Real Estate Price Predictor

Juan Lopez

jxl1581@miami.edu

ABSTRACT

This paper encapsulates a semester worth long of Statistical Learning/ Machine Learning algorithms all put into a Regression type model for data analysis on Real Estate Data. My idea was to use and optimize the models for better Mean Squared Errors and testing them compared to other models. Some of the specific models I decided to look at were, Multinomial Regression with all, significant and polynomial attributes, Bagging with all, significant, and polynomial attributes, and Random Forests with all, significant, and polynomial regression. My analysis concluded that Random Forests was the best learning model that could give the least Mean Squared Error. These findings were very significant and importance

Index Terms— Real Estate, Machine Learning, Regression

1. INTRODUCTION

In the volatile market of real estate, a dynamic ever changing market of based on different attributes, locations and changes around all area of life. The real estate market is one of the most crucial industries in any national/ international economy. In which people do need homes, and living places that meet their needs or reach out of their expectations. Observing the real estate market as a buyer or a seller one can clearly see trends and markets in which one can assume accurate predictions of real estate prices. The interesting thing is that most of these tells are based on user bias and user knowledge. One can never know the true measurement of ones sentimental home in terms of the homes around them.

This is why Real Estate Price Forecasting is a more difficult and complicated task when it comes to direct and indirect these factors could come in as qualitative and quantitative attributes. In general I converted these qualitative attributes into numerical levels to determine the quality of places. Some of the qualitative attributes that could impact Real Estate prices come from building styles, neighborhood, quality of different parts of a home. Some quantitative attributes could come from past sale prices, area of outdoor and indoor living space in square footage. There was at first some issues in being able to derive a good model using all the data without pre-processing some of its attributes. Data was extracted from Boston Living Dataset from Kaggle and pre-processed before applying any models. After the pre-processing step our data would be prepared for data modeling.

2. STATE OF THE ART

In an article posted by Bo-Sin Tang & Siu Wai Wong, published in February 2020. They tried to use 3 Machine Learning Algorithms including SVM, RF, GBM. In the appraisal prices of 40,000 house transactions in a span of 18 years in Hong Kong. They discovered the RF and GBM performed better than SVM.

Another meaningful article by Ping-Feng Pair & Wen-Chang Wang. Studied 4 Machine Learning models for Actual Transaction Data for Predicting Real Estate Prices. These models included Least Squares Support Vector Regression (LSSVR), Classification and Regression Tree (CART) , General Regression Neural Networks, and Back-Propagation Neural Networks (BPNN) They discovered that the LSSVR outperformed the other three Machine Learning Algorithms. And that LSSVR provided relatively competitive and Satisfactory Results.

Some of my findings done on Regression, Bagging, and Random Forests are almost identical to those of the two articles that I looked at. My conclusion on the analysis of the results are also consistent to the hypothesis of these articles so my hypothesis was also correct. Random Forests would do the best in comparison to the other models I was working on. This could mean that certain types of models could be beneficial to improve Real Estate Price forecasting in comparison to others and thus Zillow a common house hold real estate website with a 33.02 Billion Dollar worth as an example could be most benefited by using a Random Forest Model to better determine the price and well estimation of the sellers listings. This could also improve the interaction between Buyer and Seller to stop and filter out offers that are not within the meaningful range of predictions. I believe also because Real Estate data is dynamically changing and markets could go on either bullish or bearish trends one can see these changes and have better learning model predictions and more data is collected.

3. REAL ESTATE DATA

The current dataset was pulled and used from Kaggle: **House Prices - Advanced Regression Techniques** - Predict sales prices and practice feature engineering, Random Forests, and gradient boosting

This dataset consisted of 1,460 observations with about 81 different variables.

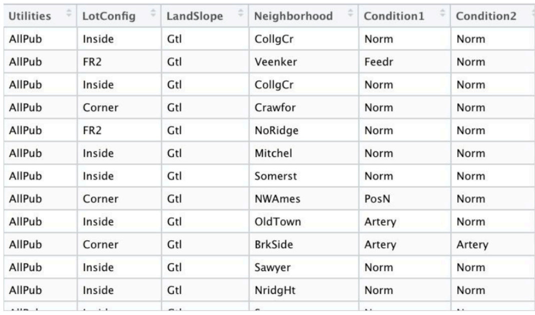
There was a lot of trial and error and reading documentation to figure out the correlation of each attribute and why it was chosen and what would happen if a home did not have this certain attribute.

Let’s just say a lot of homes did not have all attributes thus na.omit(dataset) would erase the whole dataset.

I had to erase the columns that were completely NA. And then figure out which columns had the most NA and slowly start picking off columns from there until I finally reached a workable dataset. In real life you would like to fill these NA with meaningful data for better results but for now we ignore them. Some of the attributes without their values are provided in the Dataset description provided in the Project File. The list was very long and thus copying and pasting it into this paper would take up too much space and not be useful. There were 81 attributes to decipher and after the pre-processing stage brought my use of the attributes to about 62 and used that data to train my models.

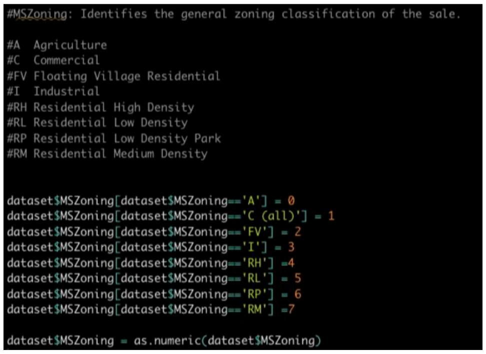
4. PRE-PROCESSING METHODS

This wasn’t an easy task at first but after having the list of the dataset description one could achieve this by using simple label encoding.An example of it is below and is as follows.



Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2
AllPub	Inside	Gtl	CollgCr	Norm	Norm
AllPub	FR2	Gtl	Veenker	Feedr	Norm
AllPub	Inside	Gtl	CollgCr	Norm	Norm
AllPub	Corner	Gtl	Crawfor	Norm	Norm
AllPub	FR2	Gtl	NoRidge	Norm	Norm
AllPub	Inside	Gtl	Mitchel	Norm	Norm
AllPub	Inside	Gtl	Somerst	Norm	Norm
AllPub	Corner	Gtl	NWAmes	PosN	Norm
AllPub	Inside	Gtl	OldTown	Artery	Norm
AllPub	Corner	Gtl	BrkSide	Artery	Artery
AllPub	Inside	Gtl	Sawyer	Norm	Norm
AllPub	Inside	Gtl	NridgHt	Norm	Norm

A small portion of the dataset is as follows and thus after applying the label encoding.



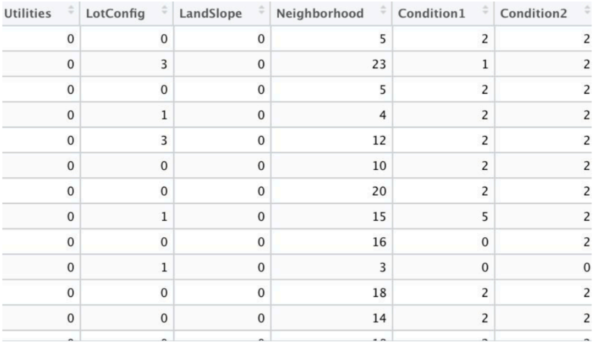
```
#MSZoning: Identifies the general zoning classification of the sale.

#A  Agriculture
#C  Commercial
#FV Floating Village Residential
#I  Industrial
#RH Residential High Density
#RL Residential Low Density
#RP Residential Low Density Park
#RM Residential Medium Density


dataset$MSZoning[dataset$MSZoning=='A'] = 0
dataset$MSZoning[dataset$MSZoning=='C (all)'] = 1
dataset$MSZoning[dataset$MSZoning=='FV'] = 2
dataset$MSZoning[dataset$MSZoning=='I'] = 3
dataset$MSZoning[dataset$MSZoning=='RH'] =4
dataset$MSZoning[dataset$MSZoning=='RL'] = 5
dataset$MSZoning[dataset$MSZoning=='RP'] = 6
dataset$MSZoning[dataset$MSZoning=='RM'] =7

dataset$MSZoning = as.numeric(dataset$MSZoning)
```

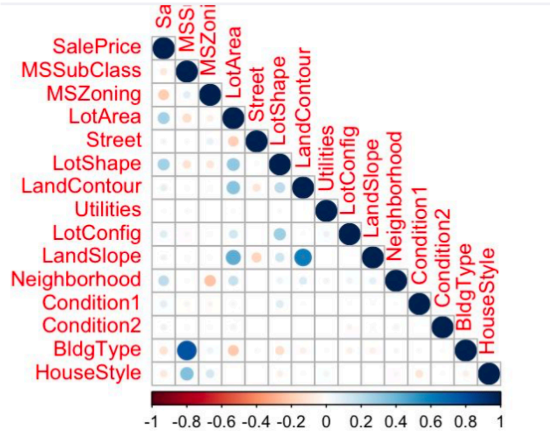
The result after the label encoding of each attribute provided us something that was viable for the pre-model analysis and model implementations. The following is the result of after the pre-process of all 62 attributes

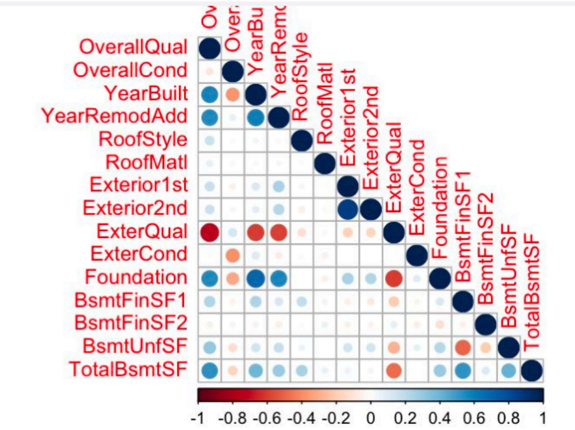


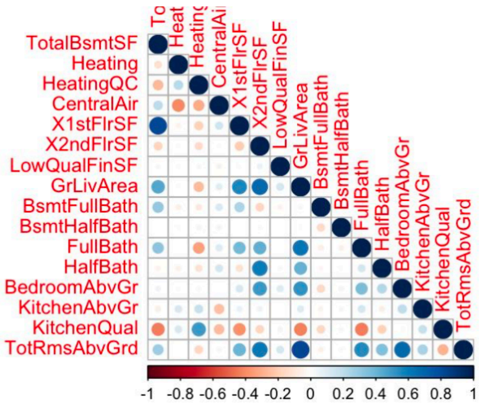
Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2
0	0	0	5	2	2
0	3	0	23	1	2
0	0	0	5	2	2
0	1	0	4	2	2
0	3	0	12	2	2
0	0	0	10	2	2
0	0	0	20	2	2
0	1	0	15	5	2
0	0	0	16	0	2
0	1	0	3	0	0
0	0	0	18	2	2
0	0	0	14	2	2

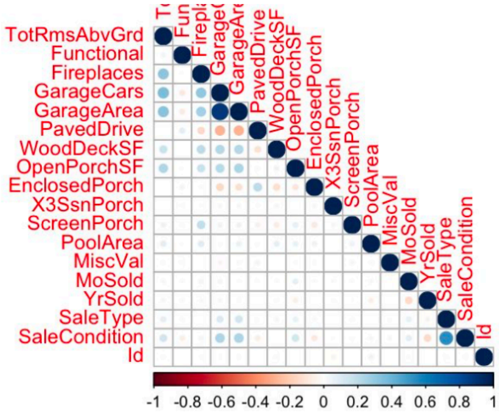
5. PRE-MODEL ANALYSIS

Some of correlations that I found are as follows. The correlation table was too large for 62 attributes so I split it into 15 attribute sets. And then computed the overall highest correlations. They are as followed.

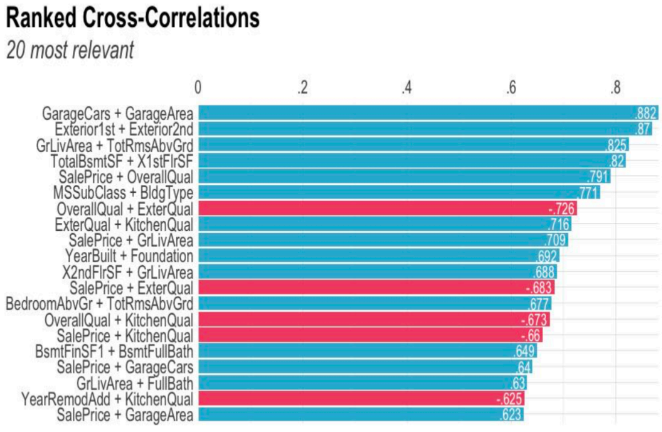




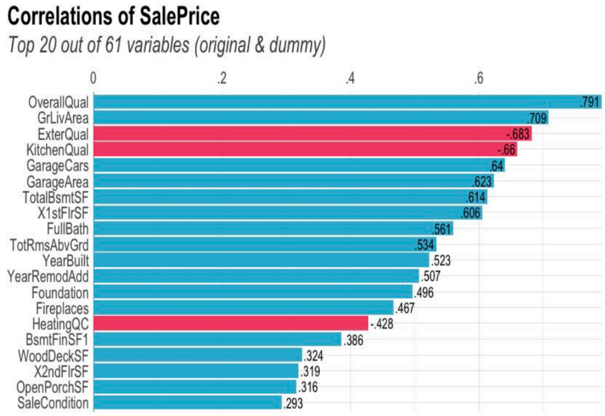




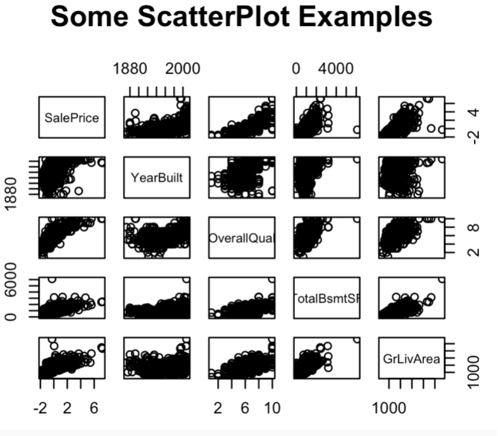
The top correlations between 2 attributes and the comparison our chosen Y value.

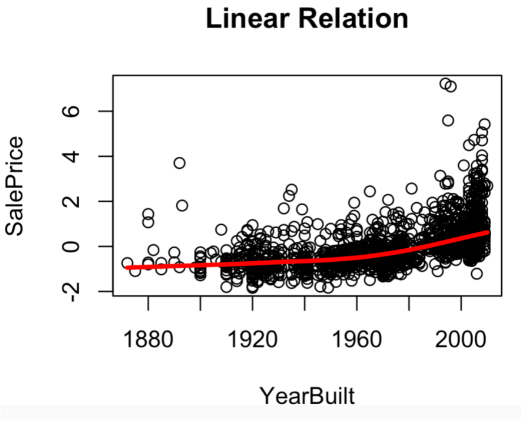


Our Y value of SalePrice also has specific attributes in which their correlation are very significant this includes.



Looking at scatterplots of the important correlations one can see clearly the Linear Correlations between them. Some of the scatterplots are as follows:

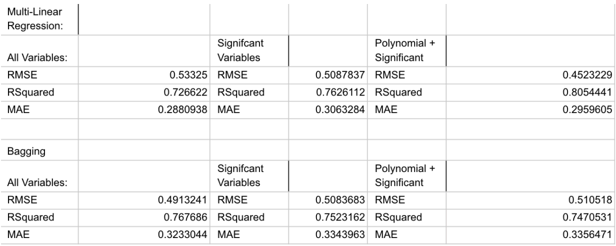




I decided to do a 70/30 split using a validation set approach. There were some issues in terms of Rank-Deficiency this is because of the issue of some of the columns in the dataset could be created from using scalar of certain other attributes. Some of the columns include GrLivArea, and TotalBsmtSF the example of my scalar multiplications that could be created are as an Example: GrLivArea (Total Living Area Square footage) could come from either 1st floor square footage or 1st floor square footage+2nd floor square footage (Was getting Rank-Deficiency error when using predict for these reasons)

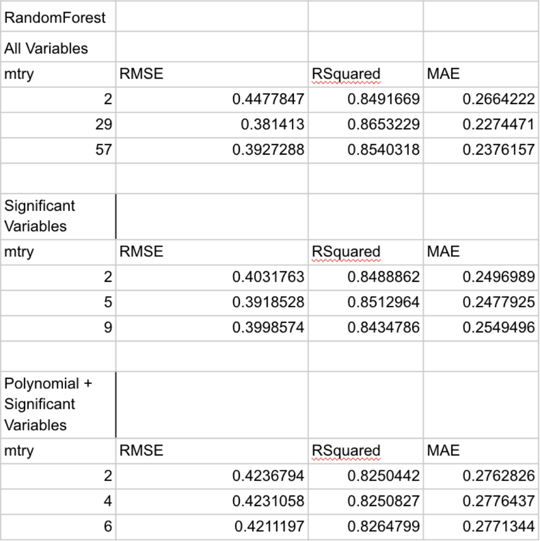
6. CROSS-VALIDATION APPROACH

To best demonstrate use the models I decided that a 10 Cross Fold with a repetition of 5 times and using some of the summary analysis on the RMSE, RSquared, and MAE to determine which model produces the best result. The significant attributes that I found were the most significant in my regression models are (Condition2, OverallQual, OverallCond, YearBuilt, BsmtFinSF1, BsmtUnfSF, X1stFlrSF, X2ndFlrSF, BedroomAbvGr). Using these significant variables and then using their polynomial variables (OverallQual, BsmtFinSF1, X1stFlrSF, X2ndFlrSF, I(OverallQual^2), I(X1stFlrSF^2)) Using these variables I decided to run my models and Compare their analysis.



Multi-Linear Regression:					
All Variables:		Signifcant Variables		Polynomial + Significant	
RMSE	0.53325	RMSE	0.5087837	RMSE	0.4523229
RSquared	0.726622	RSquared	0.7626112	RSquared	0.8054441
MAE	0.2880938	MAE	0.3063284	MAE	0.2959605
Bagging					
All Variables:		Signifcant Variables		Polynomial + Significant	
RMSE	0.4913241	RMSE	0.5083683	RMSE	0.510518
RSquared	0.767686	RSquared	0.7523162	RSquared	0.7470531
MAE	0.3233044	MAE	0.3343963	MAE	0.3356471

As we can see bagging and regression models did relatively the same compared to each other though some improvements in the Bagging using all Variables did better than the Regression using all the variables. Regression using significant and polynomial data did better overall than than the other 5 models I used. The better results on the training data was found using Random Forests Models their analysis was more consistent with the findings from the other research papers that I read. If I was to continue future work I would like to try to implement some of the models that they used including SVM or some Convolutional Neural Network to improve my results and test their analysis and compare to the ones I was researching.



RandomForest			
All Variables			
mtry	RMSE	RSquared	MAE
2	0.4477847	0.8491669	0.2664222
29	0.381413	0.8653229	0.2274471
57	0.3927288	0.8540318	0.2376157
Significant Variables			
mtry	RMSE	RSquared	MAE
2	0.4031763	0.8488862	0.2496989
5	0.3918528	0.8512964	0.2477925
9	0.3998574	0.8434786	0.2549496
Polynomial + Significant Variables			
mtry	RMSE	RSquared	MAE
2	0.4236794	0.8250442	0.2762826
4	0.4231058	0.8250827	0.2776437
6	0.4211197	0.8264799	0.2771344

This was the data that I found from using Random Forests as we can see the most viable Random Forest Model was with all variables and the best tune is mtry=29. The RMSE was at 0.381413 which is a low enough value and one that we can assume.

7. RESULTS AND CONCLUSION

Holding on to my Testing Set before applying it to a model I wanted to use Cross Validation to find the best Model. I found the best model to be Random Forest with an mtry=29. After applying my testing set for the perditions and calculating the RMSE I found that my RMSE on my testing set was 0.329767 which I thought was a very successful project. With the amount of data that there is for Real Estate where it can almost be categorized as Big Data using a Random Forests from a small example of my 1,500 data points a 32% Testing RMSE was enough to show the validity of these models.

8. FUTURE WORK

As this data is not streamed but recorded and saved connecting a Machine Learning Statistical Analysis application to a Real Estate software one would be able to constantly have models learn and predict future trends improve business needs. But this goes into the field of Big Data as this example alone was I thought very large in terms of # Variables. I believe also with this data streaming of what would seem to be unstructured data could also use the benefit of a type of neural network to also determine the linear relation with price and other attributes. Improvements in the field of Deep Learning could greatly improve overall Real Estate Businesses

12. REFERENCES

[1] Bo−sin Tang, Winky K.O. Ho, Siu Wai Wong. (2021) Sustainable development scale of housing estates: An economic assessment using machine learning approach. *Sustainable Development* 20.

[2] Jussi Kalliola, Jurgita Kapočiūtė-Dzikienė, Robertas Damaševičius. (2021) Neural network hyperparameter optimization for prediction of real estate prices in Helsinki. *PeerJ Computer Science* 7, pages e444.

.

[3] Ping-Feng Pai, Wen-Chang Wang (2020) Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices, *Applied Sciences*

[4] https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data