ST 516 MID TERM PROJECT
FALL 2021

SUBMITTED BY:

ARUN ARUNACHALAM MAHESWARAN (200379075)

CHINMAY KURDEKAR (200379150)

ASHWATH MUKUNDAN (200382936)

SAI PRASATH KUPPUSAMY JAYASREE (200377777)

AMEYA VIRENDRA PARALKAR (200371690)

OCTOBER 7TH 2021

EXECUTIVE SUMMARY

This Report investigates the relationship between various attributes of houses in order to create a predictive model that accurately predicts the purchasing price for the houses. The Report also identifies a handful of attributes that have the most significant association with house prices as well as an estimate of the quantitative relationship between those attributes and house price. A Historical data from the 2000's provided by Ames, Iowa association of realtors was determined to fit most accurately for a Best subset selection model with k-fold cross validation. This Model was evaluated and arrived as the best prediction model since it could account for 91.84% of the variations in the data with a total of 24 best subset predictors. Some of the most significant attributes that influenced the Purchase prices were identified to be GrLivArea, OverallCond, OverallQual, YearBuilt and TotalBsmtSF. The Root Mean Squared Error (RMSE) for the Model was found to be one of the least among the other models, thereby giving minimum prediction error and maintaining the right balance of interpretability and accuracy.

INTRODUCTION

Ames, Iowa association of realtors want to understand if there are any significant differences which drives the housing prices before and after the economic crash. The task is to work with the data and run different models to analyze which parameters out of the many in the dataset is driving the housing prices and suggest a model which can accurately predict the housing prices in the future and also to have an understanding on how these attributes relate to each other.

In order to accurately predict the Housing prices, the data was first preprocessed to adjust for the Null values in the data set and also to normalize the sale prices using log transform. The Categorical variables were transformed using model.matrix() function and then we proceed to work on building a predictive model which will help us with the objectives. We run different models such as linear (normal, reduced), regularized models (ridge, lasso) , best subset selection (forward, backward, k-fold cross validation) and tree based models. To select the best model, we find out the Root Mean squared error (RMSE) to find out how the model performs on the test data, R-squared to find out how well it explains the variance and the number of predictors used in the model to understand the complexity. Also, Diagnostic plots were plotted whenever needed to understand the model more. Correlation and Significance of the models were also found out in selective models to understand relationship between the parameters

DATA

The Ames housing data set consists of 1437 observations and 25 variables (15 Numerical and 10 Categorical variables) to predict the SalesPrice. The dataset had 259, 36 and 77 missing values in LotFrontage, BsmtFinType1 and GarageType categories respectively. The total missing values between these attributes contribute to around 12% of the entire dataset observations and removing all the observations with missing values will lead to a considerable loss of data which might affect the prediction capability of the

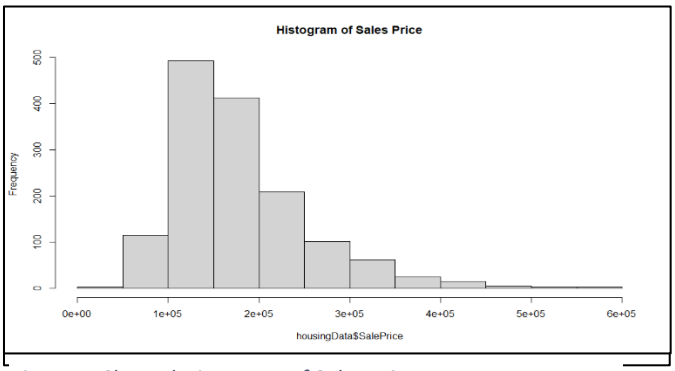


Figure 1: Skewed Histogram of Sales Price

model. Hence, the average of LotFrontage is assigned to the observations where the values are missing. As the BsmtFinType1 and GarageType are categorical variables, the observation with missing values is removed since aggregation is not possible. The response variable 'SalesPrice' was ***found to have a right skewed data***. The 'SalesPrice' variable is ***subjected to log transformation to*** compensate for the skewness that was evident before the transformation. The Scatter plot grid for Numerical attributes in the given data was investigated along with the Heat Map of attributes. These graphs helped in identifying the presence of collinearity among the predictors and infer the extent of correlation of the attributes with Sale Price.

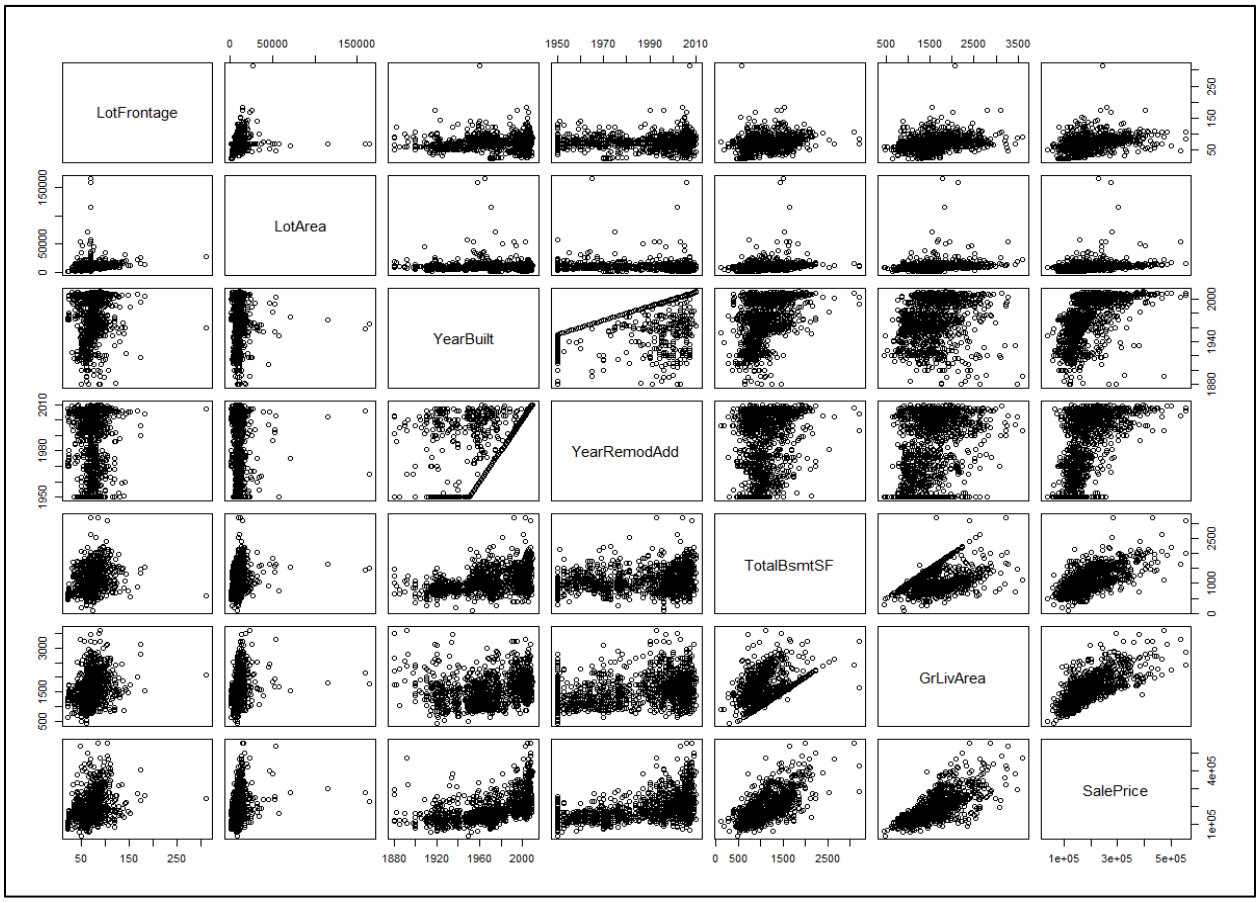


Figure 3: Scatter Plot of Numerical attributes in the given dataset

The correlation matrix was plotted for the numerical data in the database. It was done to see if there is any collinearity in-between the data points and to form a hypothesis on which numerical parameters in the model affects the sale prices heavily.

We can see the sale prices are influenced heavily by the parameters like YearBuilt, YearRemodAdd, TotalBsmtSF, GrtLivArea from the correlation matrix.

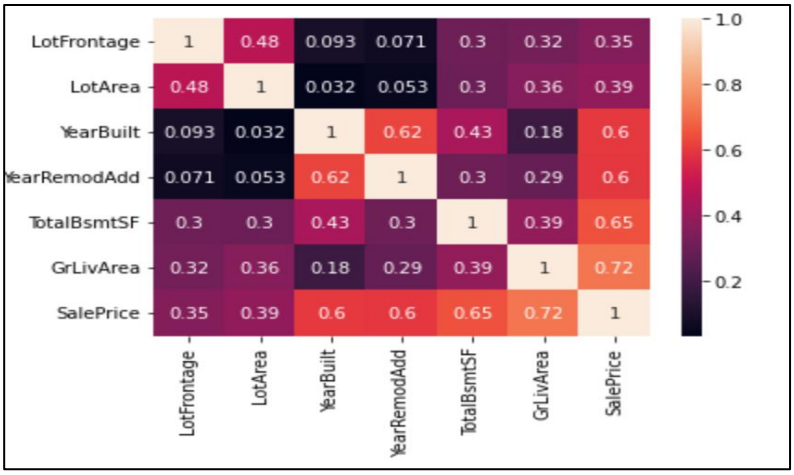
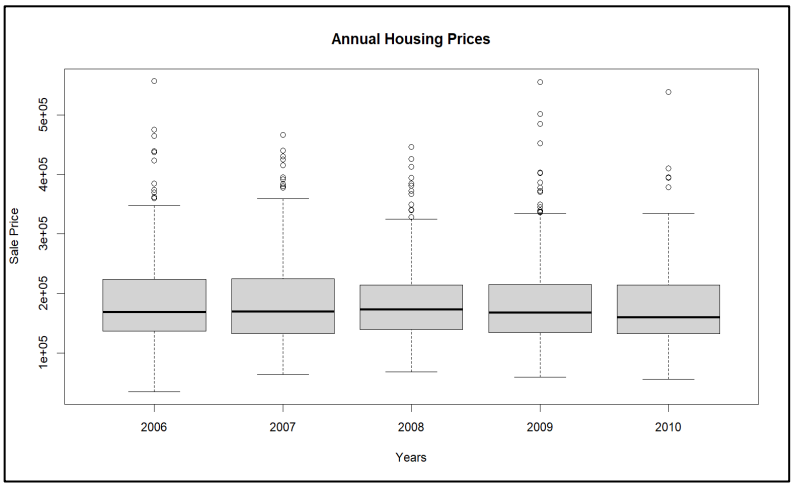


Figure 4: Heat Map Correlation of Different Numerical attributes

Also box plot was plotted for the Sale Prices through different years to see how the economic crash affected it. It was observed that the Lowest house prices are reducing post the crash and the median housing prices also seem to go down. This vaguely explains impact of the economic crash



METHODS

Different predictive model approaches have been used with all the variables taken into consideration. Based on the further results from each model, the variables which are insignificant were dropped. The models are of different methods including Linear, Best Subset, Regularisation and Tree-based.

Linear Model - Complete :

The initial model which was run on the given data after the preprocessing steps was the Linear Regression model with the response variable as SalePrice and the remaining 87 Variables as predictors (including dummy columns for Categorical variables). The Idea of the Regression model was to use its ability to predict the continuous response variable given the relationship between the predictor variables and the response. The result of the regression model was positive. The Obtained R-squared value was 0.9302 and the RMSE was 21532. The summary of the model gives us the important predictors based on the p-value. We find out that the hypothesis we had from the correlation matrix and the significant from this model are similar, thereby validation the hypothesis. The Diagnostic Plots indicate that the Data is Normally distributed and that the Variance is constant & centered around the Zero mean line.

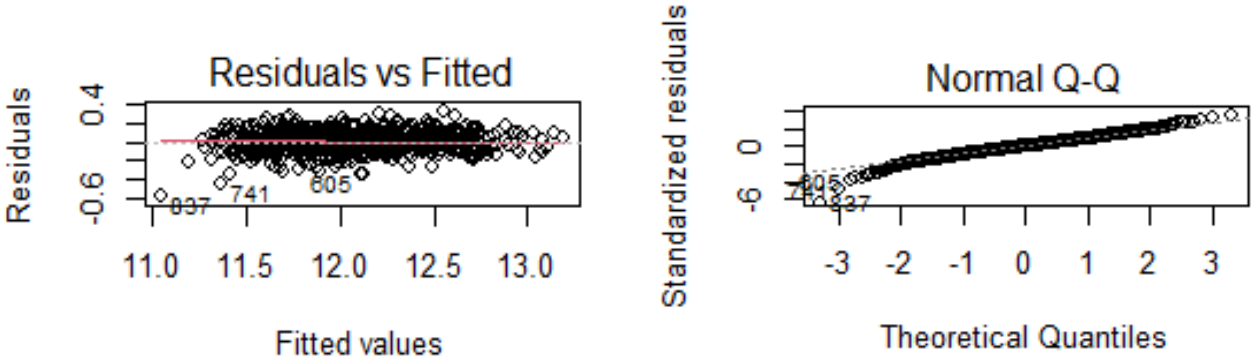


Figure 5: Diagnostic Plots for Linear complete model

Although this model gives us good results which can predict the sale prices in a reasonable way. The number of predictors used in this model is 87 which is very high and not every one of that was significantly affecting the response. So the next step was to sort the significant predictors and fit a new model only with those Predictors.

Linear regression - Reduced

The summary function was used to sort the significant predictors based on the p-value. We find out that only 31 variables are important predictors. So, we decided to drop the rest of the variables and ran a regression model with only these 31 predictors and SalePrice as response. This resulted in a better RMSE value and a slightly less R^2 value which was expected because the model explains less variability than the default model. Comparing the prediction accuracy with the complete linear model, it is observed that there is no significant difference and therefore the prediction accuracy is not compromised in spite of reducing the number of predictors in this model.

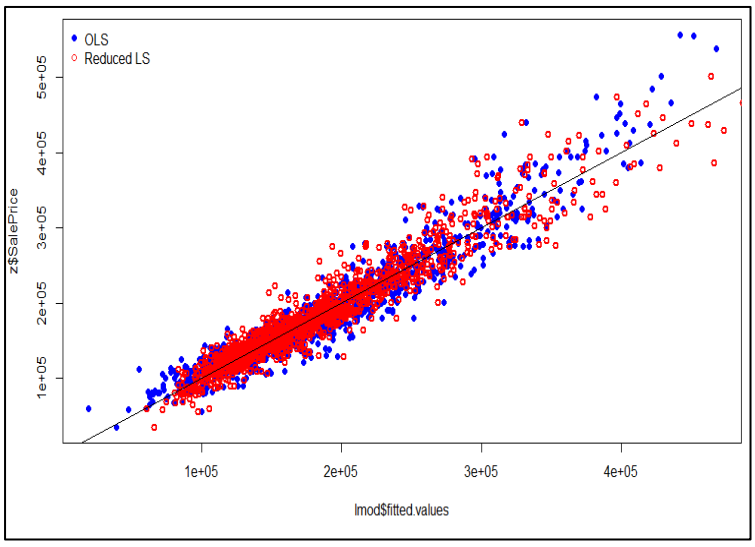


Figure 6: Actual vs Fitted data for Complete and Reduced Linear Models

The Linear model is fairly a simple model and we need to see if we can use the learnings from this and fit a more complex model which can help us predict the prices better with an even lesser RMSE. Although Linear models gave us good results, there is always a fear of overfitting the model and also the multi collinearity between the predictor variables which will give us false understanding. To remove this, the next regression models were applied

Regularization - Ridge & Lasso Models

There are some known disadvantages with the linear model which were discussed above. To remove these, Regularization was done. Regularization is expected to perform better because it doesn't require unbiased estimators; While complete least squares produces unbiased estimates, variances can be so large that the Model can result in Overfitting the data. Ridge regression adds just enough bias to make the estimates reasonably reliable approximations to

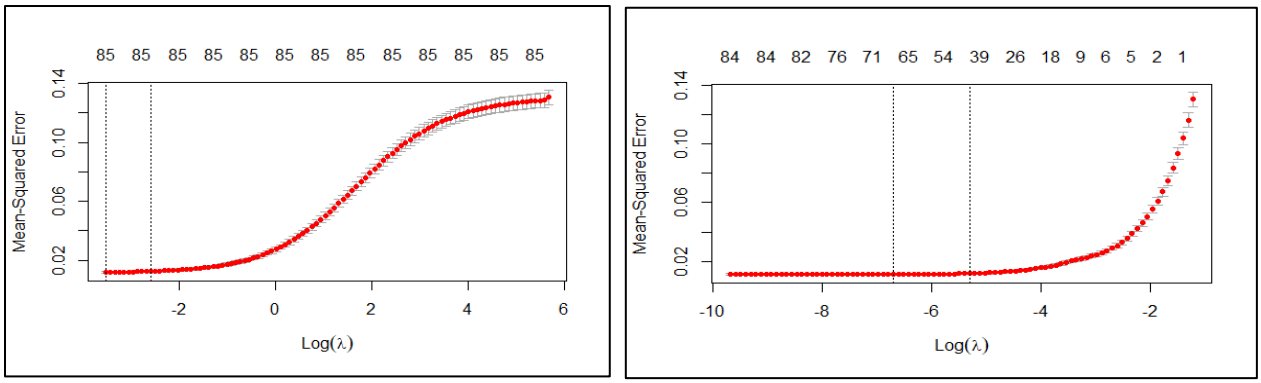


Figure 7: CV for Lambda - Ridge and Lasso Models

true population values. Ridge regression model was run on the model with 31 predictor variables instead of the original data after ignoring the insignificant predictors. The R^2 values for Ridge and Lasso models were estimated to be 0.9201 and 0.9229 and the RMSE values were 21813 and 21437 respectively. The regularization models gave a better RMSE value than the linear models, which indicates that L1 and L2 penalties applied is accounting for the slight overfitting in the linear data.

Although the regularization models gave us a better result, The model does not accommodate for splitting the testing and training data. There is a chance where based on coincidence, the model might split the data into a perfect test and train set which will not reflect the future data from the client. To remove such biases and the possibility of coincidence, we should run this data in a Cross validation model which addresses this exact problem.

K-fold Cross Validation

After the regularization models, we try to use k-fold cross validation, because we know that this model will help remove the training and test biases and also ensures that all the data is equally used for training and testing data. We figured that k-fold cross validation will be a good starting point for this data set. The K value was set to 5 meaning there were 5 different splits and hence 5 combinations of training and testing data sets.

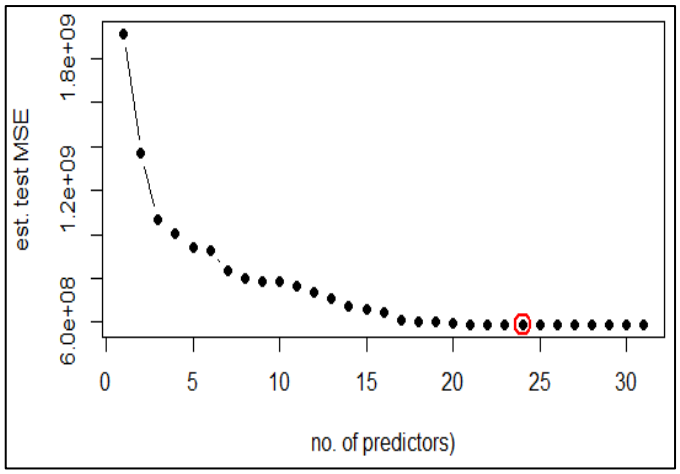


Figure 8: Estimating number of predictors in CV

The Number of predictors was chosen based on the minimum MSE value which was a result of cross validation using 5 folds. The K-fold cross validation resulted in a R^2 value of 0.9184 and RMSE of 22583 which is relatively better than the other models estimated so far. The Q-Q Plot in the Diagnostic plots fall in a straight line roughly affirming that the data is normally distributed. The Residuals vs Fitted graph indicate that the variances of the error terms are equal. The Cooks plot suggests that there are no influential points in the data thereby suggesting that the overall Cross validation model does not have any abnormalities to be addressed.

Best Subset : Forward and Backward Stepwise Selection

Stepwise regression is a way of selecting important variables to get a simple and easily interpretable model.

The forward stepwise selection begins with a model that contains no variables (called the *Null Model*). It then starts adding the most significant variables one after the other until a pre-specified stopping rule is reached or until all the variables under consideration are included in the model. We let AIC choose the threshold since it is best for prediction.

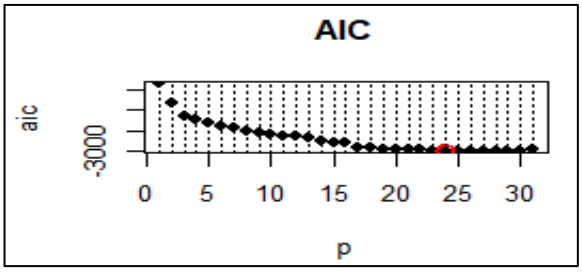


Figure 9: Estimating number of predictors using AIC

Backward Stepwise Selection begins with a model that contains all variables under consideration (called the *Full Model*). It then starts removing the least significant variables one after the other until a pre-specified stopping rule is reached or until no variable is left in the model.. Similar to forward stepwise selection, we let AIC choose the threshold. The Forward and Backward Stepwise selection models gave a R^2 of 0.9199 and RMSE values of 22202 and 22180 respectively. The Model gives a good balance of both accuracy and interpretability which are evident from the R^2 and the RMSE Values along with relatively lower number of predictor variables (24) ; Between Forward and Backward selection, Forward Selection is the preferred model as observed from the output

Tree Based Models

To explore models with easy interpretation and visualization and also to identify the most significant variables quickly, we decided to investigate the data with Tree based models.

Regression Trees :

Using the regression tree model, the data is split into 11 sets primarily by the predictors "OverallQual", "GrLivArea", "YearBuilt", LotArea", "TotalBsmtSF" and "OverallCond". The data is split by binary recursive method so that the data is bifurcated to two major portions by a single predictor. The complexity of the tree is chosen by cross-validation method and it is found that tree of complexity of 9 is found to have lowest MSE value. The RMSE is found to be 39932 which is significantly higher when compared to the models used previously.

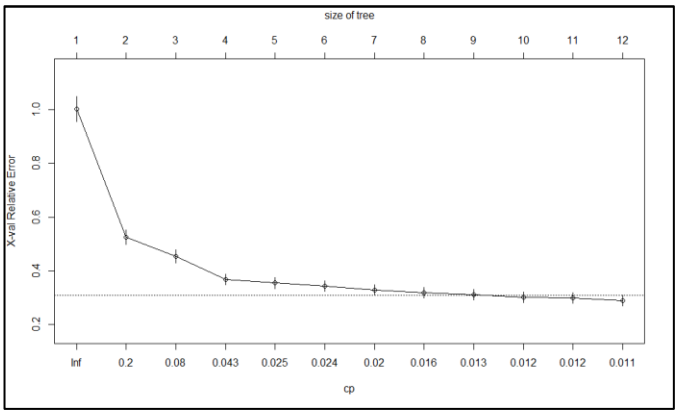


Figure 10: Estimating complexity in tree based regression

Pruned Regression Trees:

The regression tree model splits the data into 11 porti ons with 29 nodes. In order to reduce the complexity of the tree, pruning is done to lower the number of nodes and splits. However the RMSE value of 43677, is not an improvement from the regression tree model without pruning.

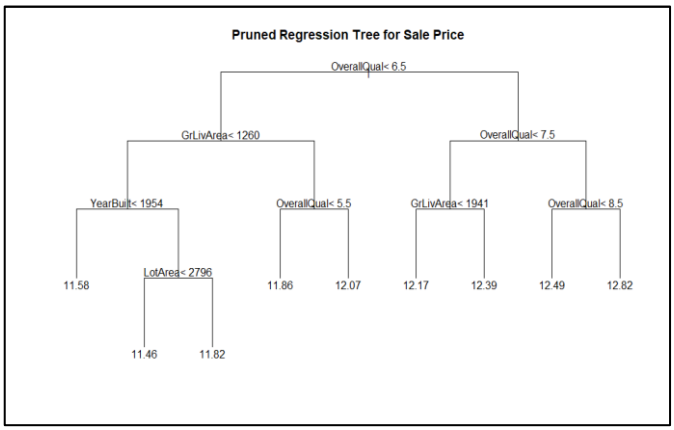


Figure 11:: Pruned Regression Tree

Bagging approach:

In bagging approach the tree is built by using 10 random predictors at each split and nearly 2000 trees are generated and the trees are aggregated in such a way that the MSE is minimum. The RMSE is found to be 27876 which is an *improvement from Pruned Regression tree Model. However, the RMSE is still higher as compared to the previously evaluated models.*

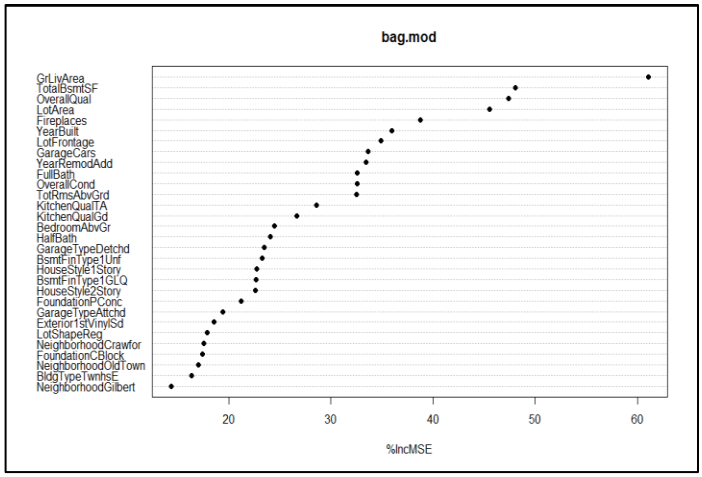


Figure 12: Importance of Predictors in Bagging Approach

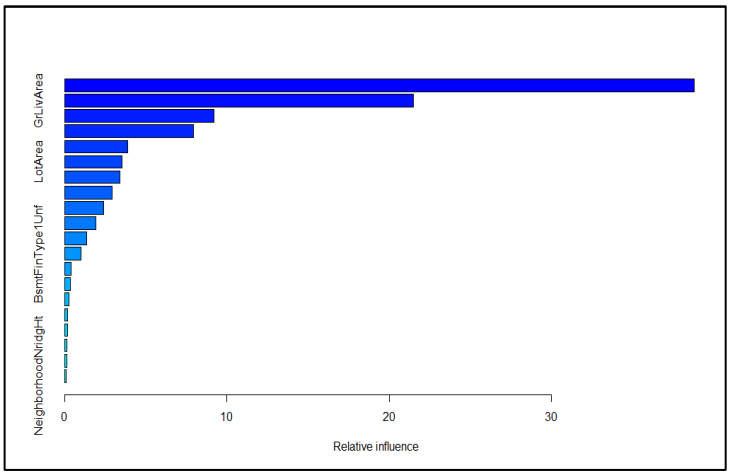


Figure 13: Importance of Predictors in Gradient Boosting

Gradient Boosting:

The Model is tuned for the hyper parameters n.trees, interaction.depth, shrinkage, n.minobsinnode using grid search. It is found that gradient boosting model yields minimum MSE when entries is 5000, shrinkage is 0.005, interaction depth is 3 and n.minobsinnode is 1.The Summary function of the gradient boosting model shows that OverallQual, GrLivArea, TotalBsmtSF, YearBuilt contribute to nearly 60% of the variations in SalesPrice.

RESULTS :

S.No	Model	Model Type	R^2	Actual RMSE	Predictors	Remarks
1	Linear Model	Complete	0.9302	21532	87	Fitting Linear Model with all predictor variables
2	Linear Model	Reduced	0.9208	22583	31	Fitting Linear Model considering significant predictors of Complete Linear model
3	Linear Model	Ridge	0.9201	21813	85	Regularizing linear model with Ridge method
4	Linear Model	Lasso	0.92294	21437	44	Regularizing linear model with Lasso method
5	Cross Validation	k Folds exhaustive	0.9184	22583	24	5 fold cross validation with significant predictors
6	Best Subset	Forward Stepwise Selection	0.9199	22202	24	Best Subset Selection using Forward Stepwise with all predictors
7	Best Subset	Backward Stepwise Selection	0.9199	22180	26	Best Subset Selection using Backward Stepwise with all predictors
8	Tree Based	Regression Tree	0.729915	39932		Regression tree Model without pruning
9	Tree Based	Regression Tree With Pruning	0.737704	43677		Regression tree Model with pruning
10	Tree Based	Regression tree with Bagging	0.897651	27876		Regression Tree Model with Bagging approach
12	Tree Based	Regression Tree with Gradient Boosting	0.9198795	20839		Regression Tree Model with Gradient Boosting

Table 1: Summary of Results

The Data set provided was predicted using various models as listed above. The Primary comparison between these models were assessed by its prediction accuracy, Interpretability and complexity of the model. The Prediction Accuracy was correlated from the RMSE values obtained from the models. The Minimum RMSE was observed for Tree based gradient Boosting model with Hyper parameter tuning followed by the Linear Models. However, since the number of predictors in the linear models were relatively high, Linear Models are not preferred for

estimating Sale Price. In terms of Interpretability, We look specifically at the R^2 values which explains the variance of the response and also the number of predictors used by the model.

As seen from the Summary table, there is a good balance between R^2 values and acceptable error estimates (in the form of RMSE) together with number of predictor variables in specifically two models namely, Regression tree based Gradient boosting and k-fold cross validation best subset model. Considering the complexity of the model and the computational efficiency between the two models, the k-fold cross validation model has a minor advantage over the tree based model. Although there is a minor loss in the RMSE and R^2 values, The k-fold cross validation predicts the Sale Prices accurately by making use of both the Train and Testing data efficiently. While both k-Fold cross validation and Tree based Gradient Boosting are the preferred predictive models from our analysis, We would prefer to go with k-fold cross validation considering that it is relatively less computationally intensive.

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.511e+00  5.528e-01   4.542 6.24e-06 ***
OverallQual          6.117e-02  4.542e-03  13.466  < 2e-16 ***
OverallCond          5.655e-02  3.875e-03  14.593  < 2e-16 ***
GrLivArea            2.580e-04  9.507e-06  27.134  < 2e-16 ***
YearBuilt            3.492e-03  2.205e-04  15.840  < 2e-16 ***
TotalBsmtSF          1.351e-04  1.140e-05  11.848  < 2e-16 ***
BsmtFinType1Unf     -6.898e-02  7.501e-03  -9.196  < 2e-16 ***
KitchenQualGd       -7.464e-02  1.571e-02  -4.751 2.31e-06 ***
GarageCars           4.093e-02  7.301e-03   5.606 2.66e-08 ***
KitchenQualFa       -1.462e-01  3.348e-02  -4.365 1.40e-05 ***
Fireplaces           4.047e-02  6.094e-03   6.640 5.06e-11 ***
BldgTypeTwnhs       -1.413e-01  2.197e-02  -6.433 1.92e-10 ***
KitchenQualTA       -8.166e-02  1.811e-02  -4.510 7.22e-06 ***
LotArea              2.635e-06  3.916e-07   6.728 2.85e-11 ***
NeighborhoodStoneBr  1.724e-01  2.737e-02   6.299 4.42e-10 ***
BldgTypeTwnhsE      -7.954e-02  1.424e-02  -5.584 3.01e-08 ***
BldgTypeDuplex      -1.207e-01  2.236e-02  -5.399 8.32e-08 ***
Exterior1stBrkFace   9.291e-02  1.942e-02   4.784 1.97e-06 ***
NeighborhoodMeadowV -9.475e-02  3.976e-02  -2.383 0.017365 *
Exterior1stMetalSd   4.267e-02  1.151e-02   3.708 0.000220 ***
NeighborhoodCrawfor  1.293e-01  1.809e-02   7.150 1.64e-12 ***
NeighborhoodNridgHt  7.662e-02  1.667e-02   4.596 4.84e-06 ***
HouseStyleSLvl       3.538e-02  1.511e-02   2.341 0.019408 *
FoundationPConc      4.356e-02  1.074e-02   4.057 5.35e-05 ***
Exterior1stVinylSd   2.856e-02  1.088e-02   2.624 0.008824 **
LotFrontage          3.833e-04  1.887e-04   2.031 0.042499 *
Exterior1stCemntBd   2.309e-02  2.208e-02   1.046 0.295916
Exterior1stWd.Sdng   2.276e-02  1.224e-02   1.859 0.063333 .
Exterior1stPlywood   3.902e-02  1.470e-02   2.654 0.008074 **
Exterior1stWdShing   4.096e-02  2.720e-02   1.506 0.132411
YearRemodAdd         6.591e-04  2.590e-04   2.545 0.011077 *
NeighborhoodSomerst  5.792e-02  1.494e-02   3.878 0.000112 ***
```

Figure 14: Significant predictors from the concluded Models

In terms of the significant predictors, Both the models have very similar results. We observe that OverallQual, OverallCond, grLivArea, YearBuilt and TotalBsmtSF are the key predictors of the Sale Price.

CONCLUSION

In this project, we analyzed the Ames housing data to build a predictive model and understand how different variables influence the Sales Price of the houses in the region. The emphasis was on finding a model with best prediction accuracy and good interpretability. Initially, Exploratory Analysis was done to understand more about the database and the variables. After accounting for the null values, we used model.matrix command to convert categorical variables to dummy variables. A total of 87 levels were generated by using model.matrix command (categorical and numerical data included). It was found that the SalePrices which is the response, was skewed to the right and therefore, Log transform was done to normalize the response.

First, Linear regression model was built with all 87 levels as predictors. 31 significant predictors (95% confidence) were filtered out and this reduced linear model was used as the base for further analysis. Regularization models were then used to account for any overfitting or collinearity in the data. There was a major overlap between the significant predictors obtained from the reduced linear model and the predictors obtained from the Lasso model. Finally, tree based models such as Bagging, Boosting and RandomForest were also built to optimize the model.

K-Fold Cross Validation and Tree Based Gradient Boosting Models were determined to best fit for the dataset. K-fold Cross validation resulted in a R^2 value of 0.9184 and RMSE of 22583 using 24 Predictors and Tree based Gradient Boosting model resulted in 0.9198 and RMSE of 20839. These models were chosen because of its best prediction capability (least test MSE) and its ability to explain 91.84% and 91.98% of variability in the model. K-fold cross validation and tree based gradient boosting models also depicted a good balance between predictive accuracy and interpretability.

The Recommendation to Ames, Iowa Association of Realtors would be to go with a k-fold cross validation Model to predict Housing Prices. If Computation power permits, Tree based gradient boosting model can be preferred. As a future scope of the project, these models can be used to estimate and understand Housing Prices post economic crash. Also, various other statistical techniques such as Hypothesis Testing can be used to check how significant is the difference between the Sale Price before and after the housing market crash.