Applying Predictive Analytics for Housing Market - House Price predictions

Scott Breitbach, Pushkar Chougule, and Rachel Nelson

Bellevue University

Executive Summary

After the impact of COVID-19 pandemic, the economy is looking to get back on track. One of the indicators of recovery is the housing market demand / supply and prices. With so many buyers in the market, with different level of needs – one common question everyone has on their mind is - *Is this house priced right?* Due to the current turbulent market conditions, it is hard to know what is considered a “good deal”. Whether looking to buy your first home, sell a home or invest in a rental property, today’s housing market moves fast and being able to gauge the fairness of the deal, it would be great to have a ready reckoner to go to. So, we decided to devise a mechanism a.k.a. prediction system, to aid the potential buyers in the market with appropriate level of guidance in terms of appropriate pricing levels of the house they would be interested in buying / selling the property. By leveraging power of modern predictive analytics, we have developed system to predict how to price one’s home, or what offer to put in, which can be extremely helpful.

The prices are determined by variety of factors, each of them contributing to the housing prices in varying magnitude. Hence, in order to be able to predict home prices, we first identified which factors affect the Sale price of a house to the greater degree. To determine these factors, we used a variety of techniques available within Predictive Analytics world and generated predictive system, comprising of multiple Machine Learning models, to bring the idea into real.

Initially we aimed to devise a system with over 80% accuracy levels. With meticulous preparation and trials of system consisting of variety of Machine Learning models, we did achieve the prediction accuracy of 88% in predicting a home price. By predicting house prices based on factors, we can compare our predicted price to the asking price of a home for sale to help us place a proper bid on a house.

Technical Report

Background

Following the impact of the COVID-19 pandemic, the economy is looking to get back on track. One of the leading indicators of recovery is the housing market demand, supply and prices. In this paper, we discuss the methods and results of a detailed preliminary analysis of the housing market and determine what all different factors, if any, contribute to the pricing of housing. We will leverage the findings of our analysis to create a housing price prediction model and compare models to select our final model.

Data Understanding: Defining the Data

In order to go about finding a proper data set, we looked at different options including scraping data from real estate websites like Zillow or Trulia or reviewing various existing data sets already available online.

While scraping Zillow data could provide us up-to-date information, it also presents its own challenges in obtaining a consistent set of features about the homes. The data available from initial scraping was home price, number of bedroom and bathrooms, square footage of the home and the home address, and the listing status. The benefit of scraping the data also would allow us to select what location we wanted to perform our analytics on.

The second option we looked at was the data available on Kaggle. We found a large data set that had extensive collection of factors and boasting 81 columns of data, separated into training and test sets, along with a currently active Kaggle competition. Due to the sheer size of the data, we decided to move forward with this data set found On Kaggle titled "*House Prices* -

*Advanced Regression Techniques Predict sales prices and practice feature engineering, RFs, and gradient boostin*g".

This data set, included both categorical and numerical values. The numerical values were integers and the categorical values were strings. The data columns available in the training set included: Id, MSSubClass, MSZoning, LotFrontage, LotArea, Street, Alley, LotShape, LandContour, Utilities, LotConfig, LandSlope, Neighborhood, Condition1, Condition2, BldgType, HouseStyle, OverallQual, OverallCond, YearBuilt, YearRemodAdd, RoofStyle, RoofMatl, Exterior1st, Exterior2nd, MasVnrType, MasVnrArea, ExterQual, ExterCond, Foundation, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinSF1, BsmtFinType2, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, Heating, HeatingQC, CentralAir, Electrical, 1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, BedroomAbvGr, KitchenAbvGr, KitchenQual, TotRmsAbvGrd, Functional, Fireplaces, FireplaceQu, GarageType, GarageYrBlt, GarageFinish, GarageCars, GarageArea, GarageQual, GarageCond, PavedDrive, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea, PoolQC, Fence, MiscFeature, MiscVal, MoSold, YrSold, SaleType, SaleCondition, as well as our target variable, which is SalePrice.

Using the CRISP-DM methodology to data analysis, we started with trying to understand the housing market, which has been increasing at a record setting pace (2021, Smart). We looked through the data set to understand the factors available. We hypothesized that certain factors would have an impact on the price of the home, including square feet, lot size, and neighborhood.

Data Preparation

For data preparation, we first combined the training and test data sets. This ended up being in error as the test data set did not include the target variable and was for submitting results to the Kaggle competition. We looked at the number of records (rows) in the training data, which was 1460.

We encoded ordinal categorical variables using OrdinalEncode in the sklearn.preprocessing python package. Variables handed in this method included lotShape, LandSlope, ExterQual, ExterCond, BsmtQual, BsmtCond, HeatingQC, KitchenQual, FireplaceQu, GarageQual, GarageCond, PoolQC, BsmtExposure, BsmtFinType1, BsmtFinType2, Functional, GarageFinish and PavedDrive. We set up the list in order of their categories and encoded lot irregularity.

For the rest of the categorical variables, we created dummy variables.

The other data preparation that needed to be addressed was reviewing and handling of null values. After getting a sorted list of columns with null values, we replaced NA with 'None' or '0'. We imputed any remaining numerical null vales with the mean or the mode.

Preliminary Analysis

We looked at the numerical variables basic statistics including count, mean, standard deviation, min, and max to get an idea of the data we were working with.

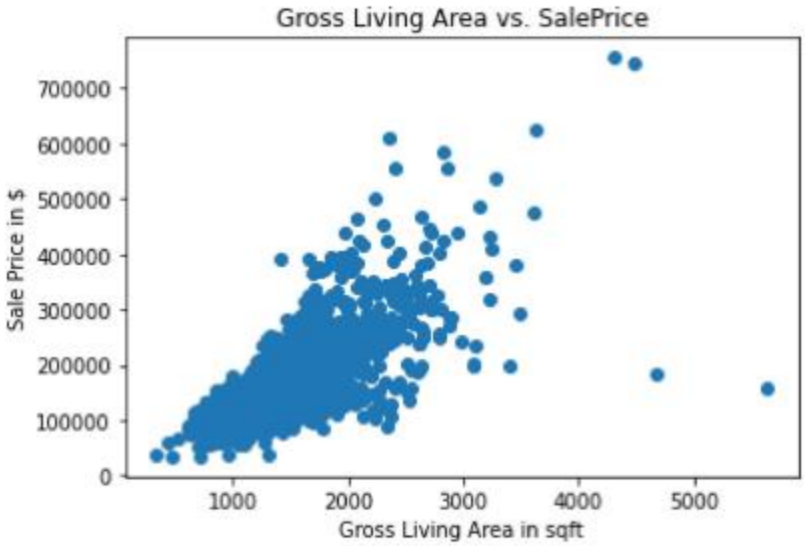
Next, we performed a correlation matrix on the data set and identified which values had high correlation values to our data set. A list of values with medium-high correlation to SalesPrice is listed below:

Data Points with high correlation to SalesPrice:

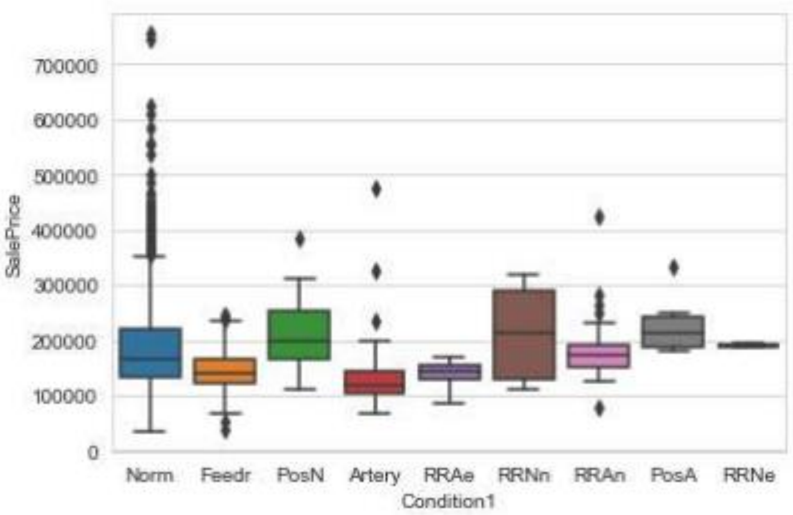
• OverallQual	0.790982
• YearBuilt	0.522897

YearRemodAdd	0.507101
ExterQual	0.682639
BsmtQual	0.585207
TotalBsmtSF	0.613581
1stFlrSF	0.605852
GrLivArea	0.708624
FullBath	0.560664
KitchenQual	0.659600
TotRmsAbvGrd	0.533723
FireplaceQu	0.520438
GarageFinish	0.549247
GarageCars	0.640409
GarageArea	0.623431

We looked at these values in a scatter chart, comparing the target variable (Sale Price) to the numerical categories including Lot Frontage, Lot Area, Overall Quality, Year Built, Year Remodeled, Gross Living Area, Total Basement Area Masonry Veneer Area, WoodDeck Area and Open Porch Area. The best representation we found of a good relationship was Gross Living Area to SalePrice as seen below.

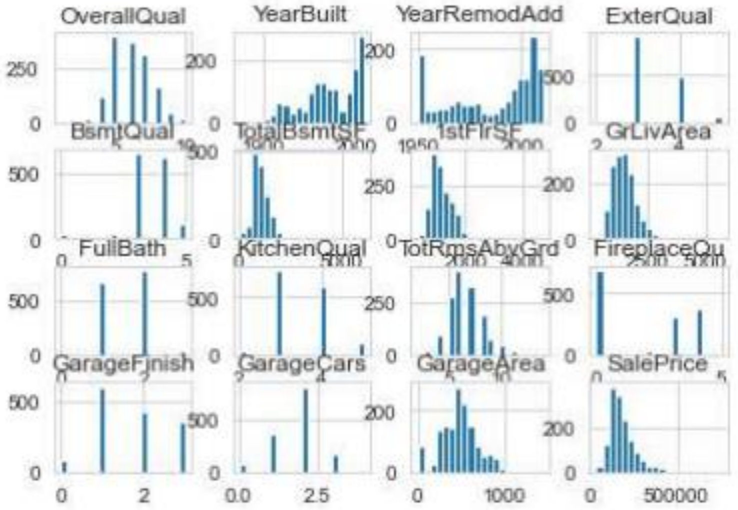


Next, we reviewed box plots for categorical variables. This allowed us to check on the distribution of values based on the categorical variable.



The distribution of high-correlating variables were also reviewed, showing that many of the variables including TotalBsmtSF, 1stFlrSF, GrLvArea, GarageArea and SalePrice were all skewed to the right.

Histograms: All variables



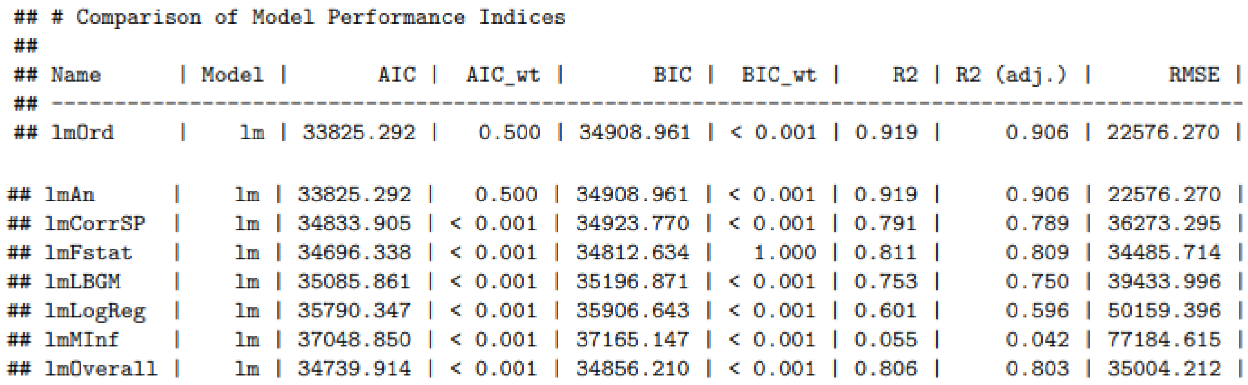
Lastly, we created feature selected data subsets for variables with high correlation, variables with high F-statistic value, variables with high LightGBM value, variables with high logistic regression value, variables with high mutual information values and overall top variables from feature selections. When creating the data sets, for the subsets, we used the top 30 variables for each subset.

Modeling and Results

The next step was loading the data frames into R studio for modeling. We had our cleaned and prepped data sets loaded into R. For each data set, we created a regression model to determine which linear model performed the best and checked collinearity of the model variables. We also checked the accuracy, and goodness of fit test. We picked the best models from regression and decision trees to create an ensemble model

Linear Modeling

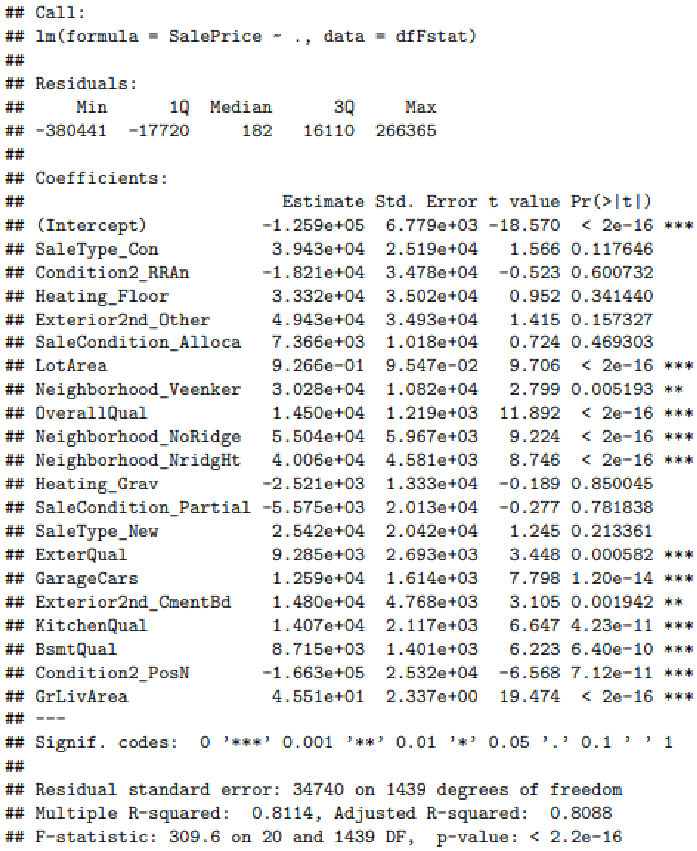
For the linear regression models, we compared the performance of each model. The highest performing model was the one created from the Entire Dataset with Ordinal Variables converted. The second highest performing model was the one created with the dummy variables converted. Wanting to have a reduction in features and not using all features, the best forming model was the lmFstat model, which had an R-Squared of .811, which met our expectation of having 80% or more of the variability explained by the model.



```
## # Comparison of Model Performance Indices
##
## Name       | Model |       AIC |  AIC_wt |       BIC |  BIC_wt |    R2 | R2 (adj.) |      RMSE |
## ---------------------------------------------------------------------------------------------------
## lmOrd      |    lm | 33825.292 |   0.500 | 34908.961 | < 0.001 | 0.919 |     0.906 | 22576.270 |

## lmAn       |    lm | 33825.292 |   0.500 | 34908.961 | < 0.001 | 0.919 |     0.906 | 22576.270 |
## lmCorrSP   |    lm | 34833.905 | < 0.001 | 34923.770 | < 0.001 | 0.791 |     0.789 | 36273.295 |
## lmFstat    |    lm | 34696.338 | < 0.001 | 34812.634 |   1.000 | 0.811 |     0.809 | 34485.714 |
## lmLBGM     |    lm | 35085.861 | < 0.001 | 35196.871 | < 0.001 | 0.753 |     0.750 | 39433.996 |
## lmLogReg   |    lm | 35790.347 | < 0.001 | 35906.643 | < 0.001 | 0.601 |     0.596 | 50159.396 |
## lmMInf     |    lm | 37048.850 | < 0.001 | 37165.147 | < 0.001 | 0.055 |     0.042 | 77184.615 |
## lmOverall  |    lm | 34739.914 | < 0.001 | 34856.210 | < 0.001 | 0.806 |     0.803 | 35004.212 |
```

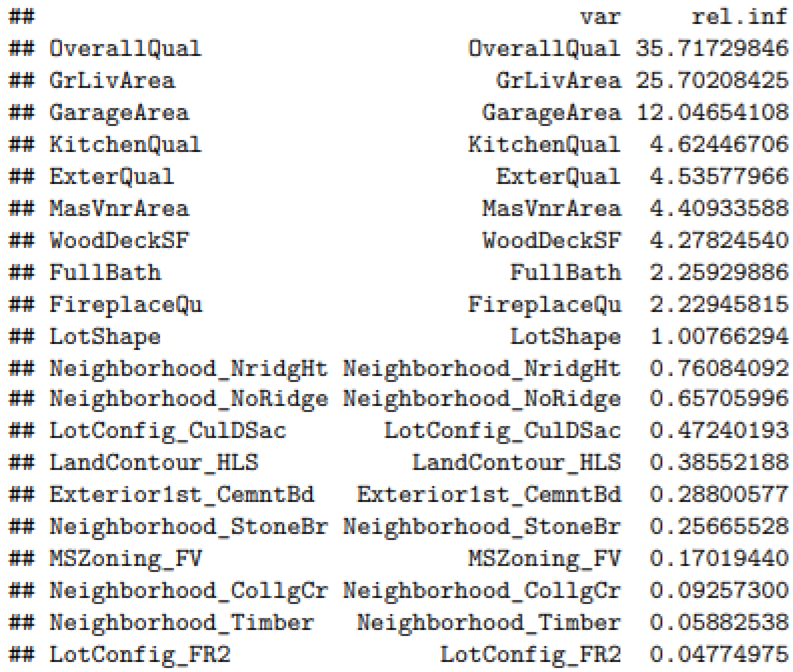
Looking into the lmStat model deeper, we can see that the which cariables had the highest significance to the target variable as seen below:



```
## Call:
## lm(formula = SalePrice ~ ., data = dfFstat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -380441  -17720     182   16110  266365
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.259e+05  6.779e+03 -18.570  < 2e-16 ***
## SaleType_Con           3.943e+04  2.519e+04   1.566 0.117646
## Condition2_RRAn       -1.821e+04  3.478e+04  -0.523 0.600732
## Heating_Floor          3.332e+04  3.502e+04   0.952 0.341440
## Exterior2nd_Other      4.943e+04  3.493e+04   1.415 0.157327
## SaleCondition_Alloca   7.366e+03  1.018e+04   0.724 0.469303
## LotArea                9.266e-01  9.547e-02   9.706  < 2e-16 ***
## Neighborhood_Veenker   3.028e+04  1.082e+04   2.799 0.005193 **
## OverallQual            1.450e+04  1.219e+03  11.892  < 2e-16 ***
## Neighborhood_NoRidge   5.504e+04  5.967e+03   9.224  < 2e-16 ***
## Neighborhood_NridgHt   4.006e+04  4.581e+03   8.746  < 2e-16 ***
## Heating_Grav          -2.521e+03  1.333e+04  -0.189 0.850045
## SaleCondition_Partial -5.575e+03  2.013e+04  -0.277 0.781838
## SaleType_New           2.542e+04  2.042e+04   1.245 0.213361
## ExterQual              9.285e+03  2.693e+03   3.448 0.000582 ***
## GarageCars             1.259e+04  1.614e+03   7.798 1.20e-14 ***
## Exterior2nd_CmentBd    1.480e+04  4.768e+03   3.105 0.001942 **
## KitchenQual            1.407e+04  2.117e+03   6.647 4.23e-11 ***
## BsmtQual               8.715e+03  1.401e+03   6.223 6.40e-10 ***
## Condition2_PosN       -1.663e+05  2.532e+04  -6.568 7.12e-11 ***
## GrLivArea              4.551e+01  2.337e+00  19.474  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34740 on 1439 degrees of freedom
## Multiple R-squared:  0.8114, Adjusted R-squared:  0.8088
## F-statistic: 309.6 on 20 and 1439 DF,  p-value: < 2.2e-16
```

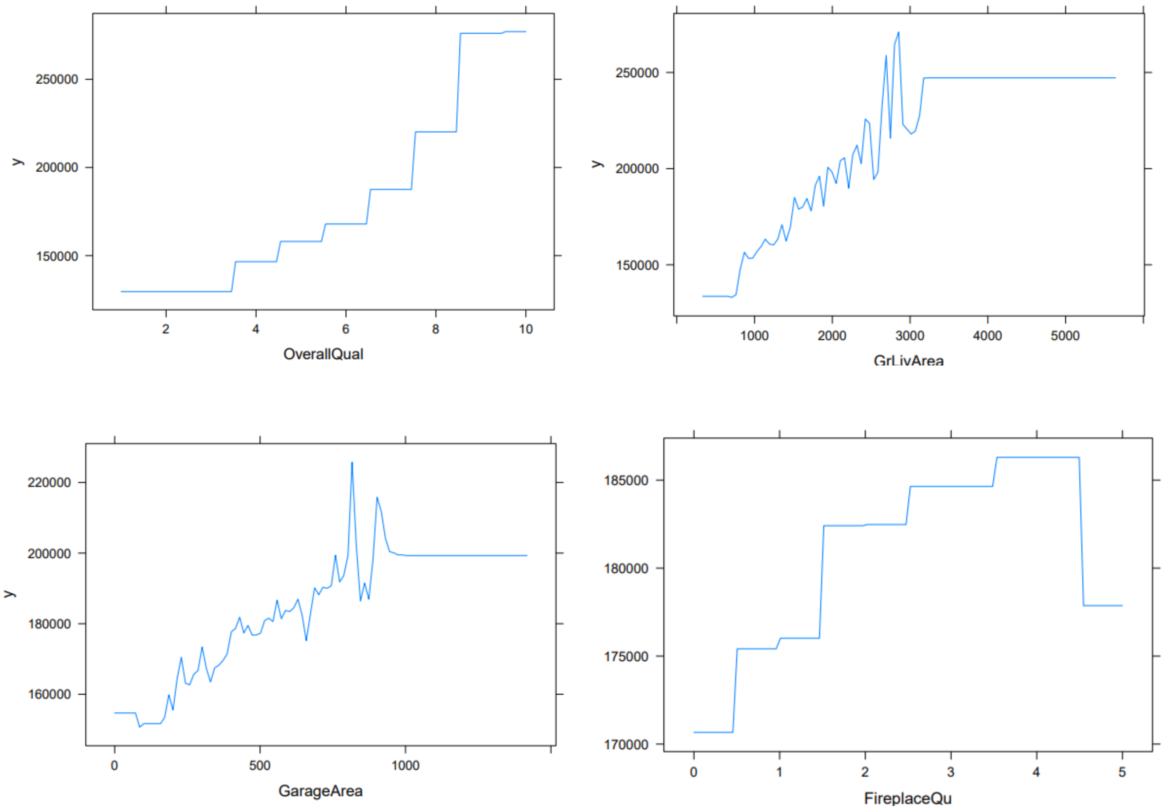
Decision Tree Modeling

Using our overall dataset, we created a decision tree model and boosted the trees using gradient boosted modeling.

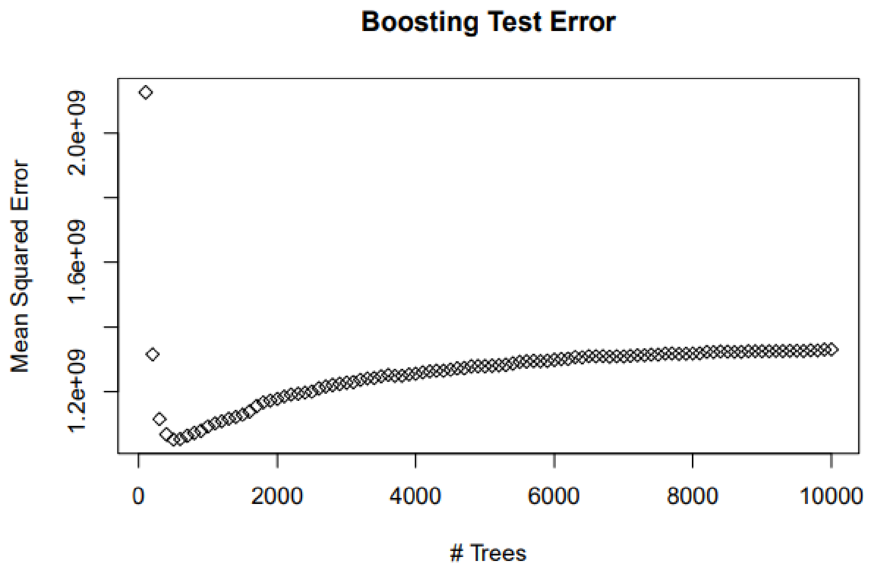


```
##                                          var     rel.inf
## OverallQual                      OverallQual 35.71729846
## GrLivArea                          GrLivArea 25.70208425
## GarageArea                        GarageArea 12.04654108
## KitchenQual                      KitchenQual  4.62446706
## ExterQual                          ExterQual  4.53577966
## MasVnrArea                        MasVnrArea  4.40933588
## WoodDeckSF                        WoodDeckSF  4.27824540
## FullBath                            FullBath  2.25929886
## FireplaceQu                      FireplaceQu  2.22945815
## LotShape                            LotShape  1.00766294
## Neighborhood_NridgHt Neighborhood_NridgHt  0.76084092
## Neighborhood_NoRidge Neighborhood_NoRidge  0.65705996
## LotConfig_CulDSac       LotConfig_CulDSac  0.47240193
## LandContour_HLS           LandContour_HLS  0.38552188
## Exterior1st_CemntBd   Exterior1st_CemntBd  0.28800577
## Neighborhood_StoneBr Neighborhood_StoneBr  0.25665528
## MSZoning_FV                      MSZoning_FV  0.17019440
## Neighborhood_CollgCr Neighborhood_CollgCr  0.09257300
## Neighborhood_Timber   Neighborhood_Timber  0.05882538
## LotConfig_FR2                  LotConfig_FR2  0.04774975
```

We were able to determine that our model is largely influenced by only a handful of variables which include OverallQual, GrLivArea and GarageArea. When reviewing deeper, most of these variables appear to show a roughly linear increase in correlation with SalePrice, but notice that FirePlaceQu has a huge jump between 0/1 rating compared to 2+ rating. In the future, this variable could probably be binned into a binary good/bad rating.

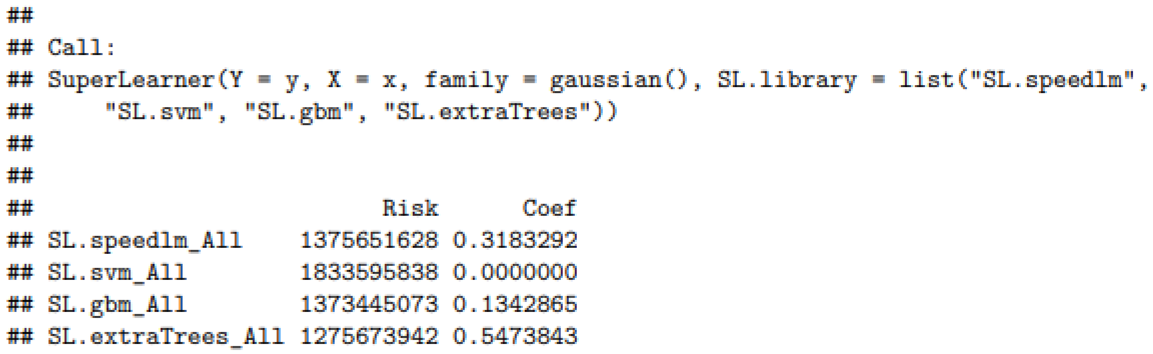


Predicting the boosted model on the data set gave us the boosting test error. The boosting managed to drop the error below the best error from the Random Forest model.

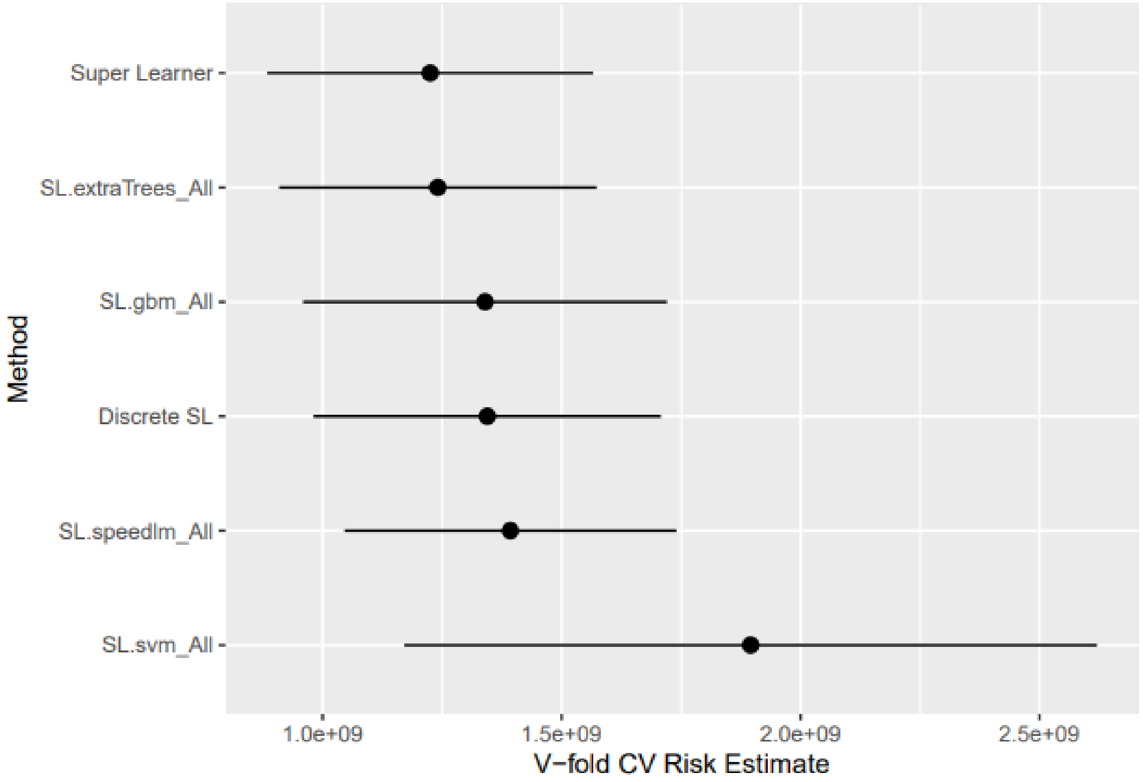


Ensemble

The final step was to create an ensemble model using the created regression and decision tree models. The extraTrees model in the ensemble has the highest coefficient, indicating it is weighted highly in the ensemble. Following this, speedlm and gbm are the next highest weighted models.

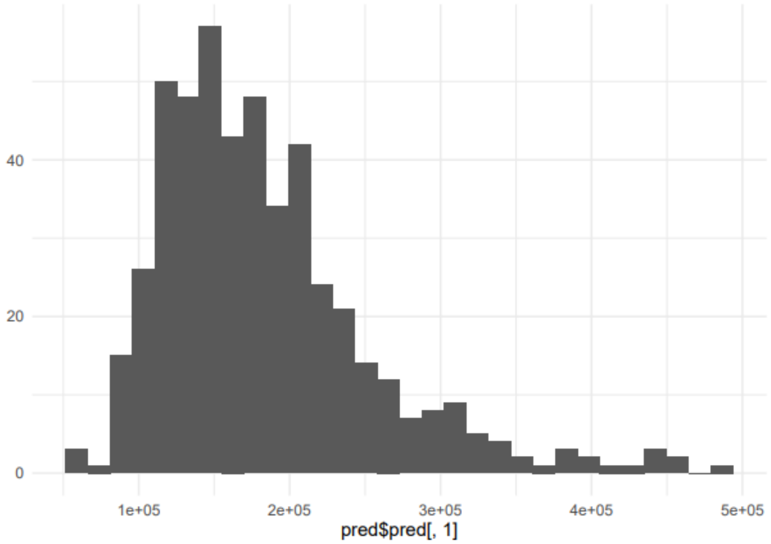


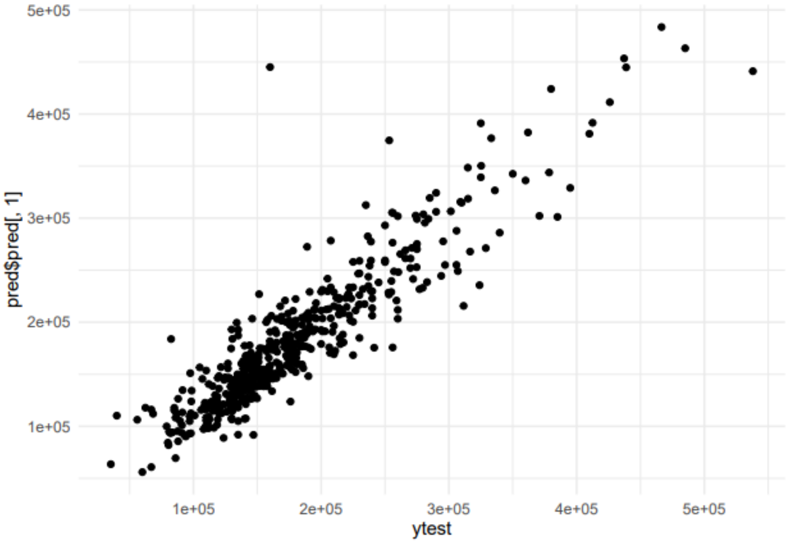
```
##
## Call:
## SuperLearner(Y = y, X = x, family = gaussian(), SL.library = list("SL.speedlm",
##     "SL.svm", "SL.gbm", "SL.extraTrees"))
##
##
##                           Risk      Coef
## SL.speedlm_All    1375651628 0.3183292
## SL.svm_All        1833595838 0.0000000
## SL.gbm_All        1373445073 0.1342865
## SL.extraTrees_All 1275673942 0.5473843
```



Discussion/Conclusion

Final predictions were created using the SuperLearner package. This took the best of each model and put them together for a strong prediction.





Below are some of the conclusions / findings we would like to share:

1) We learned that although using all of the available features in the entire training dataset, would be helpful in achieving higher level of accuracy (above 90%), preparing and

running the models would not be efficient and time friendly. So, we utilized various techniques for feature reduction / feature selection to hone in on the critical features.

2) We also realized that using ordinal encoding was more efficient and produced better accuracy levels rather than using dummy variables creation for categorical features

3) Some of the key factors affecting the Home prices were Overall Quality of the house, Gross Living area and Garage area which had greater impact

4) Some of the other factors were the Lot area, Basement area, Year Built / Remodeled (with recent buildings / remodeling having higher prices), Fireplace quality, number of rooms.

5) Predictive Analytics system consisting of combination of multiple Machine Learning models brought the best out of all the constituting models and improved the accuracy levels at the same time - around 88% which was well above our target levels.

Acknowledgements

We wish to thank all the people whose assistance was a milestone in the completion of this project with special recognition to our professor, Fadi Alsaleem.

References

House prices - advanced regression techniques. Kaggle. (n.d.). Retrieved September 30, 2021, from https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview.

Smart, T. (2021, September 28). *Home prices continue record-setting pace, rising 19.7 % in ...* USNews.com. Retrieved September 30, 2021, from https://www.usnews.com/news/economy/articles/2021-09-28/home-prices-continue-record-setting-pace-rising-197-in-july.

https://cran.r-project.org/web/packages/performance/performance.pdf

https://www.datacamp.com/community/tutorials/decision-trees-R (Random Forest)

https://cran.r-project.org/web/packages/gbm/gbm.pdf

https://cran.r-project.org/web/packages/SuperLearner/vignettes/Guide-to-SuperLearner.html

https://www.datacamp.com/community/tutorials/ensemble-r-machine-learning