USA Housing listing price Prediction

Pratik Patil(202118023) Devanshi Shah(202118042)

Abstract

House price listing is an important topic of real estate. Our project attempts to derive useful knowledge from historical data of property markets. Machine learning techniques are applied to analyze historical property transactions in USA to discover useful models for house buyers and sellers. Revealed is the high discrepancy between house prices in the most expensive and most affordable suburbs in the USA. Moreover, experiments demonstrate that the Linear Regression that is based on mean squared error measurement is a competitive approach and also applied Random Forest Regression.

1 Introduction

This study uses two machine learning algorithms including, Linear Regression and Random Forest Regression in the appraisal of property prices. It applies these methods to examine a data sample of about 40,000 housing transactions in a period of over 18 years in USA, and then compares the results of these algorithms. In terms of predictive Linear Regression and Random Forest Regression have achieved better performance. The two performance metrics including root mean squared error (RMSE) and R-squared associated with these two algorithms. Our conclusion is that machine learning offers a promising, alternative technique in property valuation and appraisal research especially in relation to property price prediction.

2 Data Description

The data set is taken from kaggle (BDP Project/USA Housing Listings/housing.csv) its of marketing agencies having customers who are using their service to publish ads for client websites. The data set has fields like ID and Sale Price.

3 Data Exploration

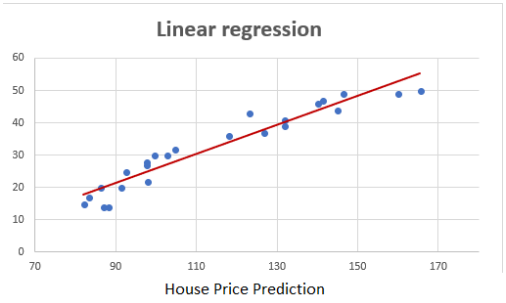
Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software Python. Before it can conduct analysis on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working.We divided the data for Training and Testing purpose respectively.

4 Model Implementation

4.1 Linear Regression

Linear Regression is a supervised machine learning model that attempts to model a linear relationship between dependent variables (Y) and independent variables (X). Every evaluated observation with a model, the target (Y)'s actual value is compared to the target (Y)'s predicted value, and the major differences in these values are called residuals. The Linear Regression model aims to minimize the sum of all squared residuals.

- Linear Regression is a machine learning algorithm based on supervised learning.
- It performs a regression task. Regression models a target prediction value based on independent variables.
- It is mostly used for finding out the relationship between variables and forecasting.



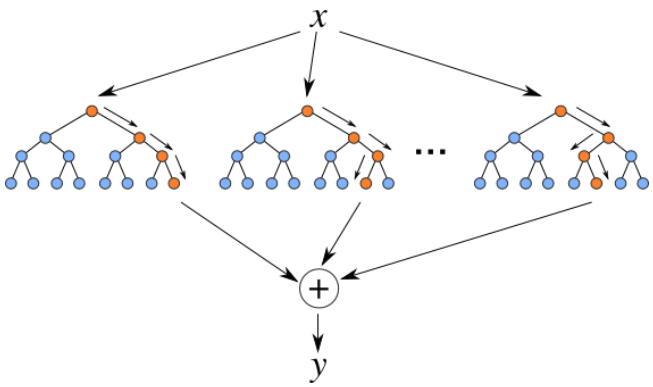
4.2 Random Forest Regression

A random forest is a machine learning technique for classifying and predicting outcomes. It employs en- semble learning, a method for resolving complex problems by merging many classifiers. Many decision trees make up a random forest algorithm. Bagging or bootstrap aggregation is used to train the 'forest' generated by the random forest method. Bagging (Bootstrap Aggregation) Decision trees are extremely sensitive to the data they're trained on, and even minor modifications to the training set might result in drastically different tree architectures. Random forest takes use of this by enabling each tree to sample from the dataset at random with replacement, resulting in unique trees. Bagging is the term for this procedure.

This method determines the outcome based on the decision trees' predictions. It forecasts by averaging or averaging the output of various trees. The precision of the result improves as the number of trees grows.

The disadvantages of a decision tree algorithm are avoided by using a random forest technique. It improves precision while reducing data set overfitting.

Before feeding the data to the model, lets first divide it into training and testing sets. Make sure that the data is shuffled and randomly chosen in these sets so that the model covers maximum cases and bias is avoided.



5 Model Evaluation

Evaluation metrics are a measure of how good a model performs and how well it approximates the relationship. Let us look at R Squared and RMSE.

R Squared

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

$$R^2 = 1 - \frac{UnexplainedVariation}{TotalVariation}$$

RMSE

Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors,

This is the square root of the average of the squared difference of the predicted and actual value. R-squared error is better than RMSE. This is because R-squared is a relative measure while RMSE is an absolute measure of fit (highly dependent on the variables — not a normalized measure). Basically, RMSE is just the root of the average of squared residuals. We know that residuals are a measure of how distant the points are from the regression line. Thus, RMSE measures the scatter of these residuals.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

6 Conclusion and Future work

The conclusion drawn from this project is that we have used a complete data set that has accurate information regarding the houses.Improvement in computing technology has made it possible to examine social information that cannot previously be captured, processed and analysed.

New analytical techniques of machine learning can be used in property research. This study is an exploratory attempt to use two machine learning algorithms in estimating housing prices, and then compare their results. Majorly, all of the above codes and libraries we have used are not unique as there is a specific procedure to perform the house prediction procedure by linear regression and Random Forest Regression

To conclude, the application of machine learning in property research is still at an early stage. We hope this study has moved a small step ahead in providing some methodological and empirical contributions to property appraisal, and presenting an alternative approach to the valuation of housing prices. Future direction of research may consider incorporating additional property transaction data from a larger geographical location with more features, or analysing other property types beyond housing development.

References

[1] House Price Index. Federal Housing Finance Agency. https://www.fhfa.gov/ (accessed September 1, 2019).

[2] Fan C, Cui Z, Zhong X. House Prices Prediction with Machine Learning Algorithms. Proceedings of the 2018 10th International Conference on Machine Learning and Computing - ICMLC 2018. doi:10.1145/3195106.3195133.

[3] Phan TD. Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. 2018 International Conference on Machine Learning and Data Engineering (ICMLDE) 2018. doi:10.1109/icmlde.2018.00017.

[4] hen T, Guestrin C. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16 2016. doi:10.1145/2939672.2939785.