House Price Prediction Using Machine Learning Algorithm

Akshata Kudavkar
Symbiosis International University

House Price Prediction Using Machine Learning Algorithm

Akshata Ashok Kudavkar
BE.IT, Pillai HOC College Of Engineering and Technology
Navi Mumbai, India
kudavkarakshata27@gmail.com

Manali Namdev Nhavkar
BE.IT, Pillai HOC College Of Engineering and Technology
Navi Mumbai, India
manalinhavkar12@gmail.com

Samiksha Subhash Wagh
BE.IT, Pillai HOC College Of Engineering and Technology
Navi Mumbai, India
samikshawagh09@gmail.com

Shamna Sadanand
Pillai HOC College Of Engineering and Technology
Navi Mumbai, India
shamnas@mes.ac.in

ABSTRACT

The real estate market is one of the most competitive in terms of prices and it tends to vary significantly based on a lot of factors, hence it becomes one of the prime fields to apply the concepts of machine learning. Therefore in this project, we present various algorithms while predicting house prices with good accuracy. We tested a regression models such as Simple Linear Regression, Ridge Regression, Lasso Regression, Support Vector Regression, Random Forest Regression, Decision Tree Algorithm and selected the best fit among the algorithm. This project directs us that it can be best application of machine learning models in order to optimize the result.

Keywords - House prediction: regression analysis

1. INTRODUCTION

Machine learning has been used for many years to offer image recognition, spam detection, natural speech comprehension, product recommendations and medical diagnoses. Today, machine learning algorithms can help us to enhance cyber security, ensure public safety, and improve medical outcomes. In this project we used a machine learning concept, For example, if we're going to sell a house, we need to know what price tag to put on it. Here the machine learning algorithm can give us an accurate estimation or prediction.

Predicting housing prices has always been a challenge for many machine learning engineers. Several researchers have tried to come with a model to accurately predict housing prices with high accuracy and least error. Our goal for this project was to use regression models and classification techniques in order to predict the sale price of a house.

These models are created using various features such as square feet of the house, number of bedrooms, year of construction, property type etc. Some of the researchers have used techniques like clustering for grouping same houses together and then estimating the price. In this project we tested a regression models like Simple Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regression, Support Vector Regression, Decision Tree Regression and will choose the best fit among the calculation.

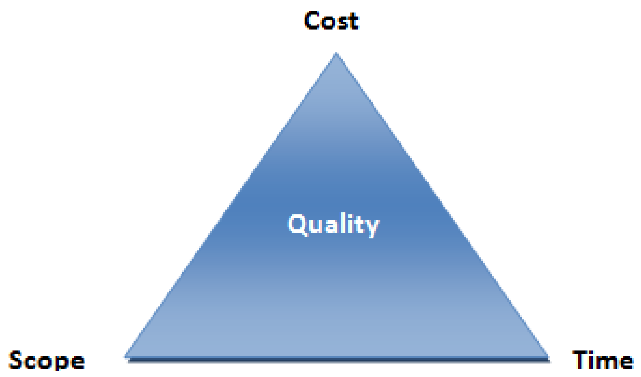
2. PURPOSE

The purpose of the project is to predict house prices with the help of various machine learning-regression algorithms and select the models with the best accuracy score to predict house prices. To evaluate the utility of machine learning models to estimate prices on samples of their housing dataset.

To develop user friendly house price predicting system which reduces the man power. House price prediction can help the developer to determine the selling price of a house and can help the customer to arrange the right time to purchase a house.

3. CONSTRAINTS

We here define the constraint using the triple constraint of project management:



Management

Cost:As no hardware elements are required in this system, the constraints are very few. So the cost of the requirements Cost: As no hardware elements are required in this system, the constraints are very few. So, the cost of the requirements for the project are very less. Machine learning algorithms require systems with high processing power i.e. systems with high Random Access Memory (RAM) are required For the smooth functioning of all the machine learning models, Anaconda - Python data science platform is installed. Python libraries - Numpy, Pandas and seaborn should be installed in the system. Tableau - Data visualization tool is also to be installed on the system.

Time: The time required for developing the project is depended on the complexity of the project and also the number of modules. According to the current specification of the system it will require almost 3 – 4 months for deploying the modules on the respective machine.

Scope: The House Prediction dataset is imported from Kaggle in Comma Separated Values (csv) format. The dataset is analyzed with the help of pandas, numpy and scikit-learn. Tableau is used as a data visualization tool. After drawing insights from the dataset with the help of Tableau, we identify the important factors i.e. factors majorly affecting the change in prices. The factors adding insignificant values to the overall result are omitted. The dataset is divided into two parts - training set and testing set. The various machine learning models are trained with the help of the training set. The testing set is then used to check the performance of all the machine learning models. Accuracy score is calculated and ıfusion matrix is generated. Root Mean Square ·or of all the models is calculated. In the final p the model with the highest accuracy score ı the least RMSE (Root Mean Square Error) ue is used for predicting house prices. The :come of our project will be predicting house ces with good accuracy and that can help the ıtomer as well as developer.

4. OVERALL SYSTEM DESCRIPTION

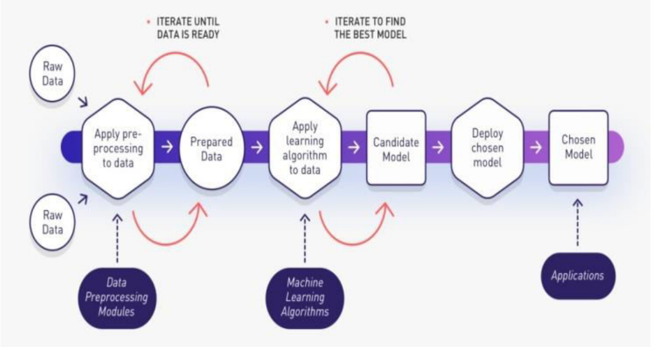
A.Existing System

In the existing system there are so many solutions for house’s sales price prediction problem for one of the Kaggle competitions, in which they combine standard machine learning algorithms with their original ideas like residual regression, logit transform and neural network machine. But during data analysis the results show that the house price variation prediction results is not accurate enough. Sometimes the Standard deviation of the results is very high because of small dataset size.

B. Proposed System

The system that we proposed solves most of the problems that we have with the existing system. In our system the accuracy of the prediction is mostly correct as compared to existing system. There are several factors that affect house prices. We divide these factors into three main groups, there are physical condition, concept and location. Physical conditions are properties possessed by a house thatcan be observed by human senses, including the size of the house, the number of bedrooms, the availability of kitchen and garage, the availability of the garden, the area of land and buildings, and the age of the house while the concept is an idea offered by developers who can attract potential buyers, for example, the concept of a minimalist home, healthy and green environment. Location is an important factor in shaping the price of a house. This is because the location determines the regnant land price. Additionally, the area likewise decides the easy access to public facilitates, for example, schools, grounds, clinics and wellbeing focuses, just as family diversion facilities, for example, shopping centers, culinary visits, or even offers wonderful view. By using this all factors we optimized the result.

C. System Architecture


ITERATE UNTIL DATA IS READY
ITERATE TO FIND THE BEST MODEL
Raw Data
Raw Data
Apply pre-processing to data
Prepared Data
Apply learning algorithm to data
Candidate Model
Deploy chosen model
Chosen Model
Data Preprocessing Modules
Machine Learning Algorithms
Applications

5. METHODOLOGY

In our project, the House Prediction dataset is imported from Kaggle in Comma Separated Values (csv) format. The dataset is analyzed with the help of pandas, numpy and scikit-learn. Tableau is used as a data visualization tool. After drawing insights from the dataset with the help of Tableau, we identify the important factors i.e. factors majorly affecting the change in prices. The factors adding insignificant values to the overall result are omitted. The dataset is divided into two parts - training set and testing set. The various machine learning models are trained with the help of the training set. The testing set is then used to check the performance of all the machine learning models. Accuracy score is calculated. Root Mean Square Error of all the models is calculated. In the final step the model with the highest accuracy score and the least RMSE (Root Mean Square Error) value is used for predicting house prices.

1. Simple Linear Regression

In simple linear regression, we predict scores of one variable from the scores on a second variable. The formula for a regression line is

Y' = bX + A

Where, Y' is the predicted score i.e. dependent variable, x is an independent variable, b is the slope of line and A is the Y-intercept.

When there is only one predictor variable x the prediction method is called as a simple linear regression.

2. Ridge Regression

Ridge Regression: In ridge regression, the loss or errors is altered by adding a penalty equivalent to square of the magnitude of the coefficients. i.e. it reduces the loss or errors by adding a penalty. Ridge regression is used to reduce complexity and cost function. The formula for ridge regression is,

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}w_j^2$$

In simple terms,

Ridge R = loss+ λ||w||^2

Here, λ is constant,

||w||^2= w$_1^2$+w$_2^2$+w$_3^2$……. Here, w is a vector of coefficient.

So ridge regression puts limitations on the coefficients *(w).* The penalty term (lambda) regularizes the coefficients with the end goal that if the coefficients take expansive qualities the enhancement work is penalized. Along these lines, ridge relapse shrivels the coefficients and it helps the model unpredictability and multi-collinearity.

3. Lasso Linear Regression

Lasso Regression: For the Lasso regression i.e.(least absolute shrinkage and selection operator) regression the cost function can be written as

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|w_j|$$

In simple terms,Lasso = loss+ λ||w||.Lasso regression coefficients; subject to similar constrain as Ridge.

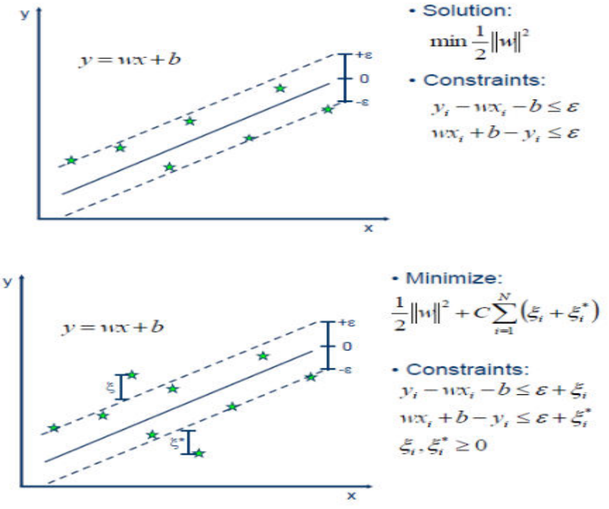
Just like Ridge regression cost function, for λ=0. The only difference is instead of taking the square of the coefficients, magnitudes are taken into account. This type of regularization can lead to zero coefficients i.e. some of the features are avoidedto generate the outputs. So Lasso regression not only helps in reducing loss/errors of models but it can help us in feature selection.

4. Support Vector Regression

SVM is a supervised learning algorithm which is widely used for classification algorithms. SVM is applicable only for those data that are linearly separable. For non-linear data kernel functions are used.

SVM classifies the two classes using "hyperplane". The hyperplane should have the largest margin in a high dimensional space to separate given data into classes.

The margin between the two classes represents the longest distance between closest data points of those classes.


y
$y = wx + b$
+ε
0
-ε
x
• Solution:
$\min \frac{1}{2}\|w\|^2$
• Constraints:
$y_i - wx_i - b \le \varepsilon$
$wx_i + b - y_i \le \varepsilon$
y
$y = wx + b$
+ε
0
-ε
ξ_i
ξ_i^*
x
• Minimize:
$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$
• Constraints:
$y_i - wx_i - b \le \varepsilon + \xi_i$
$wx_i + b - y_i \le \varepsilon + \xi_i^*$
$\xi_i, \xi_i^* \ge 0$

5. Decision Tree Regression

Decision tree is a tree shaped figure which is used to determine a course of action. Each branch of the tree represents a possible decision, transpire or reaction.

This algorithm makes a classification decision for a test sample with the help of tree like structure.

The nodes in the tree are attribute names of the given data. Branches are attribute values and leaf nodes are the class labels.

The advantages of using this algorithm in house price prediction are:-

1. It is simple to understand, interpret and visualize.
2. Little effort required for data preparation.
3. It can handle both numerical and categorical data.

6. Random Forest Regression

Random forest regression develops lots of decision tree based on random selection of variables. It provides the class of dependent variable based on many trees.

1. Random selection of data:-

original data= subset 1+subset 2+subset 3+.....

This subsets each can have different size of the observation, there can be some overlapping or cannot.

2. Random selection of variables:-

If we have variables x1, x2 ...xn independent variables, which can be used for developing decision tree.

We divide this variable into different sets like,

variable set 1- x1,x3,...

variable set 2- x3,x4,...

As the trees are based on random selection of data as well as variables, these are random tree. Many such random trees leads to a random forest. When we have many trees we get a forest, similarly when we have many decision trees it is a random forest. There are two major belief that helps us to used this trees:

1. Most of the trees can provide correct prediction of class for most part of the data.
2. The tree are making mistakes at different places.

- Regression Results:

Regression	Accuracy Score
Linear Regression	88.82
Lasso	78.56
Ridge	88.83
Random Forest Regression	89.56
SVR (Gaussian kernel)	11.138
Decision Tree Regression	79.56

Prediction and Real Value of a test case with different Regression methods:

Regression	Real Value	Predicted Value
Linear Regression	11.767	11.622
Lasso	11.767	11.566
Ridge	11.767	11.621
Random Forest Regression	11.767	11.462
SVR (Gaussian kernel)	11.767	12.04
Decision Tree Regression	11.767	11.462

- **Models and Results**

For regression models, we try to solve the following problem: given a processed list of features for a house, we would like to predict its potential sale price. Linear regression is a natural choice of baseline model for regression problems. So we first run linear regression including all features, using our 81 features and 1461 training samples. The model is then used to predict sale prices of houses given features in our test data and is compared to the actual sale prices of houses given in test data set. The performance was measured by the Accuracy Score of the predicted results and the actual results. Our baseline model generated an accuracy score of 88.82.

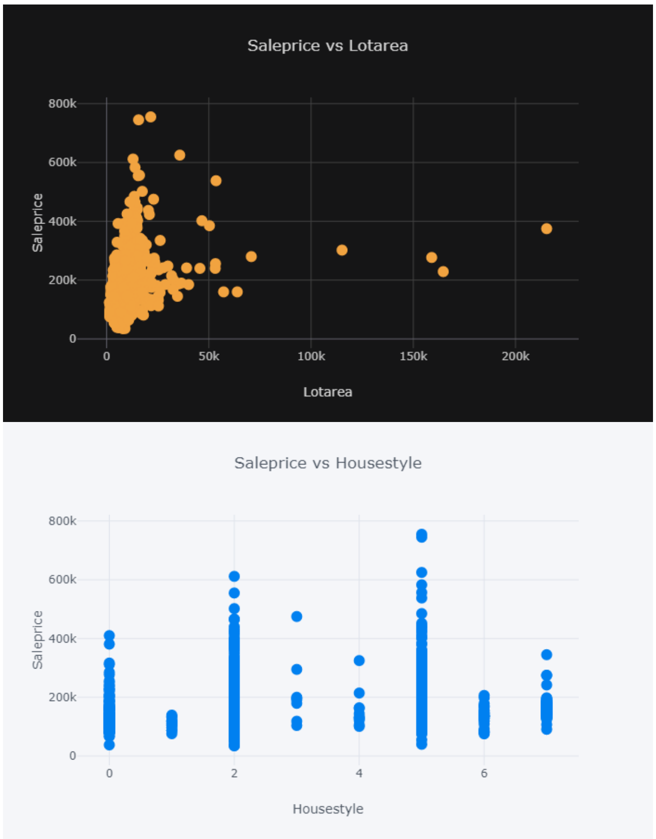
After using linear regression model as the baseline model, we included the regularization parameters in linear regression models to reduce overfitting. Linear regression with Lasso after 5-fold cross validation generated an accuracy score of 78.56, which is lesser than our baseline model. Also, linear regression with lasso automatically picked 56 variables and eliminated the other 35 variables to fit in the model.

Other than lasso regularizer, we also applied ridge regularizer with cross validation in our linear regression model, which generated an accuracy score of 88.83. This score is also better than our baseline model, meaning that regularized linear regression helped with overfitting.

Support vector regression (SVR) with Gaussian kernel is also fitted to the features. Parameters Cs of both models are cross validated to pick the best performing parameters. SVR with Gaussian kernel model generated a score of 11.138.

Then, we fitted our training dataset with random forest regression model, with max_depth parameter cross validated to be 150. Our random forest regression model generated an accuracy score of 89.56, which is also better than our baseline model. Lastly, we applied the Decision Tree Classifier to the dataset which gave an accuracy score of 79.56. Overall, the random forest classifier and ridge classifier models performed better than the basic linear regression model. The highest accuracy score is achieved by the Random Forest Classifier. We suggest that this regression model be used for future house price predictions.

- Visualizations andAnalysis:



This histogram, plot depicts the lot size in square feet and house style with respect to the price range.

Conclusion

So we conclude that the system that we proposed solves most of the problems that we have with the existing system.

After training and testing of datasets with all models, the random forest classifier and ridge classifier models performs better than the simple linear regression model. The highest accuracy score is achieved by the Random Forest Classifier. So, we suggest that this regression model be used for future house price predictions. Therefore, the outcome of our project will be predicting house prices with good accuracy which can help the customer as well as developer.

Acknowledgment

We remain immensely obliged to **Prof. Shamna Sadanand** for providing us with the moral and technical support and guiding us. We would also like to thank our guide for providing us with her expert opinion and valuable suggestions at every stage of project.

We would like to take this opportunity to thank **Prof. Monisha Mohan**, Head of Information Technology for her motivation and valuable support. This acknowledgement is incomplete without thanking teaching and non-teaching staff of department of their kind support.

We would also like to thank **Dr. Madhumita Chatterjee,** Principal of Pillai HOC college of Engineering and Technology, Rasayani for providing the infrastructure and resources required for project.

Reference

- " Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning" published in 2018 IEEE.https://www.kaggle.com/c/house-prices-advanced-regression-techniques
- "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization", (IJACSA) International Journal of Advanced Computer Science and Applications, published in 2017.
- "Prediction of Real Estate Price Variation Based on Economic Parameters" Proceedings of the 2017 IEEE International Conference on Applied System Innovation IEEE-ICASI 2017 - Meen, Prior & Lam (Eds).
- "Waiting to be Sold: Prediction of Time-Dependent House Selling Probability", 2016 IEEE International Conference on Data Science and Advanced Analytics