International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

House Rent Price Prediction

Adarsh Kumar

[1]*B.E., Department of Information Science and Engineering, The National Institute of Engineering, Mysuru, India*

Abstract - *Determining the sale price of the house is very important nowadays as the price of the land and price of the house increases every year. So our future generation needs a simple technique to predict the house price in future. The price of house helps the buyer to know the cost price of the house and also the right time to buy it. The right price of the house helps the customer to elect the house and go for the bidding of that house. There are several factors that affect the price of the house such as the physical condition, location, landmark etc. This paper uses various regression techniques to predict the house price such as Ridge, Lasso, ElasticNet regression techniques.*

Key Words: House Price, Ridge, Lasso (Least Absolute Shrinkage and Selection Operator), Regression analysis.

1.INTRODUCTION

Real property is not only a man's basic need, but it also represents a person's wealth and prestige today. Because their property values do not decline rapidly, investment in real estate generally seems to be profitable. Changes in the price of real estate can affect various investors in households, bankers, policy makers and many others. Investment in the real estate sector appears to be an attractive investment choice. Predicting the value of immovable property is therefore an important economic index.

In this study, several methods of prediction were compared t o finding the best predicted results in determining a house's selling price compared to the actual price.

This paper brings the latest research on regression technique that can be used for house prediction such as Linear regression, Ridge regression, Gradient boosting and hybrid regression.

As the initial house price prediction were challenging and require some best method to get accurate prediction. Data quality is a key factor to predict the house prices and missing features are a difficult aspect to handle in machine learning models let alone house prediction model. Therefore, feature engineering becomes an important method for creating models which will give better accuracy.

In general, the value of the property increases over time and its valued value must be calculated. During the sale of property or while applying for the loan and the marketability of the property, this valued value is required. The professional evaluators determine these valued values. However, the disadvantage of this practice is that these evaluators could be biased because buyers, sellers or mortgages have bestowed interest. We therefore need an automated model of prediction that can help to predict property values without bias. This automated model can help first - time buyers and less experienced customers to see if property rates are overrated or underrated.

Kaggle organizes a dataset "Ames housing" and it provides data with 82 explanatory variables for part of residential home transactions in Ames, Iowa, and opens to all to predict price of each covered home transaction SalePrice. The major factors are MS zoning, salePrice, Overall Qual, Overall Cond, Year Built, Year Remod/Add, Sale Condition, SalePrice. The dataset contains various missing critical features which can degrade the model performance to predict house prices.

2. DATA ANALYSIS

The dataset contains two types of variables:

1. Numeric Variables

There are 36 numerical features that are relevant. M SSubClass, which "identifies the type of residence involve d in the sale," is encoded as numeric but is a categorical va riable in reality.

There are 36 numerical features, of the following types:

Square footage: shows the square footage of some features, i.e. 1stFlrSF (first floor square footage) and GarageArea (square feet garage size).

Time: variables related to time, such as when the house was built or sold.

Room and amenities: data representing amenities such as "How many bathrooms".

Condition and quality: Most of the variables dealing with the apartment's actual physical space are positively skewed— which makes sense because people tend to live in smaller homes / apartments apart from the extremely rich.

2. Categorial Variables

There are many categorial variables as numeric variables. But, there are many sales prices that don't change with categories. There are also features that don't vary in price a lot among different categories, including the roof style and land slope. Some include the presence or absence of

central air, the neighborhood, the external quality, and the zoning.

3. METHEDOLOGY

The design approach involves pre-processing of data, creative feature engineering and the regression model such as ridge regression, Gradient boosting, Linear regression and Hybrid regression.

3.1 Data Pre-processing

The data pre-processing involves:

a) **Removing stop words**: Stop words take 70% of the text in the dataset. It is necessary to remove stop words. To remove stop words from dataset, NLTK corpus for stop words is used to check for stop words.
b) **Case folding**: At this stage, all words are made to same case such as lower case or upper case.
c) **Stemming**: Stemming is the process of producing a root / base word's morphological variants. Stemming programs are commonly called stemming or stemming algorithms.
d) **Filling NaN Values:** Many of the variables had-1 values that had to be addressed. Based on what made the most sense, those values were filled out accordingly. For instance,-1 values for values such as Alley were filled with a string ("No Alley"), whereas GarageYrBuilt-1 values were filled with the median to prevent data from being skewed.
e) **Dummy Variables:** Categorial variables are string which pose a threat to models. It is better to create dummy variables which are numeric constant for categorial variables which will help the models to operate on categorial variables.

3.2 Feature Engineering

We conducted set of feature engineering step such as:

- Reduce the number of categorical variables whenever possible / appropriate, as each categorical variable must be converted into multiple dummy variables for regular multiple linear regression models (OLS) and regularized linear regression models, for example. Ridge, Lasso, and ElasticNet, which would significantly increase the total number of variables and make prediction very inefficient.
- Add new variable promising feature based on knowledge of the domain.
- Remove trivial variables of very low prediction value.
- Adjust variables as required to ensure that their values or types are fit for regression purposes, i.e. help to predict the target variable accurately.

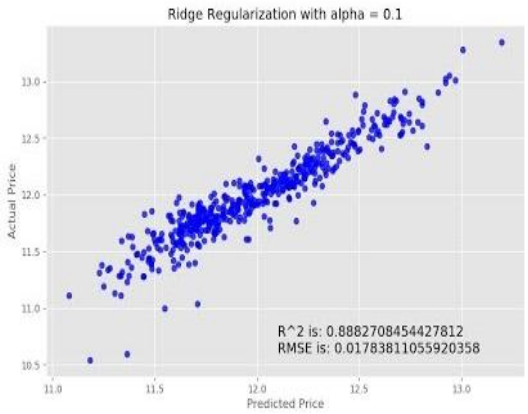

Fig -1: Ridge regression with alpha= 0.1

Specifically, we ran RF model fit on the training data in R for the first time and collected information about the feature importance of all variables. Then, using Caret's recursive feature elimination (RFE) function, we sorted all the feature variables from the most important to the least, and performed backward selection of features by dropping 5 variables at a time and comparing their predictive performance of cross-validation (CV). As such, we can find the optimal set of features that gives error of prediction to the lowest RMSE (root mean squared error). The result figure is shown below that will give the best RMSE performance by dropping 10 least important variables.

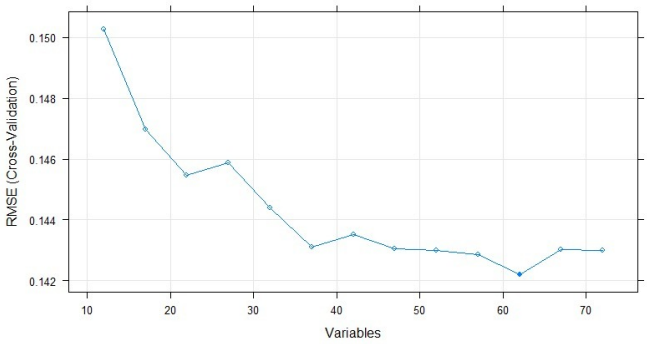

Fig -2: Comparison Graph

3.3 Regression Algorithms

a) **Linear Regression:** For finding a relationship between two continuous variables, Linear regression is useful. One variable is predictor or independent, and the other variable is variable response or dependent. It looks for a relationship that is

statistical but not deterministic. It is said that the relationship between two variables is deterministic if the other can express one variable accurately.

$$Y = \theta_1 * X + \theta_0$$

where Y is the dependent variable, X is the independent variable. Theta is the coefficient factor.

b) Ridge Regression: Ridge Regression is a multi-regression data analyzing technique suffering from multicollinearity. Most square estimates are unbiased when multicollinearity occurs, but their variances are large so that they may be far from the true value. The ridge regression reduces standard errors by adding a degree of bias to the regression estimates. It is hoped that the net effect will be to provide more reliable estimates. Ridge regression is modifying the least squares method which to allow to have biased estimators of the regression coefficients in the regression model. Ridge regression put a particular form of constraints on parameters.

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

c) Gradient Boosting: Gradient boosting is a machine learning technique for regression and classification issues that generates a predictive model in the form of a set of weak predictive models, typically decision trees.

The objective is to define a loss function and reduce it. For gradient boosting, MSE is as follows:

$$Loss = MSE = \sum (y_i - y_i^p)^2$$

where, y_i = ith target value, y_i^p = ith prediction, $L(y_i, y_i^p)$ is Loss function

The idea behind gradient boosting is to leverage the residual patterns and strengthen a model with weak predictions and make it better. Once we reach a stage that residuals do not have any pattern that could be modeled, we can stop modeling residuals.

d) Lasso Regression: Lasso (Least Absolute Shrinkage and Selection Operator), similar to Ridge Regression, also penalizes the absolute size of the coefficients of regression. It is also capable of reducing variability and enhancing linear regression models ' accuracy. Look at the equation below:

$$\sum_{i=1}^{n} (y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

The regression of Lasso differs from the regression of the ridge in a way that uses absolute values instead of squares in the penalty function. This leads to a penalization (or equivalent limitation of the sum of the absolute values of the estimates) which causes some estimates of the parameters to turn out to be exactly zero. The greater the penalty applied, the estimates are further reduced to absolute zero.

4. RESULT

To train the dataset and make predictions separately, I used LASSO (least absolute shrinkage and selection operator) and Gradient boosting regression models.

LASSO is a model of regression that selects and regularizes variable. The model LASSO uses a parameter that penalizes too many variables for fitting. It allows variable coefficients to be reduced to 0, which essentially results in the model having no effect on those variables, thus reducing dimensionality. Because there are quite a few explanatory variables, reducing the number of variables can increase the accuracy of interpretation and prediction.

One of Kaggle's most popular algorithms is gradient boosting models. For many recent competitions, a variant of GBMs known as the XGBoost was a clear favorite. Out of the box, the algorithm works well. Such as the random forest, it is a type of ensemble model where multiple decision trees are used and optimized over some cost function. The popularity and ability to compete well are reasons enough to use this type of model for the problem of house price prediction.

I used cross validation and RMSLE (Root means logarithmic error) to see how well each model is performing.

Here are the scores I got from my two models

Lasso Score: 0.1115(0.0078)

GBoost Score: 0.1159(0.0088)

Went ahead and stacked the two models together, several research papers have proved that stacking two models together can improve the accuracy of what is called ensembling.

Average base model score:0.1091(0.0077)

Finally, I trained the stacked regressor and predicted it.

RMSLE score on the train data: 0.06870805

Accuracy score: 0.9795672

dataset	model	name	preprocessing	score
test	LassoCV	lasso	scaled	0.8708811803709471
test	ElasticNetCV	elastic net	scaled	0.8552059992855636
test	RidgeCV	ridge	scaled	0.8325365372218565

Table -1: Top performing models score

5. CONCLUSIONS

The most value is removed by roofing a home with clay tile. Interestingly, being close to a park or other outdoor feature also lowers the home's value. Alternatively, the value is increased by a few neighborhoods. On this dataset, regularized models perform well. A note on the tradeoff bias/variance: the model fit by the Ordinary Least Squares (OLS) is the least biased estimator of all possible estimators, according to the Gauss-Markov theorem. In other words, it fits the data it has seen better than all possible models.

REFERENCES

[1] https://www.kaggle.com/c/house-prices-advancedregression-techniques/

[2] http://scikit-learn.org/stable/install.html

[3] https://github.com/dmlc/xgboost/

[4] Eli Beracha, Ben T Gilbert, Tyler Kjorstad, Kiplan womack, "On the Relation between Local Amenities and House Price Dynamics", Journal of Real estate Economics, Aug. 2016.

[5] Stephen Law, "Defining Street-based Local Area and measuring its effect on house price using a hedonic price approach: The case study of Metropolitan London", Cities, vol. 60, Part A, pp. 166–179, Feb. 2017.

[6]https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

[7]https://www.researchgate.net/publication/323135322_A_hybrid_regression_technique_for_house_prices_prediction