House Price Prediction: Regression Techniques

Syed Fahim Shahariar
ID: 2015-3-60-031
Department of CSE
East West University
Dhaka, Bangladesh
2015-3-60-031@std.ewubd.edu

Towhiduzzaman
ID: 2016-1-60-031
Department of CSE
East West University
Dhaka, Bangladesh
2016-1-60-031@std.ewubd.edu

Md. Khalid Hassan
ID: 2016-2-60-072
Department of CSE
East West University
Dhaka, Bangladesh
2016-2-60-072@std.ewubd.edu

Abstract— In today's world, buying houses has to be one of the biggest expenditures one makes in their lifetime. Often time people chose agents to show them their desired type of house within the budget. As the industry is booming, oftentimes people are misled because of a higher commission pay. In this report, we will be analyzing Housing prices by their features and attributes so that you get an estimate of what the real price of the house may be compared to what the realtor might be offering you using some of the basic regression algorithms like Linear Regression, Random Forest, and XG Boost and compare the results.

Keywords— Regression Techniques, Machine Learning, Algorithms.

I. INTRODUCTION

As buying a new house has to be one of the biggest purchasing decision that people make in their lifetime, our project aims to predict sale prices of houses by means of various features within residential homes. The information gathered could be used by individuals to complement their decision making process when purchasing a house, which cuts dependency from realtors with overpriced houses.

We aim to use simple Machine Learning algorithms to help successfully determine housing prices so that people have it easier when looking into the houses they desire matching up with their budget. The dataset we currently are using is a version uploaded in Kaggle for an open competition with fixed amount of variables stored as integers and factors.

II. DATA SET: HOUSING PRICES IN KAGGLE

The Kaggle dataset that is used in this project is mainly composed of multiple categorical variables stored as integers and factors, both discrete and continuous. Here the mechanism will be using is mainly descriptive and predictive analytics, with suggestions on how additional data could be obtained to conduct more experiments to enhance the purchasing experience.

The data set used in this project describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. The original training dataset has a total of 1460 observations with eighty one variables while the test dataset has one less variable since we have excluded the predicted variable. Both of these datasets are combined together to ensure consistency in pre-processing.

Id	MSSubClas	MSZoning	LotFrontag	LotArea	Street	Alley	LotShape	LandContor
1	60	RL	65	8450	Pave	NA	Reg	Lvl
2	20	RL	80	9600	Pave	NA	Reg	Lvl
3	60	RL	68	11250	Pave	NA	IR1	Lvl
4	70	RL	60	9550	Pave	NA	IR1	Lvl
5	60	RL	84	14260	Pave	NA	IR1	Lvl
6	50	RL	85	14115	Pave	NA	IR1	Lvl
7	20	RL	75	10084	Pave	NA	Reg	Lvl
8	60	RL	NA	10382	Pave	NA	IR1	Lvl
9	50	RM	51	6120	Pave	NA	Reg	Lvl
10	190	RL	50	7420	Pave	NA	Reg	Lvl
11	20	RL	70	11200	Pave	NA	Reg	Lvl
12	60	RL	85	11924	Pave	NA	IR1	Lvl
13	20	RL	NA	12968	Pave	NA	IR2	Lvl
14	20	RL	91	10652	Pave	NA	IR1	Lvl
15	20	RL	NA	10920	Pave	NA	IR1	Lvl

Utilities	LotConfig	LandSlope	Neighborhc	Condition1	Condition2	BldgType	HouseStyle	OverallQua
AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7
AllPub	FR2	Gtl	Veenker	Feedr	Norm	1Fam	1Story	6
AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7
AllPub	Corner	Gtl	Crawfor	Norm	Norm	1Fam	2Story	7
AllPub	FR2	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8
AllPub	Inside	Gtl	Mitchel	Norm	Norm	1Fam	1.5Fin	5
AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	8
AllPub	Corner	Gtl	NWAmes	PosN	Norm	1Fam	2Story	7
AllPub	Inside	Gtl	OldTown	Artery	Norm	1Fam	1.5Fin	7
AllPub	Corner	Gtl	BrkSide	Artery	Artery	2fmCon	1.5Unf	5
AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5
AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	2Story	9
AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5
AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	7
AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	6

Figure 1: Snapshot of the Training Data Set, Unprocessed

In this dataset there are a lot of N/A included as many of these information were not available for many of the instances. For these cases, we will be using numeric values to fill in and have an approximate prediction. These variables leave room for more predictive analysis for future researchers.

The variables that are more considered than others in this research work for analyzing house prices within this data set are:

- Total number of bathrooms
- Additional Remodeling Status
- Age of house
- Landscape of the house
- Garage quality (Room for number of cars)
- Total square feet of the house

III. METHODOLOGY & DATA ANALYSIS

In this section we are going to be looking into the algorithms that we have used for the predictive analysis of housing prices using our dataset. But first the data was required to be pre-processed, and for our project we applied the following techniques:

- Render out missing values with numerical ones
- Delete any anomaly and outliers
- Convert categorical values into discreet ones

- Transform Skewed attributes
- Finalize Attributes

Our Skewness:

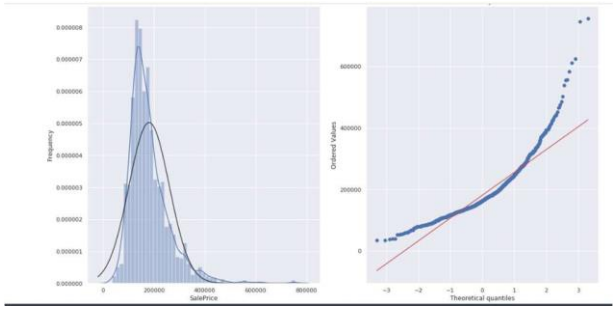

Figure 2: Before pre-processed Skewness From The Data Set

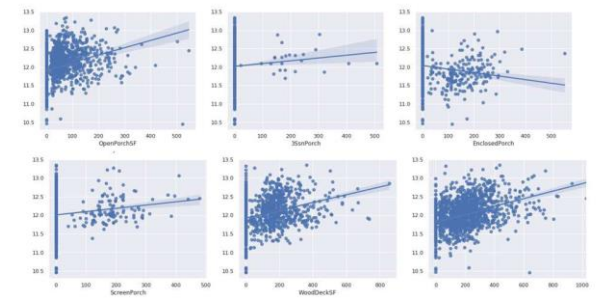

Figure 3: After pre-processed Skewness

A. Linear Regression

Linear Regression is an ML algorithm based on supervised learning, which is what we are doing in this project. Regression model generates a forecast value based on independent variables. It is mostly used for finding out the relationship between variables and predicting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

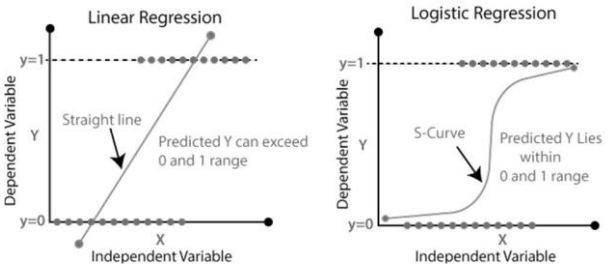

Figure 4: Linear Regression Vs. Logical Regression

Linear regression performs the work to predict an assigned variable value based on a given independent variable. So, this regression technique finds out a linear relationship between input and output. We also have should point out that since there were many "N/As", some variables turn out to be perfectly correlated to each other, this is a weakness in this form of regression since other variables are not taken into account, as compared to a logistic regression, it gives out a more dynamic result.

B. XG Boost

XG Boost is a decision-tree-based resembling ML algorithm that uses a gradient boosting framework. In forecasting problems involving unstructured data such as images, text, etc. and artificial neural networks is more likely to outperform all other algorithms or frameworks.

However, when it comes to small to medium structured/tabular data such as ours given in our data sets, decision tree based algorithms are considered best-in-class right now.

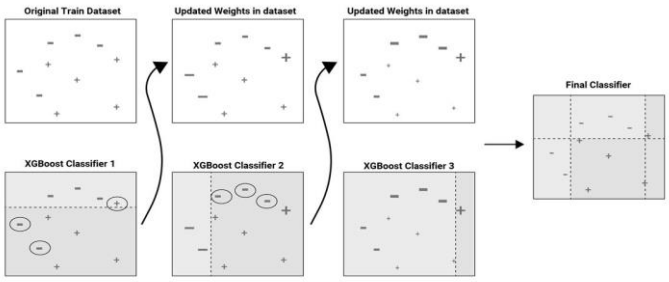

Figure 5: XG Boosting Approach

Being a non-parametric learning model, XG Boosting does not require variables to follow a normal distribution line. And this is crucial since housing prices cannot be explained simply through a straight line. For this out input data is required to be in a matrix format.

Due to the preprocessing of the data we have performed earlier, there were no longer any N/As in the data and we could proceed on with the analysis.

C. Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

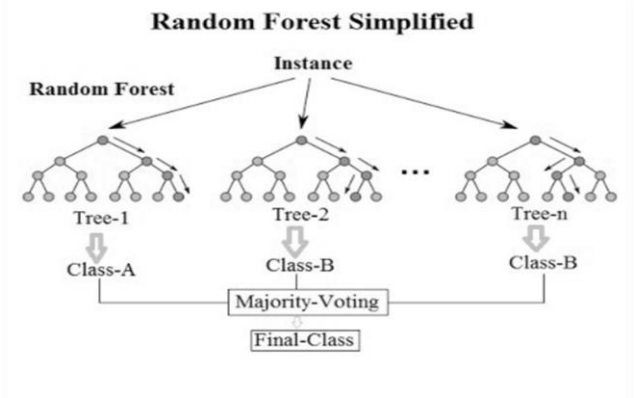

Figure 6: Random Forest Procedure

Random decision forests correct for decision trees' habit of overfitting to their training set. Here we use minimum weight for the procedure using the given equation.

$$MSE = \frac{\sum_{samples} (Price_{sample_i}{}^2 - value^2)}{total\ no\ of\ smples}$$

$$Weighted\ MSE = \frac{(NumSamplesNode1 * MSE1) + (NumSamplesNode2 * MSE)}{NumSamplesNode1 + NumSamplesNode2}$$

Minimum Weighted MSE Calculation for First Split of above.

D. LASSO & RIDGE Regrassion

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models.

Ridge Regression is a technique for analyzing multiple regression data that suffer from multi-co-linearity. When multi-co-linearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value.

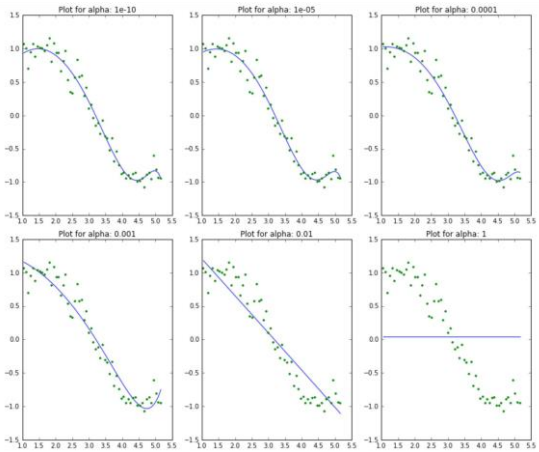

Figure 7: Example of Lasso & Ridge Regression Models

IV. DATA ANALYSIS

After implementation of five mentioned algorithms in our dataset it is cleared which algorithms gave us the best accuracy. Our score from the whole analysis within dataset:

Model	Accuracy
ElesticNet	0.111987
Lasso	0.112035
Ridge	0.112406
STACK	0.113569
XGBOOST	0.113794
LightGBM	0.117132

Boosting Algorithms like XG Boost and Light GBM have performed the best, whereas the more regression, because many attributes are merged due to N/A values present in the dataset. We have represented the following data in a chart as a snapshot from out project:

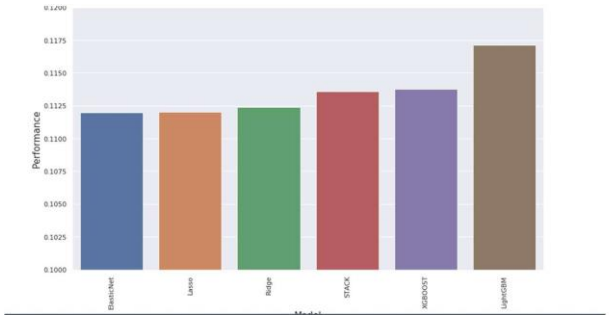

Figure 8: Performance Output

V. CONCLUSION

In this report, we have presented dataset gathered from Kaggle. We presented our method of collecting and annotating the dataset. We investigated the properties and the statistics of the dataset and tried to implement some regression algorithms. The dataset in this examination is relied upon to be utilized for arrangements which utilized machine learning based statistical calculations, for example, Linear Regression, Random Forest, and Boosting Algorithms. We also learned that algorithms that adapt more features and attributes are more likely to give a better performance than the ones that minimize anomalies.

Finally, it can be said that through machine learning if machine can analyze the housing prices prediction, people can minimize the use of realtors and get more accurate house settings in their preferred budget.

REFERENCES

[1] H. Hirose, Y. Soejima and K. Hirose, "NNRMLR: A Combined Method of Nearest Neighbor Regression and Multiple Linear Regression," 2012 IIAI International Conference on Advanced Applied Informatics, Fukuoka, 2012, pp. 351-356, doi: 10.1109/IIAI- AAI.2012.76

[2] X. Sun, Z. Ouyang and D. Yue, "Short-term load forecasting based on multivariate linear regression," 2017 IEEE Conference on Energy Internet and Energy System Integration (EI2), Beijing, 2017, pp. 1-5, doi: 10.1109/EI2.2017.8245401.

[3] M. Ross, C. A. Graves, J. W. Campbell and J. H. Kim, "Using Support Vector Machines to Classify Student Attentiveness for the Development of Personalized Learning Systems," 2013 12th International Conference on Machine Learning and Applications, Miami, FL, 2013, pp. 325-328, doi: 10.1109/ICMLA.2013.66.

[4] Yi Tan and Guo-Ji Zhang, "The application of machine learning algorithm in underwriting process," 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 2005, pp. 3523-3527 Vol. 6, doi: 10.1109/ICMLC.2005.1527552.