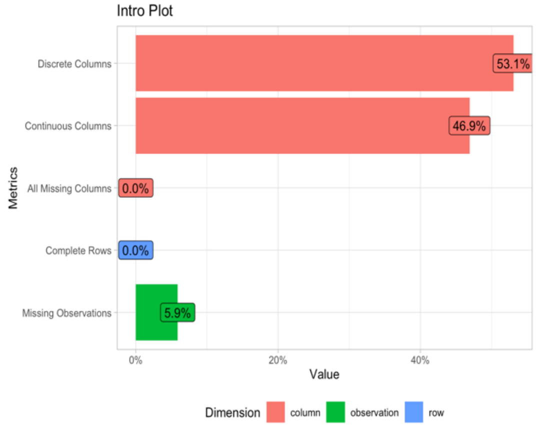
House Prices - Advanced Regression Techniques

Group: Dharaneesh Subramani, Gnaneshwaran Vadivelsureshbabu, Sai Sadhankumar

Introduction

To fill the information gap and improve the efficiency of the Real Estate market, the House Price prediction model is essential. House price predictions are useful for real estate agents (realtors) and investors to determine the trend in housing prices. Additionally, the prediction is expected to be useful to people who are planning to buy a house, so that they will be able to plan their finances well. Our goal is to predict the final price of each home using the data which provides 1460 observations and 79 explanatory variables describing almost every aspect of residential homes in Ames, Iowa. In this predictive model, you can find out what aspects should be taken into account for better rates.

Model Description - Importing, Counting the number of missing data and Cleaning



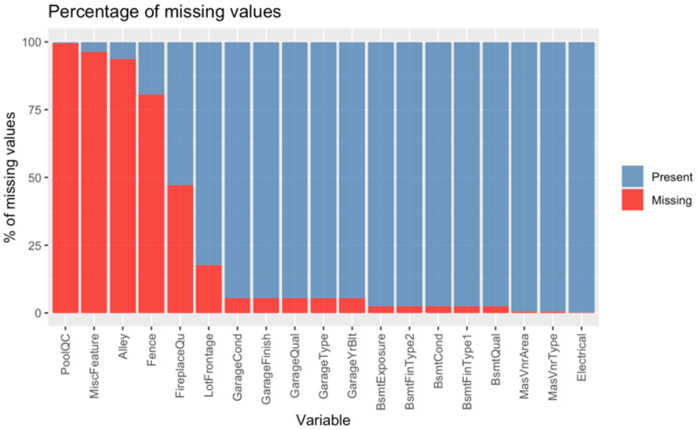


Figure 1: Plot_intro() gives us a different way of visualizing our metadata. It gives us the percentage of discrete columns, continuous columns, missing columns, missing observations, etc. in bar graph format. This makes data exploration a lot easier and gives us guidance when cleaning our dataset.

Figure 2: This plot gives a way to look more in-depth into how much data you're missing and where it's missing. It groups by columns and finds all of the missing rows for each one. There are lot of missing values, however most of these are not actual missing values as the 'NA' usually means a specific feature is missing.

It's helpful to understand why the data could be missing in order to handle missing data successfully. Making educated guesses and replacing the NA can be a helpful approach. With the help of the data dictionary, I recoded NA's appropriately. When we look at Pool Area, there is no missing value; however, if we look at Pool Quality, there are about 1453 NA values. The pool area has a value of 0 for some particular ID, which indicates that the pool does not exist. Hence, I have replaced NAs with 0.

The quality variables such as kitchenQual, ExterQual, ExterCond, BsmtQual, BsmtCond can be modeled either as an ordinal factor or as a numeric variable. We also converted all the remaining character variables into factors that could be grouped into dummy variables in the model. Even though it is a numeric integer, this is a category, so it is used with MSSubClass.

Finally, Log transforming is used to reduce the scew so the data can be understood easier. we converted the variables with skewed(distorted) data. The area of most homes is fairly average, but some homes are distorted to the right. We also removed some of the minimum variance variables that added noise to the model.

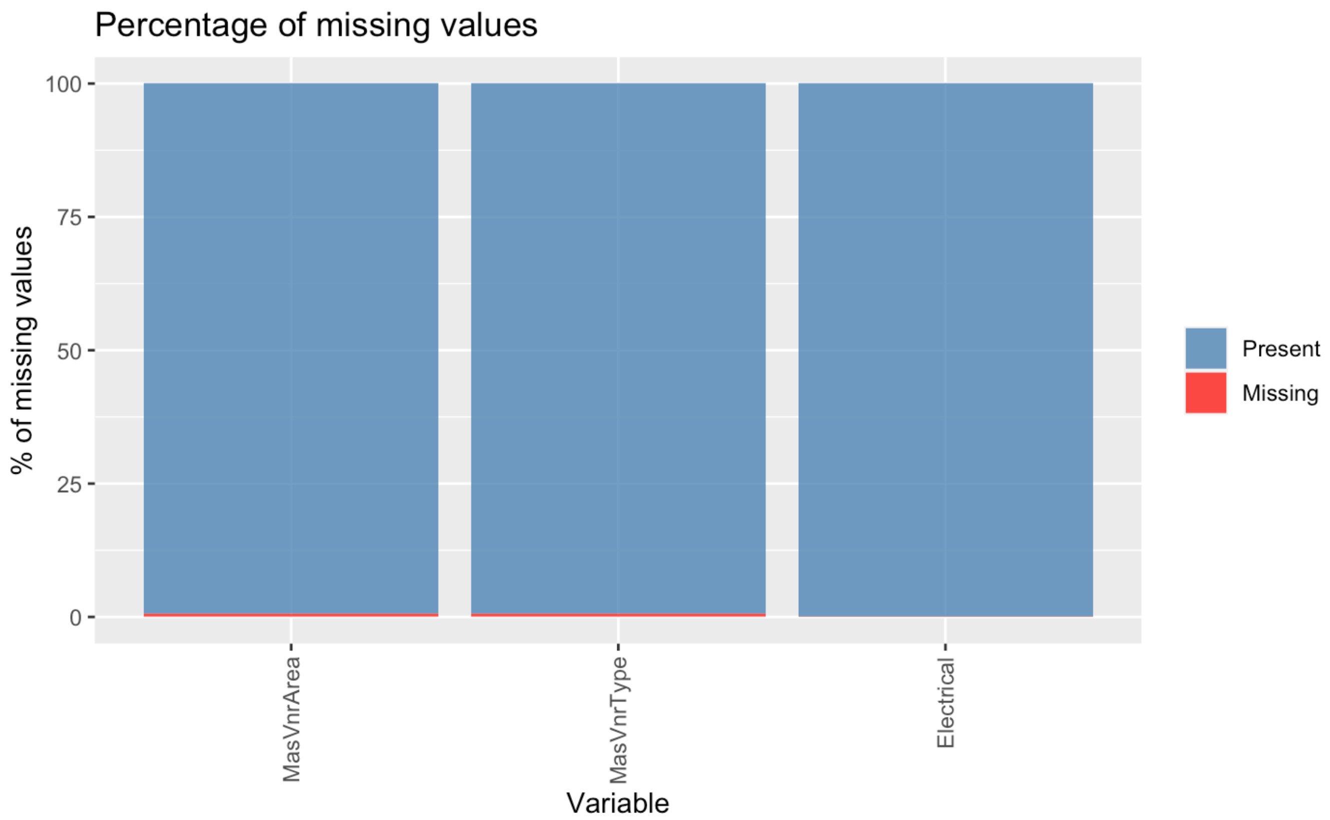


Figure 3: After recoding the NAs into their respective categories and cleaning up the data, we find that there are really few true missing values that need to be entered.

Imputing Missing Values in Data Sets

Missing values are considered to be the first obstacle in predictive modeling. Hence, it's important to master the methods to overcome them. We Imputed legitimately missing data in both the training and test datasets. 'Impute Missings' package has great embed functionality. When using the median/mode method, the character vectors and factors are imputed using the mode. Numeric and integer vectors are imputed with the median.

Creating The Model

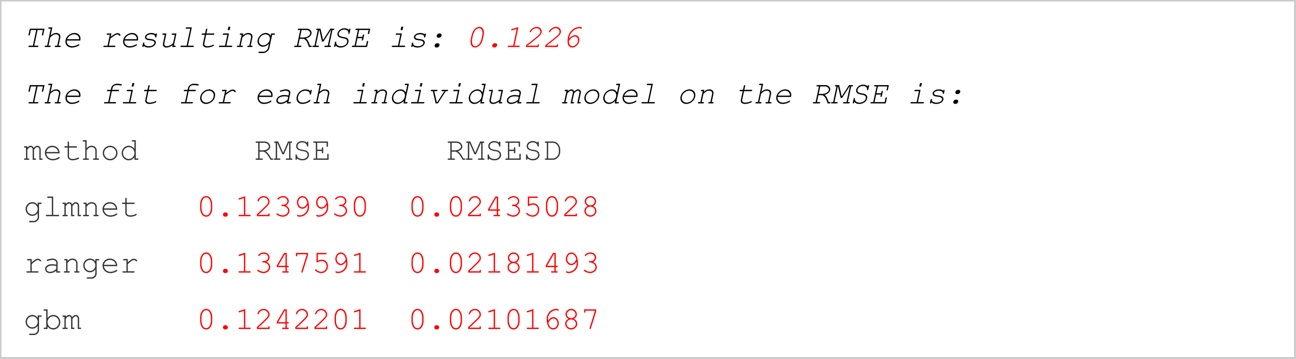
We tested multiple predictive models, including random forests, neural nets, linear models and, stochastic gradient boosting machine model.

glmnet:. Glmnet is a package that fits generalized linear and similar models via penalized maximum likelihood A generalized linear model uses regularized least squares to fit models with numeric outcomes. It is excellent when you have a huge quantity of predictors in which it can decrease the weights of every coefficient as wished.

ranger: Random Forest developed by an aggregating tree, and this can be used for classification and regression. One of the major advantages is it avoids overfitting. The random forest can deal with many features, and it helps to identify the important attributes. It can predict a continuous output and forming each tree individually and averaging them at the end.

gbm: Gradient boosting machines (GBMs) are an extremely popular machine learning algorithm that have proven successful across many domains and is one of the leading methods for winning Kaggle competitions. A stochastic gradient boosting machine model is also a collection of simple decision trees which can create a complex set of decision trees, capable of continuous predictions. Forming and adding the results along the way.

Whereas random forests build an ensemble of deep independent trees, GBMs build an ensemble of shallow and weak successive trees with each tree learning and improving on the previous. When combined, these many weak successive trees produce a powerful "committee" that are often hard to beat with other algorithms.



```
The resulting RMSE is: 0.1226
The fit for each individual model on the RMSE is:
method      RMSE       RMSESD
glmnet   0.1239930  0.02435028
ranger   0.1347591  0.02181493
gbm      0.1242201  0.02101687
```

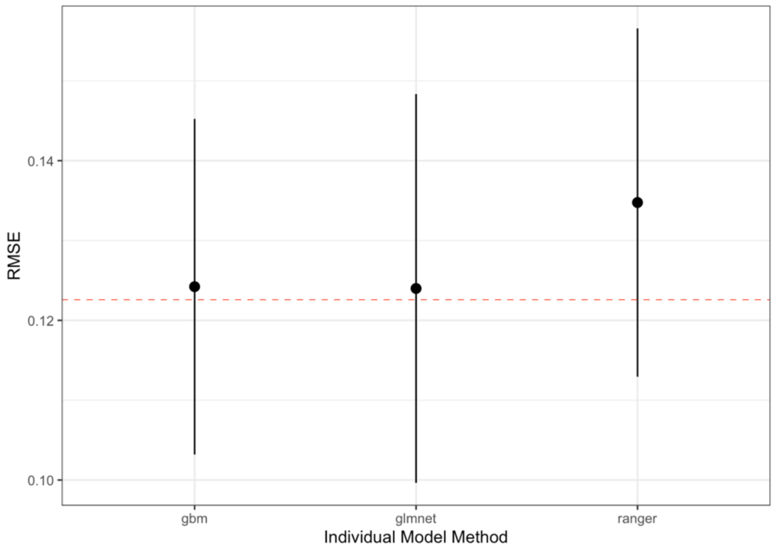
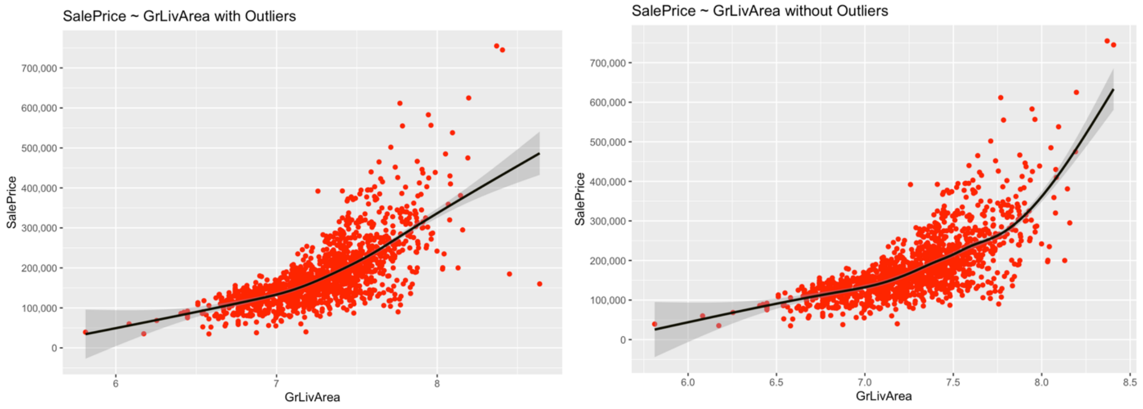


Figure 4: Plotted the fit for each individual model on the RSME

From the fit for each individual model the ‘glmnet’ and ‘gbm’ are performing good with an out of sample RMSE of .124, while the ‘ranger’ random forest models also performed well with an RMSE of .135 respectively. Combining these three models the resulting RMSE of .123.

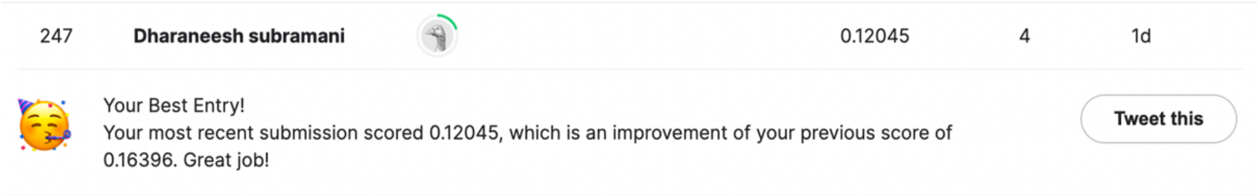
Model Performance



Plotted the data with and without outliers. By removing the outliers, the model has improved, and we can see the standard error has also decreased. The trend line is more in line with the data density points. In any case, if one of our focuses in our testing information incorporates a low SalePrice with a high GrLivArea, we may not be able to anticipate it as well as in the event that we kept the outliers. It would result in overestimating the house value.

```
A glm ensemble of 3 base models: glmnet, ranger, gbm
Ensemble results:
Generalized Linear Model
7300 samples
3 predictor
No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 7300, 7300, 7300, 7300, 7300, 7300, ...

Resampling results:
RMSE       Rsquared   MAE
0.1225764  0.9059921  0.07907188
```



247 Dharaneesh subramani 0.12045 4 1d

Your Best Entry!
Your most recent submission scored 0.12045, which is an improvement of your previous score of 0.16396. Great job!

In sample RMSE value is 0.087 and R^2 value is 0.952. It looks confident. The estimated-out sample RMSE value 0.122 and R^2 value is 0.905. The model is slightly overfitting with a drop of .035 RMSE. The R^2 is also good with a drop of only 0.047. Also the model is below .15 log RSME. The Kaggle score was pretty good with Score: .12045, ranking #247.