | -2.56 | 0.0104 |
| **LandContour HLS** | 0.0065481 | 0.01288898 | 0.51 | 0.6115 |
| **LandContour Low** | -0.0318795 | 0.01893724 | -1.68 | 0.0924 |
| **Utilities 1** | 0.1283668 | 0.11607644 | 1.11 | 0.2689 |
| **LotConfig Corner** | 0.00335215 | 0.00588709 | 0.57 | 0.5691 |
| **LotConfig CulDSac** | 0.00901222 | 0.010117 | 0.89 | 0.3731 |
| **LotConfig FR2** | -0.0401491 | 0.01308361 | -3.07 | 0.0022 |
| **LotConfig FR3** | -0.0275982 | 0.03095046 | -0.89 | 0.3726 |
| **LandSlope Gtl** | 0.02225109 | 0.03533362 | 0.63 | 0.5289 |
| **LandSlope Mod** | 0.03635623 | 0.03541788 | 1.03 | 0.3048 |
| **Neighborhood Blmngtn** | 0.03086582 | 0.0359216 | 0.86 | 0.3903 |
| **Neighborhood Blueste** | 0.04523956 | 0.04706404 | 0.96 | 0.3365 |
| **Neighborhood BrDale** | 0.0183319 | 0.03863358 | 0.47 | 0.6352 |
| **Neighborhood BrkSide** | 0.01986321 | 0.03053227 | 0.65 | 0.5154 |
| **Neighborhood ClearCr** | 0.02350655 | 0.03130661 | 0.75 | 0.4528 |
| **Neighborhood CollgCr** | -0.011161 | 0.02637026 | -0.42 | 0.6722 |
| **Neighborhood Crawfor** | 0.08898677 | 0.02835451 | 3.14 | 0.0017 |
| **Neighborhood Edwards** | -0.0751434 | 0.02716354 | -2.77 | 0.0057 |
| **Neighborhood Gilbert** | -0.0319219 | 0.02768746 | -1.15 | 0.249 |
| **Neighborhood IDOTRR** | -0.0759074 | 0.03165476 | -2.4 | 0.0166 |
| **Neighborhood MeadowV** | -0.0678992 | 0.03634107 | -1.87 | 0.0618 |
| **Neighborhood Mitchel** | -0.0324618 | 0.02759501 | -1.18 | 0.2396 |
| **Neighborhood NAmes** | -0.0273121 | 0.0262525 | -1.04 | 0.2983 |
| **Neighborhood NPkVill** | 0.05346702 | 0.03748956 | 1.43 | 0.1539 |
| **Neighborhood NWAmes** | -0.0414847 | 0.02726297 | -1.52 | 0.1282 |
| **Neighborhood NoRidge** | 0.10336373 | 0.02880613 | 3.59 | 0.0003 |
| **Neighborhood NridgHt** | 0.08388769 | 0.02806754 | 2.99 | 0.0028 |
| **Neighborhood OldTown** | -0.0551426 | 0.03035046 | -1.82 | 0.0693 |
| **Neighborhood SWISU** | -0.0014269 | 0.03231008 | -0.04 | 0.9648 |
| **Neighborhood Sawyer** | -0.007707 | 0.02744711 | -0.28 | 0.7789 |
| **Neighborhood SawyerW** | -0.0299668 | 0.0272352 | -1.1 | 0.2713 |
| **Neighborhood Somerst** | 0.0725289 | 0.03141331 | 2.31 | 0.021 |
| **Neighborhood StoneBr** | 0.1267939 | 0.03063867 | 4.14 | <.0001 |
| **Neighborhood Timber** | 0.00802611 | 0.02895956 | 0.28 | 0.7817 |
| **Condition1 Artery** | -0.0549216 | 0.04092928 | -1.34 | 0.1798 |
| **Condition1 Feedr** | -0.0283923 | 0.04003914 | -0.71 | 0.4783 |
| **Condition1 Norm** | 0.01713806 | 0.03910393 | 0.44 | 0.6612 |
| **Condition1 PosA** | 0.00916148 | 0.04780822 | 0.19 | 0.848 |
| **Condition1 PosN** | 0.0467305 | 0.04329056 | 1.08 | 0.2805 |
| **Condition1 RRAe** | -0.0534539 | 0.0451815 | -1.18 | 0.2369 |
| **Condition1 RRAn** | -0.0192339 | 0.04181401 | -0.46 | 0.6456 |
| **Condition1 RRNe** | -0.0115377 | 0.06062632 | -0.19 | 0.8491 |
| **Condition2 Artery** | -0.0420036 | 0.09675695 | -0.43 | 0.6642 |
| **Condition2 Feedr** | -0.0693281 | 0.08741277 | -0.79 | 0.4278 |
| **Condition2 Norm** | -0.0258336 | 0.08128237 | -0.32 | 0.7506 |
| **Condition2 PosA** | 0.10965546 | 0.10149725 | 1.08 | 0.2801 |
| **Condition2 PosN** | -0.366901 | 0.10112926 | -3.63 | 0.0003 |
| **Condition2 RRAe** | 0.00263066 | 0.16299037 | 0.02 | 0.9871 |
| **Condition2 RRAn** | -0.1187428 | 0.13788341 | -0.86 | 0.3892 |
| **BldgType 1Fam** | 0.02926032 | 0.0165647 | 1.77 | 0.0774 |
| **BldgType 2fmCon** | 0.02958867 | 0.02264368 | 1.31 | 0.1914 |
| **BldgType Duplex** | 0.00892155 | 0.02547054 | 0.35 | 0.7262 |
| **BldgType Twnhs** | -0.0387228 | 0.01655408 | -2.34 | 0.0194 |
| **RoofStyle Flat** | 0.01606514 | 0.07820829 | 0.21 | 0.8373 |
| **RoofStyle Gable** | 0.08716487 | 0.06126558 | 1.42 | 0.1549 |
| **RoofStyle Gambrel** | 0.07234142 | 0.06609493 | 1.09 | 0.2738 |
| **RoofStyle Hip** | 0.08547756 | 0.06144267 | 1.39 | 0.1643 |
| **RoofStyle Mansard** | 0.02995047 | 0.07018429 | 0.43 | 0.6696 |
| **RoofMatl ClyTile** | -1.7948574 | 0.14598971 | -12.29 | <.0001 |
| **RoofMatl CompShg** | -0.1286712 | 0.04635386 | -2.78 | 0.0055 |
| **RoofMatl Membran** | -0.0211476 | 0.13517173 | -0.16 | 0.8757 |
| **RoofMatl Metal** | -0.0022046 | 0.13492997 | -0.02 | 0.987 |
| **RoofMatl Roll** | -0.0522651 | 0.1242012 | -0.42 | 0.6739 |
| **RoofMatl Tar&Grv** | -0.0922007 | 0.06336548 | -1.46 | 0.1458 |
| **RoofMatl WdShake** | -0.103441 | 0.0602185 | -1.72 | 0.086 |
| **Exterior1st** | -0.0013622 | 0.00091439 | -1.49 | 0.1364 |
| **Exterior2nd 1** | -0.0742806 | 0.02188276 | -3.39 | 0.0007 |
| **Exterior2nd 2** | 0.01607078 | 0.05916696 | 0.27 | 0.7859 |
| **Exterior2nd 4** | 0.01203186 | 0.01974165 | 0.61 | 0.5423 |
| **Exterior2nd 5** | 0.02842765 | 0.06875286 | 0.41 | 0.6793 |
| **Exterior2nd 7** | -0.0172323 | 0.01081148 | -1.59 | 0.1111 |
| **Exterior2nd 8** | -0.022676 | 0.03106655 | -0.73 | 0.4655 |
| **Exterior2nd 9** | 0.00536376 | 0.0094701 | 0.57 | 0.5712 |
| **Exterior2nd 10** | -0.0642047 | 0.11261519 | -0.57 | 0.5686 |
| **Exterior2nd 11** | -0.0162115 | 0.01031394 | -1.57 | 0.1161 |
| **Exterior2nd 13** | 0.01567566 | 0.05126966 | 0.31 | 0.7598 |
| **Exterior2nd 14** | 0.03002067 | 0.01825366 | 1.64 | 0.1002 |
| **Exterior2nd 15** | -0.0015858 | 0.00829393 | -0.19 | 0.8484 |
| **MasVnrType 1** | -0.0439254 | 0.02463458 | -1.78 | 0.0747 |
| **MasVnrType 2** | -0.0106325 | 0.0090177 | -1.18 | 0.2385 |
| **MasVnrType 4** | -0.0085051 | 0.00932881 | -0.91 | 0.362 |
| **ExterQual Ex** | 0.03972451 | 0.01737832 | 2.29 | 0.0223 |
| **ExterQual Fa** | 0.01151511 | 0.02348302 | 0.49 | 0.6239 |
| **ExterQual Gd** | 0.0043665 | 0.00812785 | 0.54 | 0.5912 |
| **ExterCond Ex** | 0.00241117 | 0.0340379 | 0.07 | 0.9435 |
| **ExterCond Fa** | -0.073118 | 0.01656777 | -4.41 | <.0001 |
| **ExterCond Gd** | -0.0073729 | 0.00764724 | -0.96 | 0.3351 |
| **ExterCond Po** | -0.3237442 | 0.07292693 | -4.44 | <.0001 |
| **Foundation BrkTil** | 0.04523736 | 0.05251868 | 0.86 | 0.3891 |
| **Foundation CBlock** | 0.05199725 | 0.05174747 | 1 | 0.3151 |
| **Foundation PConc** | 0.06625188 | 0.05137677 | 1.29 | 0.1973 |
| **Foundation Slab** | 0.07682003 | 0.0588914 | 1.3 | 0.1922 |
| **Foundation Stone** | 0.13530428 | 0.06321036 | 2.14 | 0.0324 |
| **BsmtQual 1** | 0.03456686 | 0.09177636 | 0.38 | 0.7065 |
| **BsmtQual 2** | -0.012114 | 0.09115306 | -0.13 | 0.8943 |
| **BsmtQual 3** | -0.0120284 | 0.09073873 | -0.13 | 0.8946 |
| **BsmtQual 4** | -0.0207927 | 0.09127762 | -0.23 | 0.8198 |
| **BsmtCond 2** | 0.04592886 | 0.06718553 | 0.68 | 0.4943 |
| **BsmtCond 3** | 0.0301211 | 0.06640399 | 0.45 | 0.6502 |
| **BsmtCond 4** | 0.03008769 | 0.06766598 | 0.44 | 0.6566 |
| **BsmtCond 5** | 0.1221263 | 0.09014587 | 1.35 | 0.1756 |
| **BsmtExposure 1** | 0.07387982 | 0.06496392 | 1.14 | 0.2555 |
| **BsmtExposure 2** | 0.03883825 | 0.06450959 | 0.6 | 0.5472 |
| **BsmtExposure 3** | 0.00875158 | 0.06467812 | 0.14 | 0.8924 |
| **BsmtExposure 4** | 0.00877498 | 0.06433623 | 0.14 | 0.8915 |
| **BsmtFinType1 1** | 0.14288865 | 0.17850539 | 0.8 | 0.4235 |
| **BsmtFinType1 2** | 0.13987494 | 0.17858503 | 0.78 | 0.4336 |
| **BsmtFinType1 3** | 0.13017472 | 0.17867566 | 0.73 | 0.4663 |
| **BsmtFinType1 4** | 0.11313751 | 0.17871852 | 0.63 | 0.5268 |
| **BsmtFinType1 5** | 0.11450403 | 0.17928248 | 0.64 | 0.5231 |
| **BsmtFinType1 6** | 0.09984344 | 0.17865949 | 0.56 | 0.5763 |
| **BsmtFinType2 1** | -0.003386 | 0.11354961 | -0.03 | 0.9762 |
| **BsmtFinType2 2** | -0.0438456 | 0.11248301 | -0.39 | 0.6967 |
| **BsmtFinType2 3** | -0.0715073 | 0.11221598 | -0.64 | 0.524 |
| **BsmtFinType2 4** | -0.062316 | 0.11201062 | -0.56 | 0.578 |
| **BsmtFinType2 5** | -0.0517418 | 0.11205527 | -0.46 | 0.6443 |
| **BsmtFinType2 6** | -0.0163161 | 0.12064919 | -0.14 | 0.8924 |
| **Heating Floor** | 0.15214109 | 0.12907681 | 1.18 | 0.2386 |
| **Heating GasA** | 0.12660999 | 0.05611612 | 2.26 | 0.0241 |
| **Heating GasW** | 0.18987984 | 0.05986757 | 3.17 | 0.0015 |
| **Heating Grav** | -0.0025435 | 0.06854396 | -0.04 | 0.9704 |
| **Heating OthW** | 0.05942613 | 0.09898417 | 0.6 | 0.5483 |
| **HeatingQC Ex** | 0.02756603 | 0.00662782 | 4.16 | <.0001 |
| **HeatingQC Fa** | 0.01900011 | 0.01403223 | 1.35 | 0.1758 |
| **HeatingQC Gd** | 0.02060368 | 0.00695511 | 2.96 | 0.0031 |
| **HeatingQC Po** | -0.3445821 | 0.07358061 | -4.68 | <.0001 |
| **CentralAir N** | -0.0396973 | 0.0115272 | -3.44 | 0.0006 |
| **Electrical 1** | -0.1568836 | 0.14734325 | -1.06 | 0.2871 |
| **Electrical 2** | -0.1496912 | 0.14733341 | -1.02 | 0.3097 |
| **Electrical 3** | -0.1284755 | 0.14874301 | -0.86 | 0.3878 |
| **Electrical 4** | -0.1596425 | 0.15082354 | -1.06 | 0.2899 |
| **KitchenQual 1** | 0.07845983 | 0.02003322 | 3.92 | <.0001 |
| **KitchenQual 2** | 0.01357066 | 0.01667697 | 0.81 | 0.4159 |
| **KitchenQual 3** | 0.0041782 | 0.01545924 | 0.27 | 0.787 |
| **Functional 1** | 0.19368682 | 0.08427187 | 2.3 | 0.0216 |
| **Functional 2** | 0.15611769 | 0.08524732 | 1.83 | 0.0672 |
| **Functional 3** | 0.13774585 | 0.08541304 | 1.61 | 0.1069 |
| **Functional 4** | 0.15140323 | 0.0863843 | 1.75 | 0.0798 |
| **Functional 5** | 0.14962826 | 0.08804115 | 1.7 | 0.0893 |
| **Functional 6** | -0.0405526 | 0.09305066 | -0.44 | 0.663 |
| **FireplaceQu 1** | -0.0233766 | 0.02407488 | -0.97 | 0.3316 |
| **FireplaceQu 2** | -0.0369657 | 0.0166297 | -2.22 | 0.0263 |
| **FireplaceQu 3** | -0.0414644 | 0.01693878 | -2.45 | 0.0144 |
| **FireplaceQu 4** | -0.043988 | 0.02085045 | -2.11 | 0.035 |
| **FireplaceQu 5** | -0.0317799 | 0.02252234 | -1.41 | 0.1583 |
| **GarageType 1** | -0.0723786 | 0.09122813 | -0.79 | 0.4276 |
| **GarageType 2** | -0.033341 | 0.08757109 | -0.38 | 0.7034 |
| **GarageType 3** | -0.0510327 | 0.0900888 | -0.57 | 0.5711 |
| **GarageType 4** | -0.0280664 | 0.08796091 | -0.32 | 0.7497 |
| **GarageType 5** | -0.0770689 | 0.09262906 | -0.83 | 0.4055 |
| **GarageType 6** | -0.0321615 | 0.08749324 | -0.37 | 0.7132 |
| **GarageFinish 1** | -0.1073192 | 0.10856236 | -0.99 | 0.323 |
| **GarageFinish 2** | -0.1165954 | 0.10857817 | -1.07 | 0.283 |
| **GarageFinish 3** | -0.1108471 | 0.10828039 | -1.02 | 0.3061 |
| **GarageQual 1** | 0.32393486 | 0.11676231 | 2.77 | 0.0056 |
| **GarageQual 2** | 0.09061318 | 0.07426914 | 1.22 | 0.2225 |
| **GarageQual 3** | 0.04763417 | 0.06979454 | 0.68 | 0.495 |
| **GarageQual 4** | 0.04543219 | 0.06883633 | 0.66 | 0.5093 |
| **GarageCond 1** | -0.1291919 | 0.09855429 | -1.31 | 0.19 |
| **GarageCond 2** | 0.02785326 | 0.04967868 | 0.56 | 0.5751 |
| **GarageCond 3** | 0.02617713 | 0.03847226 | 0.68 | 0.4963 |
| **GarageCond 4** | -0.0079349 | 0.038909 | -0.2 | 0.8384 |
| **PavedDrive N** | -0.0280145 | 0.01079527 | -2.6 | 0.0095 |
| **PavedDrive P** | -0.0304629 | 0.01552375 | -1.96 | 0.0498 |
| **PoolQC 1** | 0.07644158 | 0.08544541 | 0.89 | 0.3711 |
| **PoolQC 2** | 0.06988172 | 0.0988488 | 0.71 | 0.4797 |
| **PoolQC 4** | -0.0415321 | 0.13344147 | -0.31 | 0.7556 |
| **Fence 1** | -0.0012026 | 0.01135718 | -0.11 | 0.9157 |
| **Fence 2** | 0.01261484 | 0.00717367 | 1.76 | 0.0788 |
| **Fence 3** | -0.0029317 | 0.01134468 | -0.26 | 0.7961 |
| **Fence 4** | -0.0009462 | 0.03295876 | -0.03 | 0.9771 |
| **MiscFeature 2** | 0.48660215 | 0.08291582 | 5.87 | <.0001 |
| **MiscFeature 3** | 0.31280616 | 0.07166774 | 4.36 | <.0001 |
| **MiscFeature 4** | 0.30820572 | 0.04518971 | 6.82 | <.0001 |
| **MiscFeature 5** | 0.29785973 | 0.16819432 | 1.77 | 0.0767 |
| **SaleType** | 0.00022704 | 0.00131105 | 0.17 | 0.8625 |
| **SaleCondition Abnorml** | -0.1201989 | 0.01246411 | -9.64 | <.0001 |
| **SaleCondition AdjLand** | 0.09364768 | 0.03636889 | 2.57 | 0.0101 |
| **SaleCondition Alloca** | -0.0384651 | 0.02760448 | -1.39 | 0.1636 |
| **SaleCondition Family** | -0.0705787 | 0.018979 | -3.72 | 0.0002 |
| **SaleCondition Normal** | -0.0267743 | 0.00909305 | -2.94 | 0.0033 |
| **MSSubClassl** | -0.0051594 | 0.00721146 | -0.72 | 0.4744 |
| **LotFrontagel** | 0.00193586 | 0.00936228 | 0.21 | 0.8362 |
| **LotAreal** | 0.08652065 | 0.00872338 | 9.92 | <.0001 |
| **OverallQuall** | 0.30355897 | 0.02068986 | 14.67 | <.0001 |
| **OverallCondl** | 0.26365874 | 0.01737404 | 15.18 | <.0001 |
| **YearBuiltl** | 3.52683902 | 0.47105823 | 7.49 | <.0001 |
| **YearRemodAddl** | 1.32104708 | 0.34769622 | 3.8 | 0.0001 |
| **MasVnrAreal** | 0.00178729 | 0.00260955 | 0.68 | 0.4935 |
| **BsmtFinSF1l** | 0.00693995 | 0.00295493 | 2.35 | 0.0189 |
| **BsmtFinSF2l** | 0.00566058 | 0.0075024 | 0.75 | 0.4506 |
| **BsmtUnfSFl** | -0.0111554 | 0.00402696 | -2.77 | 0.0056 |
| **TotalBsmtSFl** | 0.06732635 | 0.01188512 | 5.66 | <.0001 |
| **X1stFlrSFl** | 0.05308995 | 0.01693815 | 3.13 | 0.0017 |
| **X2ndFlrSFl** | 0.00921553 | 0.00609478 | 1.51 | 0.1306 |
| **LowQualFinSFl** | -0.0020486 | 0.00296531 | -0.69 | 0.4897 |
| **GrLivAreal** | 0.40248024 | 0.02140969 | 18.8 | <.0001 |
| **BsmtFullBathl** | 0.03691079 | 0.00909549 | 4.06 | <.0001 |
| **BsmtHalfBathl** | 0.0014727 | 0.01387075 | 0.11 | 0.9155 |
| **FullBathl** | 0.05698507 | 0.01725548 | 3.3 | 0.001 |
| **HalfBathl** | 0.01796293 | 0.00978064 | 1.84 | 0.0664 |
| **BedroomAbvGrl** | -0.0320492 | 0.01601877 | -2 | 0.0455 |
| **KitchenAbvGrl** | -0.1343906 | 0.0513687 | -2.62 | 0.0089 |
| **TotRmsAbvGrdl** | -0.0115745 | 0.0220366 | -0.53 | 0.5995 |
| **Fireplacesl** | 0.0810104 | 0.02080925 | 3.89 | 0.0001 |
| **GarageYrBltl** | 0.38337148 | 0.32215017 | 1.19 | 0.2341 |
| **GarageCarsl** | 0.08125059 | 0.02102412 | 3.86 | 0.0001 |
| **GarageAreal** | 0.00787913 | 0.01235689 | 0.64 | 0.5238 |
| **WoodDeckSFl** | 0.00206942 | 0.00094473 | 2.19 | 0.0286 |
| **OpenPorchSFl** | 0.00074731 | 0.00122807 | 0.61 | 0.5429 |
| **EnclosedPorchl** | 0.00368767 | 0.00141863 | 2.6 | 0.0094 |
| **X3SsnPorchl** | 0.00084482 | 0.00375898 | 0.22 | 0.8222 |
| **ScreenPorchl** | 0.00766621 | 0.00152017 | 5.04 | <.0001 |
| **PoolAreal** | 0.0144426 | 0.01097819 | 1.32 | 0.1884 |
| **MiscVall** | -0.0496929 | 0.00691613 | -7.19 | <.0001 |

**Conclusion**

The final model with the best Kaggle score is (0.10572). There are several parameters in the model that predict a positive or negative influence on the sales price of a house:

We conclude that Crawford neighborhood, Northridge neighborhood, Northridge Heights neighborhood, Somerset neighborhood, Stone Brook neighborhood, excellent exterior quality, stone foundation, heating gas forced warm air furnace, heating gas hot water or steam heat, excellent heating, good heating, excellent kitchen, typical home functionality, excellent garage, second garage, other feature, shed feature, sale condition adjoining land purchase, lot area, overall quality, overall condition, year built, year remodeled, basement finished area, total basement area, first floor area, above grade living area, number of basement full baths, number of full baths, number of fireplaces, number of car garage, wood deck area, enclosed porch area, and screen porch area have significant parameters that predict an increase in housing sales price.

We conclude that gravel street, banked land contour, lot frontage on 2 sides of property, Edwards neighborhood, Iowa Rail Road neighborhood, near positive off-site feature proximity, building type townhouse end unit, clay or tile roof material, composite shingle roof material, exterior asbestos shingles, fair exterior condition, poor exterior condition, poor heating quality, no A/C, good fireplace, average fireplace, poor fireplace, no paved driveway, partially paved driveway, abnormal sale condition, family member sale condition, normal sale, unfinished basement area, above grade bedroom area, above grade kitchen area, and miscellaneous value have significant parameters that predict a decrease in housing sales price.

Backwards selection with CV produced the best Kaggle score. We are not presumptuous with one data over another. We used interaction and it did not seem to have any parameters eventually selected by automatic selection. We only included the best parameters by contribution where logging all the variables helped auto selection produced the best regressions based on random cloud with data points that occupies the full plot rather than a cluster bias.

Not all ideal assumptions hold when using multiple-linear regression because the end goal is to retrieve the best model and the lowest kaggle score. Variance inflation factors, normality, and diagnostics for influential points are to the wayside in the real world when seeking the best predictive model.

What was more surprising as we later learn is that of the categorical variables that had 25% or less original data, MiscFeature 2, MiscFeature 3, and MiscFeature 4 had p-values that were significant to the data (Table 3). Even though the end goal is to get the best prediction value based on kaggle score, the imputation of missing values that had 25% or less original data worked too well to have the data be significant towards the final prediction model.

Upon exploring increase of a different CV split for the best backwards selection, the model's accuracy also went down. When we truncated the model to only include significant categorical parameters within a variable instead of all categorical parameters within a model, the prediction lost accuracy in the data (Table 4). Even the GLMSELECT to GLM conversion lost relevant predictability because 282 observations are left out when fitted in the CV prediction model (Figure 3). Due to 282 observations eliminated from the fitted model, some imputated data could have been left out, where the auto selection process eliminated data points from imputation by category and by observation.

Therefore, it is only strategic to include the full model in its original PROC GLMSELECT format to create the best model for better accuracy regardless of imputed data from an original data sample of 25% or less.

**Further Work**

Another imputation technique could be used to further explore imputation to get a better Kaggle score. Different CV splits or indicator other than AIC could be used to create a better model. Random Forest or other machine learning techniques could be used to create a better model.

**Appendix**

**Variable meaning:**

• SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.

• MSSubClass: The building class

• MSZoning: The general zoning classification

• LotFrontage: Linear feet of street connected to property

• LotArea: Lot size in square feet

• Street: Type of road access

• Alley: Type of alley access

• LotShape: General shape of property

• LandContour: Flatness of the property

• Utilities: Type of utilities available

• LotConfig: Lot configuration

• LandSlope: Slope of property

• Neighborhood: Physical locations within Ames city limits

• Condition1: Proximity to main road or railroad

• Condition2: Proximity to main road or railroad (if a second is present)

• BldgType: Type of dwelling

• HouseStyle: Style of dwelling

• OverallQual: Overall material and finish quality

• OverallCond: Overall condition rating

• YearBuilt: Original construction date

• YearRemodAdd: Remodel date

• RoofStyle: Type of roof

• RoofMatl: Roof material

• Exterior1st: Exterior covering on house

• Exterior2nd: Exterior covering on house (if more than one material)

• MasVnrType: Masonry veneer type

• MasVnrArea: Masonry veneer area in square feet

• ExterQual: Exterior material quality

• ExterCond: Present condition of the material on the exterior

• Foundation: Type of foundation

• BsmtQual: Height of the basement

• BsmtCond: General condition of the basement

• BsmtExposure: Walkout or garden level basement walls

• BsmtFinType1: Quality of basement finished area

• BsmtFinSF1: Type 1 finished square feet

• BsmtFinType2: Quality of second finished area (if present)

• BsmtFinSF2: Type 2 finished square feet

• BsmtUnfSF: Unfinished square feet of basement area

• TotalBsmtSF: Total square feet of basement area

• Heating: Type of heating

• HeatingQC: Heating quality and condition

• CentralAir: Central air conditioning

• Electrical: Electrical system

• 1stFlrSF: First Floor square feet

• 2ndFlrSF: Second floor square feet

• LowQualFinSF: Low quality finished square feet (all floors)

• 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: Number of bedrooms above basement level

• Kitchen: Number of kitchens

• KitchenQual: Kitchen quality

• TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

• Functional: Home functionality rating

• Fireplaces: Number of fireplaces

• FireplaceQu: Fireplace quality

• GarageType: Garage location

• GarageYrBlt: Year garage was built

• GarageFinish: Interior finish of the garage

• GarageCars: Size of garage in car capacity

• GarageArea: Size of garage in square feet

• GarageQual: Garage quality

• GarageCond: Garage condition

• PavedDrive: Paved driveway

• WoodDeckSF: Wood deck area in square feet

• OpenPorchSF: Open porch area in square feet

• EnclosedPorch: Enclosed porch area in square feet

• 3SsnPorch: Three season porch area in square feet

• ScreenPorch: Screen porch area in square feet

• PoolArea: Pool area in square feet

• PoolQC: Pool quality

• Fence: Fence quality

• MiscFeature: Miscellaneous feature not covered in other categories

• MiscVal: $Value of miscellaneous feature

• MoSold: Month Sold

• YrSold: Year Sold

• SaleType: Type of sale

• SaleCondition: Condition of sale

**Full Model**

**SalesPrice = -33.7688\* +0.0152\*MSZoning 3 +0.0229\*MSZoning 5 +0.0129\*MSZoning 6-0.1348\*Street Grvl +0.0025\*Alley 1-0.0069\*LotShape IR1 +0.0039\*LotShape IR2-0.0044\*LotShape IR3-0.0307\*LandContour Bnk +0.0065\*LandContour HLS-0.0319\*LandContour Low +0.1284\*Utilities 1 +0.0034\*LotConfig Corner +0.0090\*LotConfig CulDSac-0.0401\*LotConfig FR2-0.0276\*LotConfig FR3 +0.0223\*LandSlope Gtl +0.0364\*LandSlope Mod +0.0309\*Neighborhood Blmngtn +0.0452\*Neighborhood Blueste +0.0183\*Neighborhood BrDale +0.0199\*Neighborhood BrkSide +0.0235\*Neighborhood ClearCr-0.0112\*Neighborhood CollgCr +0.0890\*Neighborhood Crawfor-0.0751\*Neighborhood Edwards-0.0319\*Neighborhood Gilbert-0.0759\*Neighborhood IDOTRR-0.0679\*Neighborhood MeadowV-0.0325\*Neighborhood Mitchel-0.0273\*Neighborhood NAmes +0.0535\*Neighborhood NPkVill-0.0415\*Neighborhood NWAmes +0.1034\*Neighborhood NoRidge +0.0839\*Neighborhood NridgHt-0.0551\*Neighborhood OldTown-0.0014\*Neighborhood SWISU-0.0077\*Neighborhood Sawyer-0.0300\*Neighborhood SawyerW +0.0725\*Neighborhood Somerst +0.1268\*Neighborhood StoneBr +0.0080\*Neighborhood Timber-0.0549\*Condition1 Artery-0.0284\*Condition1 Feedr +0.0171\*Condition1 Norm +0.0092\*Condition1 PosA +0.0467\*Condition1 PosN-0.0535\*Condition1 RRAe-0.0192\*Condition1 RRAn-0.0115\*Condition1 RRNe-0.0420\*Condition2 Artery-0.0693\*Condition2 Feedr-0.0258\*Condition2 Norm +0.1097\*Condition2 PosA-0.3669\*Condition2 PosN +0.0026\*Condition2 RRAe-0.1187\*Condition2 RRAn +0.0293\*BldgType 1Fam +0.0296\*BldgType 2fmCon +0.0089\*BldgType Duplex-0.0387\*BldgType Twnhs +0.0161\*RoofStyle Flat +0.0872\*RoofStyle Gable +0.0723\*RoofStyle Gambrel +0.0855\*RoofStyle Hip +0.0300\*RoofStyle Mansard-1.7949\*RoofMatl ClyTile-0.1287\*RoofMatl CompShg-0.0211\*RoofMatl Membran-0.0022\*RoofMatl Metal-0.0523\*RoofMatl Roll-0.0922\*RoofMatl Tar&Grv-0.1034\*RoofMatl WdShake-0.0014\*Exterior1st-0.0743\*Exterior2nd 1 +0.0161\*Exterior2nd 2 +0.0120\*Exterior2nd 4 +0.0284\*Exterior2nd 5-0.0172\*Exterior2nd 7-0.0227\*Exterior2nd 8 +0.0054\*Exterior2nd 9-0.0642\*Exterior2nd 10-0.0162\*Exterior2nd 11 +0.0157\*Exterior2nd 13 +0.0300\*Exterior2nd 14-0.0016\*Exterior2nd 15-0.0439\*MasVnrType 1-0.0106\*MasVnrType 2-0.0085\*MasVnrType 4 +0.0397\*ExterQual Ex +0.0115\*ExterQual Fa +0.0044\*ExterQual Gd +0.0024\*ExterCond Ex-0.0731\*ExterCond Fa-0.0074\*ExterCond Gd-0.3237\*ExterCond Po +0.0452\*Foundation BrkTil +0.0520\*Foundation CBlock +0.0663\*Foundation PConc +0.0768\*Foundation Slab +0.1353\*Foundation Stone +0.0346\*BsmtQual 1-0.0121\*BsmtQual 2-0.0120\*BsmtQual 3-0.0208\*BsmtQual 4 +0.0459\*BsmtCond 2 +0.0301\*BsmtCond 3 +0.0301\*BsmtCond 4 +0.1221\*BsmtCond 5 +0.0739\*BsmtExposure 1 +0.0388\*BsmtExposure 2 +0.0088\*BsmtExposure 3 +0.0088\*BsmtExposure 4 +0.1429\*BsmtFinType1 1 +0.1399\*BsmtFinType1 2 +0.1302\*BsmtFinType1 3 +0.1131\*BsmtFinType1 4 +0.1145\*BsmtFinType1 5 +0.0998\*BsmtFinType1 6-0.0034\*BsmtFinType2 1-0.0438\*BsmtFinType2 2-0.0715\*BsmtFinType2 3-0.0623\*BsmtFinType2 4-0.0517\*BsmtFinType2 5-0.0163\*BsmtFinType2 6 +0.1521\*Heating Floor +0.1266\*Heating GasA +0.1899\*Heating GasW-0.0025\*Heating Grav +0.0594\*Heating OthW +0.0276\*HeatingQC Ex +0.0190\*HeatingQC Fa +0.0206\*HeatingQC Gd-0.3446\*HeatingQC Po-0.0397\*CentralAir N-0.1569\*Electrical 1-0.1497\*Electrical 2-0.1285\*Electrical 3-0.1596\*Electrical 4 +0.0785\*KitchenQual 1 +0.0136\*KitchenQual 2 +0.0042\*KitchenQual 3 +0.1937\*Functional 1 +0.1561\*Functional 2 +0.1377\*Functional 3 +0.1514\*Functional 4 +0.1496\*Functional 5-0.0406\*Functional 6-0.0234\*FireplaceQu 1-0.0370\*FireplaceQu 2-0.0415\*FireplaceQu 3-0.0440\*FireplaceQu 4-0.0318\*FireplaceQu 5-0.0724\*GarageType 1-0.0333\*GarageType 2-0.0510\*GarageType 3-0.0281\*GarageType 4-0.0771\*GarageType 5-0.0322\*GarageType 6-0.1073\*GarageFinish 1-0.1166\*GarageFinish 2-0.1108\*GarageFinish 3 +0.3239\*GarageQual 1 +0.0906\*GarageQual 2 +0.0476\*GarageQual 3 +0.0454\*GarageQual 4-0.1292\*GarageCond 1 +0.0279\*GarageCond 2 +0.0262\*GarageCond 3-0.0079\*GarageCond 4-0.0280\*PavedDrive N-0.0305\*PavedDrive P +0.0764\*PoolQC 1 +0.0699\*PoolQC 2-0.0415\*PoolQC 4-0.0012\*Fence 1 +0.0126\*Fence 2-0.0029\*Fence 3-0.0009\*Fence 4 +0.4866\*MiscFeature 2 +0.3128\*MiscFeature 3 +0.3082\*MiscFeature 4 +0.2979\*MiscFeature 5 +0.0002\*SaleType-0.1202\*SaleCondition Abnorml +0.0936\*SaleCondition AdjLand-0.0385\*SaleCondition Alloca-0.0706\*SaleCondition Family-0.0268\*SaleCondition Normal-0.0052\*MSSubClassl +0.0019\*LotFrontagel +0.0865\*LotAreal +0.3036\*OverallQuall +0.2637\*OverallCondl +3.5268\*YearBuiltl +1.3210\*YearRemodAddl +0.0018\*MasVnrAreal +0.0069\*BsmtFinSF1l +0.0057\*BsmtFinSF2l-0.0112\*BsmtUnfSFl +0.0673\*TotalBsmtSFl +0.0531\*X1stFlrSFl +0.0092\*X2ndFlrSFl-0.0020\*LowQualFinSFl +0.4025\*GrLivAreal +0.0369\*BsmtFullBathl +0.0015\*BsmtHalfBathl +0.0570\*FullBathl +0.0180\*HalfBathl-0.0320\*BedroomAbvGrl-0.1344\*KitchenAbvGrl-0.0116\*TotRmsAbvGrdl +0.0810\*Fireplacesl +0.3834\*GarageYrBltl +0.0813\*GarageCarsl +0.0079\*GarageAreal +0.0021\*WoodDeckSFl +0.0007\*OpenPorchSFl +0.0037\*EnclosedPorchl +0.0008\*X3SsnPorchl +0.0077\*ScreenPorchl +0.0144\*PoolAreal-0.0497\*MiscVall**

**R Imputation**

test <- read.csv("C:\\Users\\Yao\\Dropbox\\immersion future contacts\\stats MSDS 6371 404\\stats w14 MSDS 6371 404\\test.csv",

stringsAsFactors = FALSE)

train <- read.csv("C:\\Users\\Yao\\Dropbox\\immersion future contacts\\stats MSDS 6371 404\\stats w14 MSDS 6371 404\\train.csv",

stringsAsFactors = FALSE)

total <- rbind(train, test)

total[total == 0] <- NA

colSums(is.na(total))

round((nrow(total)-colSums(is.na(total)))/nrow(total)\*100, digits = 2)

setwd("C:\\Users\\Yao\\Dropbox\\stats 2 w5 MSDS 6372 403")

write.csv(total, "total0.csv", row.names=FALSE)

str(total)

total$MSZoning[total$MSZoning=="A"]<-1

total$MSZoning[total$MSZoning=="C"]<-2

total$MSZoning[total$MSZoning=="FV"]<-3

total$MSZoning[total$MSZoning=="I"]<-4

total$MSZoning[total$MSZoning=="RH"]<-5

total$MSZoning[total$MSZoning=="RL"]<-6

total$MSZoning[total$MSZoning=="RP"]<-7

total$MSZoning[total$MSZoning=="RM"]<-8

total$MSZoning <- as.numeric(as.character(total$MSZoning))

MSZoning <- total[['MSZoning']]

MSZoning2 <- MSZoning[!is.na(MSZoning)]

MSZoning3 <- sample(MSZoning2, size=(2919-length(MSZoning2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$MSZoning[i]) == TRUE){

total$MSZoning[i] <- MSZoning3[j]

j<-j+1;

}

}

LotFrontage <- total[['LotFrontage']]

LotFrontage2 <- LotFrontage[!is.na(LotFrontage)]

LotFrontage3 <- sample(LotFrontage2, size=(2919-length(LotFrontage2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$LotFrontage[i]) == TRUE){

total$LotFrontage[i] <- LotFrontage3[j]

j<-j+1;

}

}

total$Alley[total$Alley=="Grvl"]<-1

total$Alley[total$Alley=="Pave"]<-2

total$Alley <- as.numeric(as.character(total$Alley))

Alley <- total[['Alley']]

Alley2 <- Alley[!is.na(Alley)]

Alley3 <- sample(Alley2, size=(2919-length(Alley2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$Alley[i]) == TRUE){

total$Alley[i] <- Alley3[j]

j<-j+1;

}

}

total$Utilities[total$Utilities=="AllPub"]<-1

total$Utilities[total$Utilities=="NoSewr"]<-2

total$Utilities[total$Utilities=="NoSeWa"]<-3

total$Utilities[total$Utilities=="ELO"]<-4

total$Utilities <- as.numeric(as.character(total$Utilities))

Utilities <- total[['Utilities']]

Utilities2 <- Utilities[!is.na(Utilities)]

Utilities3 <- sample(Utilities2, size=(2919-length(Utilities2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$Utilities[i]) == TRUE){

total$Utilities[i] <- Utilities3[j]

j<-j+1;

}

}

total$Exterior1st[total$Exterior1st=="AsbShng"]<-1

total$Exterior1st[total$Exterior1st=="AsphShn"]<-2

total$Exterior1st[total$Exterior1st=="BrkComm"]<-3

total$Exterior1st[total$Exterior1st=="BrkFace"]<-4

total$Exterior1st[total$Exterior1st=="CBlock"]<-5

total$Exterior1st[total$Exterior1st=="CemntBd"]<-6

total$Exterior1st[total$Exterior1st=="HdBoard"]<-7

total$Exterior1st[total$Exterior1st=="ImStucc"]<-8

total$Exterior1st[total$Exterior1st=="MetalSd"]<-9

total$Exterior1st[total$Exterior1st=="Other"]<-10

total$Exterior1st[total$Exterior1st=="Plywood"]<-11

total$Exterior1st[total$Exterior1st=="PreCast"]<-12

total$Exterior1st[total$Exterior1st=="Stone"]<-13

total$Exterior1st[total$Exterior1st=="Stucco"]<-14

total$Exterior1st[total$Exterior1st=="VinylSd"]<-15

total$Exterior1st[total$Exterior1st=="Wd Sdng"]<-16

total$Exterior1st[total$Exterior1st=="WdShing"]<-17

total$Exterior1st <- as.numeric(as.character(total$Exterior1st))

Exterior1st <- total[['Exterior1st']]

Exterior1st2 <- Exterior1st[!is.na(Exterior1st)]

Exterior1st3 <- sample(Exterior1st2, size=(2919-length(Exterior1st2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$Exterior1st[i]) == TRUE){

total$Exterior1st[i] <- Exterior1st3[j]

j<-j+1;

}

}

total$Exterior2nd[total$Exterior2nd=="AsbShng"]<-1

total$Exterior2nd[total$Exterior2nd=="AsphShn"]<-2

total$Exterior2nd[total$Exterior2nd=="BrkComm"]<-3

total$Exterior2nd[total$Exterior2nd=="BrkFace"]<-4

total$Exterior2nd[total$Exterior2nd=="CBlock"]<-5

total$Exterior2nd[total$Exterior2nd=="CemntBd"]<-6

total$Exterior2nd[total$Exterior2nd=="HdBoard"]<-7

total$Exterior2nd[total$Exterior2nd=="ImStucc"]<-8

total$Exterior2nd[total$Exterior2nd=="MetalSd"]<-9

total$Exterior2nd[total$Exterior2nd=="Other"]<-10

total$Exterior2nd[total$Exterior2nd=="Plywood"]<-11

total$Exterior2nd[total$Exterior2nd=="PreCast"]<-12

total$Exterior2nd[total$Exterior2nd=="Stone"]<-13

total$Exterior2nd[total$Exterior2nd=="Stucco"]<-14

total$Exterior2nd[total$Exterior2nd=="VinylSd"]<-15

total$Exterior2nd[total$Exterior2nd=="Wd Sdng"]<-16

total$Exterior2nd[total$Exterior2nd=="WdShing"]<-17

total$Exterior2nd <- as.numeric(as.character(total$Exterior2nd))

Exterior2nd <- total[['Exterior2nd']]

Exterior2nd2 <- Exterior2nd[!is.na(Exterior2nd)]

Exterior2nd3 <- sample(Exterior2nd2, size=(2919-length(Exterior2nd2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$Exterior2nd[i]) == TRUE){

total$Exterior2nd[i] <- Exterior2nd3[j]

j<-j+1;

}

}

total$MasVnrType[total$MasVnrType=="BrkCmn"]<-1

total$MasVnrType[total$MasVnrType=="BrkFace"]<-2

total$MasVnrType[total$MasVnrType=="CBlock"]<-3

total$MasVnrType[total$MasVnrType=="None"]<-4

total$MasVnrType[total$MasVnrType=="Stone"]<-5

total$MasVnrType <- as.numeric(as.character(total$MasVnrType))

MasVnrType <- total[['MasVnrType']]

MasVnrType2 <- MasVnrType[!is.na(MasVnrType)]

MasVnrType3 <- sample(MasVnrType2, size=(2919-length(MasVnrType2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$MasVnrType[i]) == TRUE){

total$MasVnrType[i] <- MasVnrType3[j]

j<-j+1;

}

}

MasVnrArea <- total[['MasVnrArea']]

MasVnrArea2 <- MasVnrArea[!is.na(MasVnrArea)]

MasVnrArea3 <- sample(MasVnrArea2, size=(2919-length(MasVnrArea2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$MasVnrArea[i]) == TRUE){

total$MasVnrArea[i] <- MasVnrArea3[j]

j<-j+1;

}

}

total$BsmtQual[total$BsmtQual=="Ex"]<-1

total$BsmtQual[total$BsmtQual=="Gd"]<-2

total$BsmtQual[total$BsmtQual=="TA"]<-3

total$BsmtQual[total$BsmtQual=="Fa"]<-4

total$BsmtQual[total$BsmtQual=="Po"]<-5

total$BsmtQual[is.na(total$BsmtQual)]<-6

total$BsmtQual <- as.numeric(as.character(total$BsmtQual))

total$BsmtCond[total$BsmtCond=="Ex"]<-1

total$BsmtCond[total$BsmtCond=="Gd"]<-2

total$BsmtCond[total$BsmtCond=="TA"]<-3

total$BsmtCond[total$BsmtCond=="Fa"]<-4

total$BsmtCond[total$BsmtCond=="Po"]<-5

total$BsmtCond[is.na(total$BsmtCond)]<-6

total$BsmtCond <- as.numeric(as.character(total$BsmtCond))

total$BsmtExposure[total$BsmtExposure=="Gd"]<-1

total$BsmtExposure[total$BsmtExposure=="Av"]<-2

total$BsmtExposure[total$BsmtExposure=="Mn"]<-3

total$BsmtExposure[total$BsmtExposure=="No"]<-4

total$BsmtExposure[is.na(total$BsmtExposure)]<-5

total$BsmtExposure <- as.numeric(as.character(total$BsmtExposure))

total$BsmtFinType1[total$BsmtFinType1=="GLQ"]<-1

total$BsmtFinType1[total$BsmtFinType1=="ALQ"]<-2

total$BsmtFinType1[total$BsmtFinType1=="BLQ"]<-3

total$BsmtFinType1[total$BsmtFinType1=="Rec"]<-4

total$BsmtFinType1[total$BsmtFinType1=="LwQ"]<-5

total$BsmtFinType1[total$BsmtFinType1=="Unf"]<-6

total$BsmtFinType1[is.na(total$BsmtFinType1)]<-7

total$BsmtFinType1 <- as.numeric(as.character(total$BsmtFinType1))

BsmtFinSF1 <- total[['BsmtFinSF1']]

BsmtFinSF12 <- BsmtFinSF1[!is.na(BsmtFinSF1)]

BsmtFinSF13 <- sample(BsmtFinSF12, size=(2919-length(BsmtFinSF12)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$BsmtFinSF1[i]) == TRUE){

total$BsmtFinSF1[i] <- BsmtFinSF13[j]

j<-j+1;

}

}

total$BsmtFinType2[total$BsmtFinType2=="GLQ"]<-1

total$BsmtFinType2[total$BsmtFinType2=="ALQ"]<-2

total$BsmtFinType2[total$BsmtFinType2=="BLQ"]<-3

total$BsmtFinType2[total$BsmtFinType2=="Rec"]<-4

total$BsmtFinType2[total$BsmtFinType2=="LwQ"]<-5

total$BsmtFinType2[total$BsmtFinType2=="Unf"]<-6

total$BsmtFinType2[is.na(total$BsmtFinType2)]<-7

total$BsmtFinType2 <- as.numeric(as.character(total$BsmtFinType2))

total$BsmtFinSF2[is.na(total$BsmtFinSF2)]<-0

BsmtUnfSF <- total[['BsmtUnfSF']]

BsmtUnfSF2 <- BsmtUnfSF[!is.na(BsmtUnfSF)]

BsmtUnfSF3 <- sample(BsmtUnfSF2, size=(2919-length(BsmtUnfSF2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$BsmtUnfSF[i]) == TRUE){

total$BsmtUnfSF[i] <- BsmtUnfSF3[j]

j<-j+1;

}

}

TotalBsmtSF <- total[['TotalBsmtSF']]

TotalBsmtSF2 <- TotalBsmtSF[!is.na(TotalBsmtSF)]

TotalBsmtSF3 <- sample(TotalBsmtSF2, size=(2919-length(TotalBsmtSF2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$TotalBsmtSF[i]) == TRUE){

total$TotalBsmtSF[i] <- TotalBsmtSF3[j]

j<-j+1;

}

}

total$Electrical[total$Electrical=="SBrkr"]<-1

total$Electrical[total$Electrical=="FuseA"]<-2

total$Electrical[total$Electrical=="FuseF"]<-3

total$Electrical[total$Electrical=="FuseP"]<-4

total$Electrical[total$Electrical=="Mix"]<-5

total$Electrical <- as.numeric(as.character(total$Electrical))

Electrical <- total[['Electrical']]

Electrical2 <- Electrical[!is.na(Electrical)]

Electrical3 <- sample(Electrical2, size=(2919-length(Electrical2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$Electrical[i]) == TRUE){

total$Electrical[i] <- Electrical3[j]

j<-j+1;

}

}

X2ndFlrSF <- total[['X2ndFlrSF']]

X2ndFlrSF2 <- X2ndFlrSF[!is.na(X2ndFlrSF)]

X2ndFlrSF3 <- sample(X2ndFlrSF2, size=(2919-length(X2ndFlrSF2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$X2ndFlrSF[i]) == TRUE){

total$X2ndFlrSF[i] <- X2ndFlrSF3[j]

j<-j+1;

}

}

LowQualFinSF <- total[['LowQualFinSF']]

LowQualFinSF2 <- LowQualFinSF[!is.na(LowQualFinSF)]

LowQualFinSF3 <- sample(LowQualFinSF2, size=(2919-length(LowQualFinSF2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$LowQualFinSF[i]) == TRUE){

total$LowQualFinSF[i] <- LowQualFinSF3[j]

j<-j+1;

}

}

total$BsmtFullBath[is.na(total$BsmtFullBath)] <- 0

total$BsmtHalfBath[is.na(total$BsmtHalfBath)] <- 0

total$FullBath[is.na(total$FullBath)] <- 0

total$HalfBath[is.na(total$HalfBath)] <- 0

BedroomAbvGr <- total[['BedroomAbvGr']]

BedroomAbvGr2 <- BedroomAbvGr[!is.na(BedroomAbvGr)]

BedroomAbvGr3 <- sample(BedroomAbvGr2, size=(2919-length(BedroomAbvGr2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$BedroomAbvGr[i]) == TRUE){

total$BedroomAbvGr[i] <- BedroomAbvGr3[j]

j<-j+1;

}

}

KitchenAbvGr <- total[['KitchenAbvGr']]

KitchenAbvGr2 <- KitchenAbvGr[!is.na(KitchenAbvGr)]

KitchenAbvGr3 <- sample(KitchenAbvGr2, size=(2919-length(KitchenAbvGr2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$KitchenAbvGr[i]) == TRUE){

total$KitchenAbvGr[i] <- KitchenAbvGr3[j]

j<-j+1;

}

}

total$KitchenQual[total$KitchenQual=="Ex"]<-1

total$KitchenQual[total$KitchenQual=="Gd"]<-2

total$KitchenQual[total$KitchenQual=="TA"]<-3

total$KitchenQual[total$KitchenQual=="Fa"]<-4

total$KitchenQual[total$KitchenQual=="Po"]<-5

total$KitchenQual <- as.numeric(as.character(total$KitchenQual))

KitchenQual <- total[['KitchenQual']]

KitchenQual2 <- KitchenQual[!is.na(KitchenQual)]

KitchenQual3 <- sample(KitchenQual2, size=(2919-length(KitchenQual2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$KitchenQual[i]) == TRUE){

total$KitchenQual[i] <- KitchenQual3[j]

j<-j+1;

}

}

total$Functional[total$Functional=="Typ"]<-1

total$Functional[total$Functional=="Min1"]<-2

total$Functional[total$Functional=="Min2"]<-3

total$Functional[total$Functional=="Mod"]<-4

total$Functional[total$Functional=="Maj1"]<-5

total$Functional[total$Functional=="Maj2"]<-6

total$Functional[total$Functional=="Sev"]<-7

total$Functional[total$Functional=="Sal"]<-8

total$Functional <- as.numeric(as.character(total$Functional))

Functional <- total[['Functional']]

Functional2 <- Functional[!is.na(Functional)]

Functional3 <- sample(Functional2, size=(2919-length(Functional2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$Functional[i]) == TRUE){

total$Functional[i] <- Functional3[j]

j<-j+1;

}

}

total$Fireplaces[is.na(total$Fireplaces)] <- 0

total$FireplaceQu[total$FireplaceQu=="Ex"]<-1

total$FireplaceQu[total$FireplaceQu=="Gd"]<-2

total$FireplaceQu[total$FireplaceQu=="TA"]<-3

total$FireplaceQu[total$FireplaceQu=="Fa"]<-4

total$FireplaceQu[total$FireplaceQu=="Po"]<-5

total$FireplaceQu[is.na(total$FireplaceQu)] <-6

total$FireplaceQu <- as.numeric(as.character(total$FireplaceQu))

total$GarageType[total$GarageType=="2Types"]<-1

total$GarageType[total$GarageType=="Attchd"]<-2

total$GarageType[total$GarageType=="Basment"]<-3

total$GarageType[total$GarageType=="BuiltIn"]<-4

total$GarageType[total$GarageType=="CarPort"]<-5

total$GarageType[total$GarageType=="Detchd"]<-6

total$GarageType[is.na(total$GarageType)]<-7

total$GarageType <- as.numeric(as.character(total$GarageType))

GarageYrBlt <- total[['GarageYrBlt']]

GarageYrBlt2 <- GarageYrBlt[!is.na(GarageYrBlt)]

GarageYrBlt3 <- sample(GarageYrBlt2, size=(2919-length(GarageYrBlt2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$GarageYrBlt[i]) == TRUE){

total$GarageYrBlt[i] <- GarageYrBlt3[j]

j<-j+1;

}

}

total$GarageFinish[total$GarageFinish=="Fin"]<-1

total$GarageFinish[total$GarageFinish=="RFn"]<-2

total$GarageFinish[total$GarageFinish=="Unf"]<-3

total$GarageFinish[is.na(total$GarageFinish)] <-4

total$GarageFinish <- as.numeric(as.character(total$GarageFinish))

total$GarageCars[is.na(total$GarageCars)] <-0

total$GarageArea[is.na(total$GarageArea)] <-0

total$GarageQual[total$GarageQual=="Ex"]<-1

total$GarageQual[total$GarageQual=="Gd"]<-2

total$GarageQual[total$GarageQual=="TA"]<-3

total$GarageQual[total$GarageQual=="Fa"]<-4

total$GarageQual[total$GarageQual=="Po"]<-5

total$GarageQual[is.na(total$GarageQual)]<-6

total$GarageQual <- as.numeric(as.character(total$GarageQual))

total$GarageCond[total$GarageCond=="Ex"]<-1

total$GarageCond[total$GarageCond=="Gd"]<-2

total$GarageCond[total$GarageCond=="TA"]<-3

total$GarageCond[total$GarageCond=="Fa"]<-4

total$GarageCond[total$GarageCond=="Po"]<-5

total$GarageCond[is.na(total$GarageCond)]<-6

total$GarageCond <- as.numeric(as.character(total$GarageCond))

total$WoodDeckSF[is.na(total$WoodDeckSF)]<-0

total$OpenPorchSF[is.na(total$OpenPorchSF)]<-0

total$EnclosedPorch[is.na(total$EnclosedPorch)]<-0

total$X3SsnPorch[is.na(total$X3SsnPorch)]<-0

total$ScreenPorch[is.na(total$ScreenPorch)]<-0

total$PoolArea[is.na(total$PoolArea)]<-0

total$PoolQC[total$PoolQC=="Ex"]<-1

total$PoolQC[total$PoolQC=="Gd"]<-2

total$PoolQC[total$PoolQC=="TA"]<-3

total$PoolQC[total$PoolQC=="Fa"]<-4

total$PoolQC[is.na(total$PoolQC)]<-5

total$PoolQC <- as.numeric(as.character(total$PoolQC))

total$Fence[total$Fence=="GdPrv"]<-1

total$Fence[total$Fence=="MnPrv"]<-2

total$Fence[total$Fence=="GdWo"]<-3

total$Fence[total$Fence=="MnWw"]<-4

total$Fence[is.na(total$Fence)]<-5

total$Fence <- as.numeric(as.character(total$Fence))

total$MiscFeature[total$MiscFeature=="Elev"]<-1

total$MiscFeature[total$MiscFeature=="Gar2"]<-2

total$MiscFeature[total$MiscFeature=="Othr"]<-3

total$MiscFeature[total$MiscFeature=="Shed"]<-4

total$MiscFeature[total$MiscFeature=="TenC"]<-5

total$MiscFeature[is.na(total$MiscFeature)]<-6

total$MiscFeature <- as.numeric(as.character(total$MiscFeature))

total$MiscVal[is.na(total$MiscVal)]<-0

total$SaleType[total$SaleType=="WD "]<-1

total$SaleType[total$SaleType=="CWD"]<-2

total$SaleType[total$SaleType=="VWD"]<-3

total$SaleType[total$SaleType=="New"]<-4

total$SaleType[total$SaleType=="COD"]<-5

total$SaleType[total$SaleType=="Con"]<-6

total$SaleType[total$SaleType=="ConLw"]<-7

total$SaleType[total$SaleType=="ConLI"]<-8

total$SaleType[total$SaleType=="ConLD"]<-9

total$SaleType[total$SaleType=="Oth"]<-10

total$SaleType <- as.numeric(as.character(total$SaleType))

SaleType <- total[['SaleType']]

SaleType2 <- SaleType[!is.na(SaleType)]

SaleType3 <- sample(SaleType2, size=(2919-length(SaleType2)), replace=TRUE)

j<-1

for(i in 1:2919){

if(is.na(total$SaleType[i]) == TRUE){

total$SaleType[i] <- SaleType3[j]

j<-j+1;

}

}

total75 <- total[, !(colnames(total) %in% c("Alley","MasVnrArea","BsmtFinSF1","BsmtFinSF2",

"X2ndFlrSF", "LowQualFinSF", "BsmtFullBath",

"BsmtHalfBath","HalfBath","Fireplaces","FireplaceQu",

"WoodDeckSF","OpenPorchSF","EnclosedPorch","X3SsnPorch",

"ScreenPorch","PoolArea","PoolQC","Fence","MiscFeature",

"MiscVal"))]

total50 <- total[, !(colnames(total) %in% c("Alley","MasVnrArea","BsmtFinSF2",

"X2ndFlrSF", "LowQualFinSF", "BsmtFullBath",

"BsmtHalfBath","HalfBath",

"WoodDeckSF","EnclosedPorch","X3SsnPorch",

"ScreenPorch","PoolArea","PoolQC","Fence","MiscFeature",

"MiscVal"))]

trainfull <- total[1:1460,]

testfull <- total[1461:2919,]

train50imp <- total50[1:1460,]

test50imp <- total50[1461:2919,]

train75imp <- total75[1:1460,]

test75imp <- total75[1461:2919,]

write.csv(total, "totalfull.csv", row.names=FALSE)

write.csv(total50, "total50.csv", row.names=FALSE)

write.csv(total75, "total75.csv", row.names=FALSE)

write.csv(trainfull, "trainfull.csv", row.names=FALSE)

write.csv(testfull, "testfull.csv", row.names=FALSE)

write.csv(train50imp, "train50imp.csv", row.names=FALSE)

write.csv(test50imp, "test50imp.csv", row.names=FALSE)

write.csv(train75imp, "train75imp.csv", row.names=FALSE)

write.csv(test75imp, "test75imp.csv", row.names=FALSE)

**SAS**

**Full Model**

FILENAME REFFILE '/home/yaoy890/proj 1/totalfull.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=totalfull;

GETNAMES=YES;

RUN;

proc sgscatter data=totalfull;

matrix SalePrice MSSubClass LotFrontage LotArea OverallQual OverallCond / diagonal=(histogram) ;

run;

proc sgscatter data=totalfull;

matrix SalePrice YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 / diagonal=(histogram) ;

run;

proc sgscatter data=totalfull;

matrix SalePrice BsmtUnfSF TotalBsmtSF X1stFlrSF X2ndFlrSF LowQualFinSF / diagonal=(histogram) ;

run;

proc sgscatter data=totalfull;

matrix SalePrice GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath / diagonal=(histogram) ;

run;

proc sgscatter data=totalfull;

matrix SalePrice BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt / diagonal=(histogram) ;

run;

proc sgscatter data=totalfull;

matrix SalePrice GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch / diagonal=(histogram) ;

run;

proc sgscatter data=totalfull;

matrix SalePrice X3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold / diagonal=(histogram) ;

run;

data totalfull2;

set totalfull;

MSSubClassl=log(MSSubClass+1);

LotFrontagel=log(LotFrontage+1);

LotAreal=log(LotArea+1);

OverallQuall=log(OverallQual+1);

OverallCondl=log(OverallCond+1);

YearBuiltl=log(YearBuilt+1);

YearRemodAddl=log(YearRemodAdd+1);

MasVnrAreal=log(MasVnrArea+1);

BsmtFinSF1l=log(BsmtFinSF1+1);

BsmtFinSF2l=log(BsmtFinSF2+1);

BsmtUnfSFl=log(BsmtUnfSF+1);

TotalBsmtSFl=log(TotalBsmtSF+1);

X1stFlrSFl=log(X1stFlrSF+1);

X2ndFlrSFl=log(X2ndFlrSF+1);

LowQualFinSFl=log(LowQualFinSF+1);

GrLivAreal=log(GrLivArea+1);

BsmtFullBathl=log(BsmtFullBath+1);

BsmtHalfBathl=log(BsmtHalfBath+1);

FullBathl=log(FullBath+1);

HalfBathl=log(HalfBath+1);

BedroomAbvGrl=log(BedroomAbvGr+1);

KitchenAbvGrl=log(KitchenAbvGr+1);

TotRmsAbvGrdl=log(TotRmsAbvGrd+1);

Fireplacesl=log(Fireplaces+1);

GarageYrBltl=log(GarageYrBlt+1);

GarageCarsl=log(GarageCars+1);

GarageAreal=log(GarageArea+1);

WoodDeckSFl=log(WoodDeckSF+1);

OpenPorchSFl=log(OpenPorchSF+1);

EnclosedPorchl=log(EnclosedPorch+1);

X3SsnPorchl=log(X3SsnPorch+1);

ScreenPorchl=log(ScreenPorch+1);

PoolAreal=log(PoolArea+1);

MiscVall=log(MiscVal+1);

MoSoldl=log(MoSold+1);

YrSoldl=log(YrSold+1);

SalePricel=log(SalePrice+1);

run;

proc sgscatter data=totalfull2;

matrix SalePricel MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl / diagonal=(histogram) ;

run;

proc sgscatter data=totalfull2;

matrix SalePricel YearBuiltl YearRemodAddl MasVnrAreal BsmtFinSF1l BsmtFinSF2l / diagonal=(histogram) ;

run;

proc sgscatter data=totalfull2;

matrix SalePricel BsmtUnfSFl TotalBsmtSFl X1stFlrSFl X2ndFlrSFl LowQualFinSFl / diagonal=(histogram) ;

run;

proc sgscatter data=totalfull2;

matrix SalePricel GrLivAreal BsmtFullBathl BsmtHalfBathl FullBathl HalfBathl / diagonal=(histogram) ;

run;

proc sgscatter data=totalfull2;

matrix SalePricel BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl / diagonal=(histogram) ;

run;

proc sgscatter data=totalfull2;

matrix SalePricel GarageCarsl GarageAreal WoodDeckSFl OpenPorchSFl EnclosedPorchl / diagonal=(histogram) ;

run;

proc sgscatter data=totalfull2;

matrix SalePricel X3SsnPorchl ScreenPorchl PoolAreal MiscVall MoSoldl YrSoldl / diagonal=(histogram) ;

run;

\*forward;

ods graphics on;

proc glmselect data=totalfull2 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

partition fraction(validate = 0.3 test = 0.2);

class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

model SalePricel = MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl MasVnrAreal BsmtFinSF1l BsmtFinSF2l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl X2ndFlrSFl LowQualFinSFl GrLivAreal BsmtFullBathl BsmtHalfBathl FullBathl HalfBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal WoodDeckSFl OpenPorchSFl EnclosedPorchl X3SsnPorchl ScreenPorchl PoolAreal MiscVall / selection=FORWARD(choose=CV stop=AIC) CVdetails showpvalues;

output out = totalforward p = Predict;

run; quit;

ods graphics off;

data totalforward2; set totalforward;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

ods graphics on;

proc glmselect data=totalfull2 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

model SalePricel = MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl MasVnrAreal BsmtFinSF1l BsmtFinSF2l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl X2ndFlrSFl LowQualFinSFl GrLivAreal BsmtFullBathl BsmtHalfBathl FullBathl HalfBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal WoodDeckSFl OpenPorchSFl EnclosedPorchl X3SsnPorchl ScreenPorchl PoolAreal MiscVall / selection=FORWARD(choose=CV stop=AIC) CVdetails showpvalues;

output out = totalforwardCV p = Predict;

run; quit;

ods graphics off;

data totalforwardCV2; set totalforwardCV;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

\*BACKWARD;

ods graphics on;

proc glmselect data=totalfull2 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

partition fraction(validate = 0.3 test = 0.2);

class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

model SalePricel = MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl MasVnrAreal BsmtFinSF1l BsmtFinSF2l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl X2ndFlrSFl LowQualFinSFl GrLivAreal BsmtFullBathl BsmtHalfBathl FullBathl HalfBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal WoodDeckSFl OpenPorchSFl EnclosedPorchl X3SsnPorchl ScreenPorchl PoolAreal MiscVall / selection=BACKWARD(choose=CV stop=AIC) CVdetails showpvalues;

output out = totalbackward p = Predict;

run; quit;

ods graphics off;

data totalbackward2; set totalbackward;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

ods graphics on;

proc glmselect data=totalfull2 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

model SalePricel = MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl MasVnrAreal BsmtFinSF1l BsmtFinSF2l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl X2ndFlrSFl LowQualFinSFl GrLivAreal BsmtFullBathl BsmtHalfBathl FullBathl HalfBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal WoodDeckSFl OpenPorchSFl EnclosedPorchl X3SsnPorchl ScreenPorchl PoolAreal MiscVall / selection=BACKWARD(choose=CV stop=AIC) CVdetails showpvalues;

output out = totalbackwardCV p = Predict;

run; quit;

ods graphics off;

data totalbackwardCV2; set totalbackwardCV;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

\*STEPWISE;

ods graphics on;

proc glmselect data=totalfull2 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

partition fraction(validate = 0.3 test = 0.2);

class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

model SalePricel = MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl MasVnrAreal BsmtFinSF1l BsmtFinSF2l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl X2ndFlrSFl LowQualFinSFl GrLivAreal BsmtFullBathl BsmtHalfBathl FullBathl HalfBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal WoodDeckSFl OpenPorchSFl EnclosedPorchl X3SsnPorchl ScreenPorchl PoolAreal MiscVall / selection=stepwise(choose=CV stop=AIC) CVdetails showpvalues;

output out = totalstepwise p = Predict;

run; quit;

ods graphics off;

data totalstepwise2; set totalstepwise;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

ods graphics on;

proc glmselect data=totalfull2 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

model SalePricel = MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl MasVnrAreal BsmtFinSF1l BsmtFinSF2l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl X2ndFlrSFl LowQualFinSFl GrLivAreal BsmtFullBathl BsmtHalfBathl FullBathl HalfBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal WoodDeckSFl OpenPorchSFl EnclosedPorchl X3SsnPorchl ScreenPorchl PoolAreal MiscVall / selection=stepwise(choose=CV stop=AIC) CVdetails showpvalues;

output out = totalstepwiseCV p = Predict;

run; quit;

ods graphics off;

data totalstepwiseCV2; set totalstepwiseCV;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

\*LAR;

ods graphics on;

proc glmselect data=totalfull2 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

partition fraction(validate = 0.3 test = 0.2);

class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

model SalePricel = MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl MasVnrAreal BsmtFinSF1l BsmtFinSF2l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl X2ndFlrSFl LowQualFinSFl GrLivAreal BsmtFullBathl BsmtHalfBathl FullBathl HalfBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal WoodDeckSFl OpenPorchSFl EnclosedPorchl X3SsnPorchl ScreenPorchl PoolAreal MiscVall / selection=lar(choose=CV stop=AIC) CVdetails showpvalues;

output out = totallar p = Predict;

run; quit;

ods graphics off;

data totallar2; set totallar;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

ods graphics on;

proc glmselect data=totalfull2 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

model SalePricel = MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl MasVnrAreal BsmtFinSF1l BsmtFinSF2l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl X2ndFlrSFl LowQualFinSFl GrLivAreal BsmtFullBathl BsmtHalfBathl FullBathl HalfBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal WoodDeckSFl OpenPorchSFl EnclosedPorchl X3SsnPorchl ScreenPorchl PoolAreal MiscVall / selection=lar(choose=CV stop=AIC) CVdetails showpvalues;

output out = totallarCV p = Predict;

run; quit;

ods graphics off;

data totallarCV2; set totallarCV;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

\*LASSO;

ods graphics on;

proc glmselect data=totalfull2 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

partition fraction(validate = 0.3 test = 0.2);

class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

model SalePricel = MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl MasVnrAreal BsmtFinSF1l BsmtFinSF2l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl X2ndFlrSFl LowQualFinSFl GrLivAreal BsmtFullBathl BsmtHalfBathl FullBathl HalfBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal WoodDeckSFl OpenPorchSFl EnclosedPorchl X3SsnPorchl ScreenPorchl PoolAreal MiscVall / selection=lasso(choose=CV stop=AIC) CVdetails showpvalues;

output out = totallasso p = Predict;

run; quit;

ods graphics off;

data totallasso2; set totallasso;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

ods graphics on;

proc glmselect data=totalfull2 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition;

model SalePricel = MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl MasVnrAreal BsmtFinSF1l BsmtFinSF2l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl X2ndFlrSFl LowQualFinSFl GrLivAreal BsmtFullBathl BsmtHalfBathl FullBathl HalfBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal WoodDeckSFl OpenPorchSFl EnclosedPorchl X3SsnPorchl ScreenPorchl PoolAreal MiscVall / selection=lasso(choose=CV stop=AIC) CVdetails showpvalues;

output out = totallassoCV p = Predict;

run; quit;

ods graphics off;

data totallassoCV2; set totallassoCV;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

\*proc reg from winner;

data totalfull3;

set totalfull2;

dumBE1=(BsmtExposure='1');

dumBE2=(BsmtExposure='2');

dumBE3=(BsmtExposure='3');

dumBE4=(BsmtExposure='4');

dumBF1=(BsmtFinType1='1');

dumBF2=(BsmtFinType1='2');

dumBF3=(BsmtFinType1='3');

dumBF4=(BsmtFinType1='4');

dumBF5=(BsmtFinType1='5');

dumBF6=(BsmtFinType1='6');

dumEx=(HeatingQC='Ex');

dumFa=(HeatingQC='Fa');

dumGd=(HeatingQC='Gd');

dumPo=(HeatingQC='Po');

dumK1=(KitchenQual='1');

dumK2=(KitchenQual='2');

dumK3=(KitchenQual='3');

dumBlmngtn=(Neighborhood='Blmngtn');

dumBlueste=(Neighborhood='Blueste');

dumBrDale=(Neighborhood='BrDale');

dumBrkSide=(Neighborhood='BrkSide');

dumClearCr=(Neighborhood='ClearCr');

dumCollgCr=(Neighborhood='CollgCr');

dumCrawfor=(Neighborhood='Crawfor');

dumEdwards=(Neighborhood='Edwards');

dumGilbert=(Neighborhood='Gilbert');

dumIDOTRR=(Neighborhood='IDOTRR');

dumMeadowV=(Neighborhood='MeadowV');

dumMitchel=(Neighborhood='Mitchel');

dumNAmes=(Neighborhood='NAmes');

dumNPkVill=(Neighborhood='NPkVill');

dumNWAmes=(Neighborhood='NWAmes');

dumNoRidge=(Neighborhood='NoRidge');

dumNridgHt=(Neighborhood='NridgHt');

dumOldTown=(Neighborhood='OldTown');

dumSWISU=(Neighborhood='SWISU');

dumSawyer=(Neighborhood='Sawyer');

dumSawyerW=(Neighborhood='SawyerW');

dumSomerst=(Neighborhood='Somerst');

dumStoneBr=(Neighborhood='StoneBr');

dumTimber=(Neighborhood='Timber');

dumClyTile=(RoofMatl='ClyTile');

dumCompShg=(RoofMatl='CompShg');

dumMembran=(RoofMatl='Membran');

dumMetal=(RoofMatl='Metal');

dumRoll=(RoofMatl='Roll');

dumTarGrv=(RoofMatl='Tar&Grv');

dumWdShake=(RoofMatl='WdShake');

dumAbnorml=(SaleCondition='Abnorml');

dumAdjLand=(SaleCondition='AdjLand');

dumAlloca=(SaleCondition='Alloca');

dumFamily=(SaleCondition='Family');

dumNormal=(SaleCondition='Normal');

dumGrvl=(Street='Grvl');

run;

ods graphics on ;

proc reg data = totalfull3 outest=totalfullres plots(label) = (rstudentbyleverage cooksd) ;

model SalePricel = BsmtFinSF1l BsmtFullBathl Fireplacesl GarageCarsl GarageYrBltl GrLivAreal KitchenAbvGrl LotAreal MSSubClassl OverallCondl OverallQuall ScreenPorchl TotalBsmtSFl YearBuiltl dumBE1 dumBE2 dumBE3 dumBE4 dumBF1 dumBF2 dumBF3 dumBF4 dumBF5 dumBF6 dumEx dumFa dumGd dumPo dumK1 dumK2 dumK3 dumBlmngtn dumBlueste dumBrDale dumBrkSide dumClearCr dumCollgCr dumCrawfor dumEdwards dumGilbert dumIDOTRR dumMeadowV dumMitchel dumNAmes dumNPkVill dumNWAmes dumNoRidge dumNridgHt dumOldTown dumSWISU dumSawyer dumSawyerW dumSomerst dumStoneBr dumTimber dumClyTile dumCompShg dumMembran dumMetal dumRoll dumTarGrv dumWdShake dumAbnorml dumAdjLand dumAlloca dumFamily dumNormal dumGrvl / AIC VIF ;

output out = totalfullreg predicted= Predict;

run; quit;

ods graphics off;

data totalfullreg2; set totalfullreg;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

\*remove insig;

ods graphics on ;

proc reg data = totalfull3 outest=totalfullres plots(label) = (rstudentbyleverage cooksd) ;

model SalePricel = BsmtFinSF1l BsmtFullBathl Fireplacesl GarageCarsl GarageYrBltl GrLivAreal KitchenAbvGrl LotAreal MSSubClassl OverallCondl OverallQuall ScreenPorchl TotalBsmtSFl YearBuiltl dumBF1 dumBF2 dumBF3 dumEx dumGd dumPo dumK1 dumK2 dumCrawfor dumEdwards dumIDOTRR dumMeadowV dumNoRidge dumNridgHt dumSomerst dumStoneBr dumClyTile dumCompShg dumTarGrv dumWdShake dumAbnorml dumFamily dumNormal dumGrvl / AIC VIF ;

output out = totalsigreg predicted= Predict;

run; quit;

ods graphics off;

data totalsigreg2; set totalsigreg;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

data totalfull4;

set totalfull2;

dumBE1=(BsmtExposure='1');

dumBE2=(BsmtExposure='2');

dumBE3=(BsmtExposure='3');

dumBE4=(BsmtExposure='4');

dumBF1=(BsmtFinType1='1');

dumBF2=(BsmtFinType1='2');

dumBF3=(BsmtFinType1='3');

dumBF4=(BsmtFinType1='4');

dumBF5=(BsmtFinType1='5');

dumBF6=(BsmtFinType1='6');

dumHEx=(HeatingQC='Ex');

dumHFa=(HeatingQC='Fa');

dumHGd=(HeatingQC='Gd');

dumHPo=(HeatingQC='Po');

dumK1=(KitchenQual='1');

dumK2=(KitchenQual='2');

dumK3=(KitchenQual='3');

dumBlmngtn=(Neighborhood='Blmngtn');

dumBlueste=(Neighborhood='Blueste');

dumBrDale=(Neighborhood='BrDale');

dumBrkSide=(Neighborhood='BrkSide');

dumClearCr=(Neighborhood='ClearCr');

dumCollgCr=(Neighborhood='CollgCr');

dumCrawfor=(Neighborhood='Crawfor');

dumEdwards=(Neighborhood='Edwards');

dumGilbert=(Neighborhood='Gilbert');

dumIDOTRR=(Neighborhood='IDOTRR');

dumMeadowV=(Neighborhood='MeadowV');

dumMitchel=(Neighborhood='Mitchel');

dumNAmes=(Neighborhood='NAmes');

dumNPkVill=(Neighborhood='NPkVill');

dumNWAmes=(Neighborhood='NWAmes');

dumNoRidge=(Neighborhood='NoRidge');

dumNridgHt=(Neighborhood='NridgHt');

dumOldTown=(Neighborhood='OldTown');

dumSWISU=(Neighborhood='SWISU');

dumSawyer=(Neighborhood='Sawyer');

dumSawyerW=(Neighborhood='SawyerW');

dumSomerst=(Neighborhood='Somerst');

dumStoneBr=(Neighborhood='StoneBr');

dumTimber=(Neighborhood='Timber');

dumClyTile=(RoofMatl='ClyTile');

dumCompShg=(RoofMatl='CompShg');

dumMembran=(RoofMatl='Membran');

dumMetal=(RoofMatl='Metal');

dumRoll=(RoofMatl='Roll');

dumTarGrv=(RoofMatl='TarGrv');

dumWdShake=(RoofMatl='WdShake');

dumAbnorml=(SaleCondition='Abnorml');

dumAdjLand=(SaleCondition='AdjLand');

dumAlloca=(SaleCondition='Alloca');

dumFamily=(SaleCondition='Family');

dumNormal=(SaleCondition='Normal');

dumGrvl=(Street='Grvl');

dumBQ1=(BsmtQual='1');

dumBQ2=(BsmtQual='2');

dumBQ3=(BsmtQual='3');

dumBQ4=(BsmtQual='4');

dumBQ5=(BsmtQual='5');

dumCN=(CentralAir='N');

dumCEx=(ExterCond='Ex');

dumCFa=(ExterCond='Fa');

dumCGd=(ExterCond='Gd');

dumCPo=(ExterCond='Po');

dumF1=(Functional='1');

dumF2=(Functional='2');

dumF3=(Functional='3');

dumF4=(Functional='4');

dumF5=(Functional='5');

dumF6=(Functional='6');

dumF7=(Functional='7');

dumM1=(MiscFeature='1');

dumM2=(MiscFeature='2');

dumM3=(MiscFeature='3');

dumM4=(MiscFeature='4');

dumM5=(MiscFeature='5');

dumPN=(PavedDrive='N');

dumPP=(PavedDrive='P');

dumU1=(Utilities='1');

dumU2=(Utilities='2');

dumU3=(Utilities='3');

run;

\*aic;

ods graphics on ;

proc reg data = totalfull4 outest=totalfullres plots(label) = (rstudentbyleverage cooksd) ;

model SalePricel = BedroomAbvGrl BsmtFinSF1l BsmtFullBathl BsmtUnfSFl EnclosedPorchl Fireplacesl FullBathl GarageAreal GarageCarsl GarageYrBltl GrLivAreal HalfBathl KitchenAbvGrl LotAreal MiscVall MSSubClassl OverallCondl OverallQuall PoolAreal ScreenPorchl TotalBsmtSFl WoodDeckSFl X1stFlrSFl X2ndFlrSFl YearBuiltl YearRemodAddl dumBE1 dumBE2 dumBE3 dumBE4 dumBF1 dumBF2 dumBF3 dumBF4 dumBF5 dumBF6 dumHEx dumHFa dumHGd dumHPo dumK1 dumK2 dumK3 dumBlmngtn dumBlueste dumBrDale dumBrkSide dumClearCr dumCollgCr dumCrawfor dumEdwards dumGilbert dumIDOTRR dumMeadowV dumMitchel dumNAmes dumNPkVill dumNWAmes dumNoRidge dumNridgHt dumOldTown dumSWISU dumSawyer dumSawyerW dumSomerst dumStoneBr dumTimber dumClyTile dumCompShg dumMembran dumMetal dumRoll dumTarGrv dumWdShake dumAbnorml dumAdjLand dumAlloca dumFamily dumNormal dumGrvl dumBQ1 dumBQ2 dumBQ3 dumBQ4 dumBQ5 dumCN dumCEx dumCFa dumCGd dumCPo dumF1 dumF2 dumF3 dumF4 dumF5 dumF6 dumF7 dumM1 dumM2 dumM3 dumM4 dumM5 dumPN dumPP dumU1 dumU2 dumU3/ AIC VIF ;

output out = totalaicreg predicted= Predict;

run; quit;

ods graphics off;

data totalaicreg2; set totalaicreg;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

\*proc GLM conversion;

ods graphics on;

proc glm data=totalfull2;

class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType RoofStyle RoofMatl Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleCondition;

model SalePricel = MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence MiscFeature SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl MasVnrAreal BsmtFinSF1l BsmtFinSF2l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl X2ndFlrSFl LowQualFinSFl GrLivAreal BsmtFullBathl BsmtHalfBathl FullBathl HalfBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal WoodDeckSFl OpenPorchSFl EnclosedPorchl X3SsnPorchl ScreenPorchl PoolAreal MiscVall / solution;

output out = totalbackwardglm p = Predict;

run; quit;

ods graphics off;

data totalbackwardglm2; set totalbackwardglm;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

**50 imputation**

FILENAME REFFILE '/home/yaoy890/proj 1/total50.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=imp50;

GETNAMES=YES;

RUN;

proc sgscatter data=imp50;

matrix SalePrice MSSubClass LotFrontage LotArea OverallQual OverallCond / diagonal=(histogram) ;

run;

proc sgscatter data=imp50;

matrix SalePrice YearBuilt YearRemodAdd BsmtFinSF1 BsmtUnfSF TotalBsmtSF / diagonal=(histogram) ;

run;

proc sgscatter data=imp50;

matrix SalePrice X1stFlrSF GrLivArea FullBath BedroomAbvGr KitchenAbvGr / diagonal=(histogram) ;

run;

proc sgscatter data=imp50;

matrix SalePrice TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea OpenPorchSF MoSold YrSold / diagonal=(histogram) ;

run;

data imp502;

set imp50;

SalePricel=log(SalePrice+1);

MSSubClassl=log(MSSubClass+1);

LotFrontagel=log(LotFrontage+1);

LotAreal=log(LotArea+1);

OverallQuall=log(OverallQual+1);

OverallCondl=log(OverallCond+1);

YearBuiltl=log(YearBuilt+1);

YearRemodAddl=log(YearRemodAdd+1);

BsmtFinSF1l=log(BsmtFinSF1+1);

BsmtUnfSFl=log(BsmtUnfSF+1);

TotalBsmtSFl=log(TotalBsmtSF+1);

X1stFlrSFl=log(X1stFlrSF+1);

GrLivAreal=log(GrLivArea+1);

FullBathl=log(FullBath+1);

BedroomAbvGrl=log(BedroomAbvGr+1);

KitchenAbvGrl=log(KitchenAbvGr+1);

TotRmsAbvGrdl=log(TotRmsAbvGrd+1);

Fireplacesl=log(Fireplaces+1);

GarageYrBltl=log(GarageYrBlt+1);

GarageCarsl=log(GarageCars+1);

GarageAreal=log(GarageArea+1);

OpenPorchSFl=log(OpenPorchSF+1);

MoSoldl=log(MoSold+1);

YrSoldl=log(YrSold+1);

run;

proc sgscatter data=imp502;

matrix SalePricel MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl / diagonal=(histogram) ;

run;

proc sgscatter data=imp502;

matrix SalePricel YearBuiltl YearRemodAddl BsmtFinSF1l BsmtUnfSFl TotalBsmtSFl / diagonal=(histogram) ;

run;

proc sgscatter data=imp502;

matrix SalePricel X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl / diagonal=(histogram) ;

run;

proc sgscatter data=imp502;

matrix SalePricel TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal OpenPorchSFl MoSoldl YrSoldl / diagonal=(histogram) ;

run;

\*forward;

ods graphics on;

proc glmselect data=imp502 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

partition fraction(validate = 0.3 test = 0.2);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtFinSF1l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal OpenPorchSFl MoSoldl YrSoldl / selection=FORWARD(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp50forward p = Predict;

run; quit;

ods graphics off;

data imp50forward2; set imp50forward;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

ods graphics on;

proc glmselect data=imp502 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtFinSF1l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal OpenPorchSFl MoSoldl YrSoldl / selection=FORWARD(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp50forwardCV p = Predict;

run; quit;

ods graphics off;

data imp50forwardCV2; set imp50forwardCV;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

\*BACKWARD;

ods graphics on;

proc glmselect data=imp502 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

partition fraction(validate = 0.3 test = 0.2);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtFinSF1l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal OpenPorchSFl MoSoldl YrSoldl / selection=BACKWARD(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp50backward p = Predict;

run; quit;

ods graphics off;

data imp50backward2; set imp50backward;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

ods graphics on;

proc glmselect data=imp502 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtFinSF1l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal OpenPorchSFl MoSoldl YrSoldl / selection=BACKWARD(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp50backwardCV p = Predict;

run; quit;

ods graphics off;

data imp50backwardCV2; set imp50backwardCV;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

\*STEPWISE;

ods graphics on;

proc glmselect data=imp502 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

partition fraction(validate = 0.3 test = 0.2);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtFinSF1l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal OpenPorchSFl MoSoldl YrSoldl / selection=STEPWISE(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp50stepwise p = Predict;

run; quit;

ods graphics off;

data imp50stepwise2; set imp50stepwise;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

ods graphics on;

proc glmselect data=imp502 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtFinSF1l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal OpenPorchSFl MoSoldl YrSoldl / selection=STEPWISE(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp50stepwiseCV p = Predict;

run; quit;

ods graphics off;

data imp50stepwiseCV2; set imp50stepwiseCV;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

\*LAR;

ods graphics on;

proc glmselect data=imp502 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

partition fraction(validate = 0.3 test = 0.2);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtFinSF1l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal OpenPorchSFl MoSoldl YrSoldl / selection=LAR(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp50lar p = Predict;

run; quit;

ods graphics off;

data imp50lar2; set imp50lar;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

ods graphics on;

proc glmselect data=imp502 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtFinSF1l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal OpenPorchSFl MoSoldl YrSoldl / selection=LAR(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp50larCV p = Predict;

run; quit;

ods graphics off;

data imp50larCV2; set imp50larCV;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

\*LASSO;

ods graphics on;

proc glmselect data=imp502 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

partition fraction(validate = 0.3 test = 0.2);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtFinSF1l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal OpenPorchSFl MoSoldl YrSoldl / selection=LASSO(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp50lasso p = Predict;

run; quit;

ods graphics off;

data imp50lasso2; set imp50lasso;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

ods graphics on;

proc glmselect data=imp502 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtFinSF1l BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl Fireplacesl GarageYrBltl GarageCarsl GarageAreal OpenPorchSFl MoSoldl YrSoldl / selection=LASSO(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp50lassoCV p = Predict;

run; quit;

ods graphics off;

data imp50lassoCV2; set imp50lassoCV;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

**75 imputation**

FILENAME REFFILE '/home/yaoy890/proj 1/total75.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=imp75;

GETNAMES=YES;

RUN;

proc sgscatter data=imp75;

matrix SalePrice MSSubClass LotFrontage LotArea OverallQual OverallCond / diagonal=(histogram) ;

run;

proc sgscatter data=imp75;

matrix SalePrice YearBuilt YearRemodAdd BsmtUnfSF TotalBsmtSF X1stFlrSF / diagonal=(histogram) ;

run;

proc sgscatter data=imp75;

matrix SalePrice GrLivArea FullBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd / diagonal=(histogram) ;

run;

proc sgscatter data=imp75;

matrix SalePrice GarageYrBlt GarageCars GarageArea MoSold YrSold / diagonal=(histogram) ;

run;

data imp752;

set imp75;

MSSubClassl=log(MSSubClass+1);

LotFrontagel=log(LotFrontage+1);

LotAreal=log(LotArea+1);

OverallQuall=log(OverallQual+1);

OverallCondl=log(OverallCond+1);

YearBuiltl=log(YearBuilt+1);

YearRemodAddl=log(YearRemodAdd+1);

BsmtUnfSFl=log(BsmtUnfSF+1);

TotalBsmtSFl=log(TotalBsmtSF+1);

X1stFlrSFl=log(X1stFlrSF+1);

GrLivAreal=log(GrLivArea+1);

FullBathl=log(FullBath+1);

BedroomAbvGrl=log(BedroomAbvGr+1);

KitchenAbvGrl=log(KitchenAbvGr+1);

TotRmsAbvGrdl=log(TotRmsAbvGrd+1);

GarageYrBltl=log(GarageYrBlt+1);

GarageCarsl=log(GarageCars+1);

GarageAreal=log(GarageArea+1);

MoSoldl=log(MoSold+1);

YrSoldl=log(YrSold+1);

SalePricel=log(SalePrice+1);

run;

proc sgscatter data=imp752;

matrix SalePricel MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl / diagonal=(histogram) ;

run;

proc sgscatter data=imp752;

matrix SalePricel YearBuiltl YearRemodAddl BsmtUnfSFl TotalBsmtSFl X1stFlrSFl / diagonal=(histogram) ;

run;

proc sgscatter data=imp752;

matrix SalePricel GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl / diagonal=(histogram) ;

run;

proc sgscatter data=imp752;

matrix SalePricel GarageYrBltl GarageCarsl GarageAreal MoSoldl YrSoldl / diagonal=(histogram) ;

run;

\*forward;

ods graphics on;

proc glmselect data=imp752 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

partition fraction(validate = 0.3 test = 0.2);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl GarageYrBltl GarageCarsl GarageAreal MoSoldl YrSoldl / selection=FORWARD(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp75forward p = Predict;

run; quit;

ods graphics off;

data imp75forward2; set imp75forward;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

ods graphics on;

proc glmselect data=imp752 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl GarageYrBltl GarageCarsl GarageAreal MoSoldl YrSoldl / selection=FORWARD(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp75forwardCV p = Predict;

run; quit;

ods graphics off;

data imp75forwardCV2; set imp75forwardCV;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

\*BACKWARD;

ods graphics on;

proc glmselect data=imp752 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

partition fraction(validate = 0.3 test = 0.2);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl GarageYrBltl GarageCarsl GarageAreal MoSoldl YrSoldl / selection=BACKWARD(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp75backward p = Predict;

run; quit;

ods graphics off;

data imp75backward2; set imp75backward;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

ods graphics on;

proc glmselect data=imp752 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl GarageYrBltl GarageCarsl GarageAreal MoSoldl YrSoldl / selection=BACKWARD(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp75backwardCV p = Predict;

run; quit;

ods graphics off;

data imp75backwardCV2; set imp75backwardCV;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

\*STEPWISE;

ods graphics on;

proc glmselect data=imp752 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

partition fraction(validate = 0.3 test = 0.2);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl GarageYrBltl GarageCarsl GarageAreal MoSoldl YrSoldl / selection=STEPWISE(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp75stepwise p = Predict;

run; quit;

ods graphics off;

data imp75stepwise2; set imp75stepwise;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

ods graphics on;

proc glmselect data=imp752 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl GarageYrBltl GarageCarsl GarageAreal MoSoldl YrSoldl / selection=STEPWISE(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp75stepwiseCV p = Predict;

run; quit;

ods graphics off;

data imp75stepwiseCV2; set imp75stepwiseCV;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

\*LAR;

ods graphics on;

proc glmselect data=imp752 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

partition fraction(validate = 0.3 test = 0.2);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl GarageYrBltl GarageCarsl GarageAreal MoSoldl YrSoldl / selection=LAR(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp75lar p = Predict;

run; quit;

ods graphics off;

data imp75lar2; set imp75lar;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

ods graphics on;

proc glmselect data=imp752 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl GarageYrBltl GarageCarsl GarageAreal MoSoldl YrSoldl / selection=LAR(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp75larCV p = Predict;

run; quit;

ods graphics off;

data imp75larCV2; set imp75larCV;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

\*LASSO;

ods graphics on;

proc glmselect data=imp752 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

partition fraction(validate = 0.3 test = 0.2);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl GarageYrBltl GarageCarsl GarageAreal MoSoldl YrSoldl / selection=LASSO(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp75lasso p = Predict;

run; quit;

ods graphics off;

data imp75lasso2; set imp75lasso;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;

ods graphics on;

proc glmselect data=imp752 seed=1 plots(stepAxis=number)=(criterionPanel ASEPlot ASE);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition;

model SalePricel = MSZoning Street LotShape LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType SaleCondition MSSubClassl LotFrontagel LotAreal OverallQuall OverallCondl YearBuiltl YearRemodAddl BsmtUnfSFl TotalBsmtSFl X1stFlrSFl GrLivAreal FullBathl BedroomAbvGrl KitchenAbvGrl TotRmsAbvGrdl GarageYrBltl GarageCarsl GarageAreal MoSoldl YrSoldl / selection=LASSO(choose=CV stop=AIC) CVdetails showpvalues;

output out = imp75lassoCV p = Predict;

run; quit;

ods graphics off;

data imp75lassoCV2; set imp75lassoCV;

SalePrice = exp(Predict)-1;

keep id SalePrice;

where id > 1460;

;