JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ANALYSIS AND PREDICTION OF HOUSE PRICE

K.Sathvika Reddy. Stanley College of engineering and technology for women, Hyderabad, India
Tanvi Athelli. Stanley College of engineering and technology for women, Hyderabad, India
Sayeda Kulsum. Stanley College of engineering and technology for women, Hyderabad, India

Abstract The phenomenon of the falling or rising of the house prices has attracted interest from the researcher as well as many other real-estate parties. Usually, House price index (HPI) represents the summarized price changes of residential housing. Since housing price is strongly correlated to other factors, prediction needs more accurate methods based on location, house type, size, built year, local amenities, and some other factors which could affect house demand and supply. With dataset named Ames taken from Kaggle, which consists of 81 features and 1460 rows. Exploratory Data Analysis, data pre-processing, data cleaning, creative feature engineering method, One Hot Encoding is applied in this project. The project is developed using two machine learning models, Random Forest regression model and Gradient boosting regression model to predict individual house price. The performance metrics used is RMSE (Root Mean Square Error) and R-Squared value. We fit the algorithms to the data and tried three different ratios and observed which split gave better accuracy. It was observed that 80-20 split gave better accuracy when compared to 65-30 split and 70-30 split. It was also observed that Random Forest is most accurate model than The Extreme Gradient Boosting model.

1. INTRODUCTION

1.1 Introduction

Purchasing a new house is always a big decision. Moreover, economic growth plays an important role as housing demand is often seen as elastic in terms of income, leading to an increase in revenues for households. Undoubtedly it is a tough call to consider which features should be of most importance. To ease this decision-making, in today's world, Machine Learning allows an entrepreneur on forecasting the house price with a maximum accuracy of market trend and building model of the historical dataset on 'what happened and why' to predict 'What is going to happen'.

The project endeavors to extensive data analysis and implementation of different machine learning techniques in python for having the best model with most important features of a house on insight of both business value and realistic perspective. The dataset consists of 81 different features for 1460 houses in Ames which can be used as training data to predict the sale price of the machine learning model.

1.2 Problem Statement

In the existing system Lasso regression and linear regression are used to predict house prices, comparatively these algorithms are very slow in performance and doesn't give accurate results. In real world there are many real-estate classified websites where we can see a lot of inconsistencies in terms of pricing of a house. There are many cases where similar houses are priced differently and thus there is a lot of in transparency. Sometimes the consumers may feel the pricing is not justified for a particular listed house but there is no way to confirm that either. Proper and justified prices of properties can bring in a lot of transparency and trust back to the real estate industry. Our proposed project is aimed:

- To develop a house prediction model using Random Forest, Gradient boosting algorithms.
- To analyze which is a better model by finding the accuracy of the applied two algorithms.

1.3 Objective

The proposed system brings in a lot of transparency and improve the consistency among prices of the house. We propose to use machine learning techniques and algorithms to develop a model and predict which a better model for this dataset is. Random forest and XG Boost algorithms are used in developing the model for prediction of housing prices and the most accurate model is identified.

.

2. Literature Survey

Nor Hamizah Zulkifley, Shuzlina Abdul Rahman, Nor Hasbiah Ubaidullah, Ismail Ibrahim et.al.[1]Gradient boosting was created in 1999 and is a commonly used machine learning algorithm because of its performance, consistency and interpretability. Meanwhile, researcher selects XGBoost as the best model since it provides the lowest RMSE value in contrast with other models in his study. Technically, the RMSE value of a model is highly dependent on the attributes used during the prediction process. Most of the model that are using the same attributes (locational attributes) will generate a very low RMSE value indicating the best model. The RMSE value is very low with the presence of the locational attribute only, however, the RMSE value is quite high when the structural attribute is combined with the locational attribute for the input to make a prediction. In conclusion, the impact of this research was intended to help and assist other researchers in developing a real model which can easily and accurately predict house prices. Further work on a real model needs to be done with the utilization of our findings to confirm them.

ANAND G. RAWOOL, DATTATRAY V. ROGYE, SAINATH G. RANE, DR. VINAYK A. BHARADI et.al.[2] We are creating a housing cost prediction model. By using Machine learning algorithms like Random Forest Regression. The result of this research provide that the Random Forest Regression gives maximum accuracy. In this paper, the resale price prediction of house is done using different classifications algorithms like Linear regression, Decision Tree, K-Means and Random Forest is used. Here we consider RMSE as the performance metrics for different dataset and these algorithms are applied and find out most accuracy model which predict better results. They have used step wise approach from Data Collection, Pre-Processing Data, Data Analysis, to Model Building. There might be missing values in our dataset. There are three ways to fill our missing values: 1) Get rid of the missing data points.2) Get rid of the whole attribute. 3) Set the value to some value (0, mean or median).Random Forest give a highest accuracy in prediction of housing prices. The decision to choose the algorithm is depends on the dimensions and type of data is used.

Puneet Tiwari1, Varun Singh Thakur et.al.[5] The methods used in this study consisted of simple and multiple linear regression, random forests, and gradient boosting for predictors. Their research question was to determine whether the closing price was higher or lower than the listing price.

House sellers have to formulate an estimation of the worth based on its characteristics or features in similarity to the existing market price of related houses. The mixture of the characteristics or the huge

number of features makes the challenging task to calculate approximately a satisfactory market price. Another main goal of this thesis was to inspect the significance of each predictor in illumination of price variation for a specified set of housing features. Overall, the results endow with practical information regarding the cause of various features on house prices and their corresponding analysis.

3. OVERVIEW OF THESYSTEM

3.1 Existing System

• Although service providers adhere to tight agreed terms (SLAs) with remarkable minimal downtime and processing times, history shows that even the finest providers face interruptions from time to time, which can result in significant monetary losses.
• Furthermore, in order to compete in the market, cloud services always offer varied costs. Organizations could take advantage of numerous cloud servers, known as Cloud-of-Clouds, to enhancing information dependability and save costs.

3.1.1 Disadvantages of Existing System

• In reality, though, it is not easy. This ambition is hampered by a number of variables. First, distributed file technology divides huge files into little parts, resulting in decreased Gets/Puts cost and improved costs resulting in additional networked I/Os.

3.2 Proposed System

We show that utilizing data consolidation and erasure system settings, a business may save costs and survive interruptions. Second, we present an inner-chunk erasure coding method with on-demand piece reconstructing to reduce overheads in the event of a failure. Third, we create and implement a container-based share control strategies to combine tiny data portions into a bigger unit for dynamic provisioning.

Advantages of Proposed System

✓ • DC Store is a Data center file system that we designed and implemented. DC Store is designed to allow an organization to cost-effectively and reliably outsource a big group of users' data to different clouds.
✓ • When comparison to several existing Data center storage solutions, our experimental outcomes suggest that DC Store can vastly improve productivity and economic effectiveness.

3.3 Proposed System Design

In this project work, I used five modules and each module has own functions, such as:

1. Dataset
2. Exploratory Data Analysis
3. Data Cleaning
4. Train Test Split
5. Algorithm

3.3.1 Dataset

The Ames housing price dataset on Kaggle is a one of the most popular datasets on Kaggle. Ames is a small city in the state of Iowa in the United States. Its home to Iowa State University, which is the largest university in the state.

The Ames housing dataset examines features of houses sold in Ames during the 2006–10 timeframe. The goal is to use the training data to predict the sale prices of the houses in the testing data.

3.3.2 Data Analysis

Exploratory Data Analysis (EDA), also known as Data Exploration, is a step in the Data Analysis Process, where a number of techniques are used to better understand the dataset being used.

'Understanding the dataset' can refer to a number of things including

Extracting important variables and leaving behind useless variables

Identifying outliers, missing values, or human error

Understanding the relationship(s), or lack of, between variables

Ultimately, maximizing your insights of a dataset and minimizing potential error that may occur later in the process

Exploratory Data Analysis does two main things:

1. It helps clean up a dataset.

2. It gives you a better understanding of the variables and the relationships between them.

3.3.3 *Data Cleaning*

Handling missing values: The real-world data often has a lot of missing values. The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the pre-processing of the dataset as many machine learning algorithms do not support missing values.

Delete Rows with Missing Values:

Missing values can be handled by deleting the rows or columns having null values. If columns have more than half of rows as null then the entire column can be dropped. The rows which are having one or more columns values as null can also be dropped.

Pros:

- A model trained with the removal of all missing values creates a robust model.

Cons:

- Loss of a lot of information. Works poorly if the percentage of missing values is excessive in comparison to the complete dataset.

Impute missing values with Mean/Median:

Columns in the dataset which are having numeric continuous values can be replaced with the mean, median, or mode of remaining values in the column. This method can prevent the loss of data compared to the earlier method. Replacing the above two approximations (mean, median) is a statistical approach to handle the missing values.

The missing values are replaced by the mean value in the above example, in the same way, it can be replaced by the median value.

3.3.3 Train and Test Split

Training Set:

The sample of data used to fit the model that is the actual subset of the dataset that we use to train the model (estimating the weights and biases in the case of Neural Network). The model observes and learns from this data and optimize its parameters.

Cross-Validation Set:

We select the appropriate model or the degree of the polynomial (if using regression model only) by minimizing the error on the cross-validation set.

Test set:

The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset. It is only used once the model is completely trained using the training and validation sets. Therefore test set is the one used to replicate the type of situation that will be encountered once the model is deployed for real-time use.

The test set is generally what is used to evaluate different models in competitions of Kaggle or Analytics Vidhya. Generally in a Machine Learning hackathon, the cross-validation set is released along with the training set and the actual test set is only released when the competition is about to close, and it is the score of the model on the Test set that decides the winner.

Deciding the ratio of splitting the dataset:

The answer generally lies in the dataset itself. The proportions are decided according to the size and type (for time series data, splitting techniques are a bit different) of data available with us.

If the size of our dataset is between 100 to 10,00,000, then we split it in the ratio 60:20:20. That is 60% data will go to the Training Set, 20% to the Dev Set and remaining to the Test Set.

The main aim of deciding the splitting ratio is that all three sets should have the general trend of our original dataset. If our dev set has very little data, then it is possible that we'll end up selecting some model which is biased towards the trends only present in the dev set. Same is the case with training sets — too little data will bias the model towards some trends found only in that subset of the dataset.

The models that we deploy are nothing but estimators learning the statistical trends in the data. Therefore, it is important that the data that is being used to learn and that being used to validate or test the model follow as similar statistical distribution as possible. One of the ways to achieve this as perfectly as possible is to select the subsets — here the training set, the dev set and/or the test set — randomly. For example, suppose that you are working on a face detection project and face training pictures are taken from the web and the dev/test pictures are from user's cell phone, then there will be a mismatch between the properties of train set and dev/test set.

4 Architecture

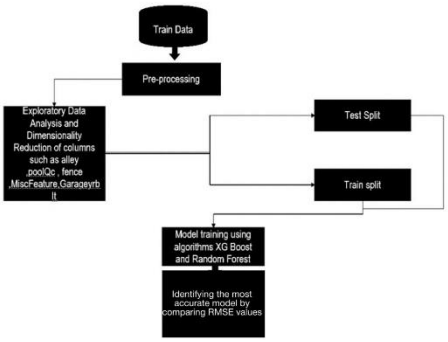


Fig 1: Frame work of DC Store

Above architecture diagram shows three stages of data flow form one module to another module. Back-end storage use data to store in cloud server proxy server manages data by handling container data share package and share cache data, Client-side uploads data and manage data in effective way which satisfies container management system.

Algorithm:

Ensembling:

It is a process by which multiple models such as classifiers or experts are strategically generated and combined to solve a particular computational intelligence problem. It is used to improve prediction, function approximation.

Bagging bootstrap aggregation:

It is a technique to reduce variance in prediction by generating additional data from datasets separately for training using combinations with repetitions

Boosting is an iterative technique which generates model with lower errors as it optimizes the advantages at the same time reduces pitfalls of the model.

5 RESULTS ANALYSIS

After observing the graphs below we can conclude that there is a deviation in XG boost algorithm that results in more RMSE value.

In random forest algorithm, for 80-20 split we find lower RMSE value when compared to other splits such 65-35 and 70-30.

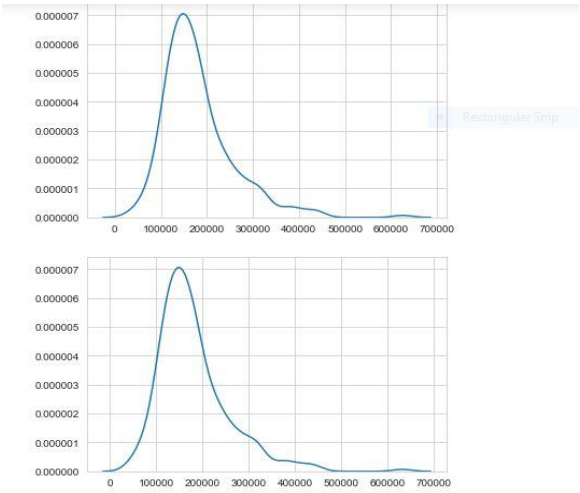


Fig 80-20 split GBDT distribution graph

RMSE Value	2505.5648294351527	6302.694367666291
R-Squared Value	0.9989668214456358	0.9934624264200033

Performance Metrics	Random Forest Algorithm	Extreme Gradient Boosting (Xgboost)Algorithm
RMSE Value	3138.502354797816	6323.908589667352
R-Squared Value	0.9983442357988913	0.9932775939976918

Table: 5.2 Performance Metrics of both Algorithms for 70-30 Split

Performance Metrics	Random Forest Algorithm	Extreme Gradient Boosting (Xgboost)Algorithm
RMSE Value	1084.5796574312487	4212.215370967135
R-Squared Value	0.9998008885371564	0.9969967292156099

Table: Performance Metrics of both Algorithms for 80-20 Split

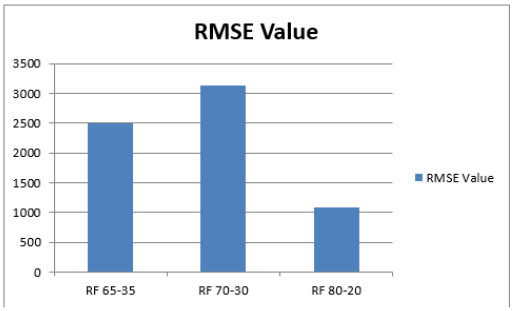
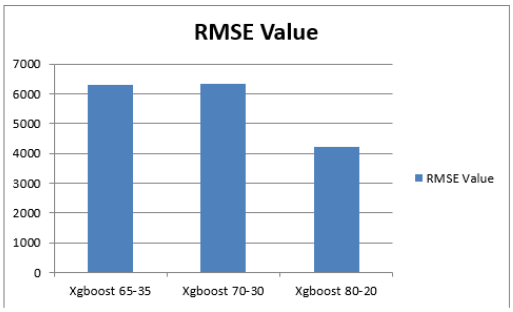


Fig 5.7 RMSE Value representation in bar graph for all three splits of RF



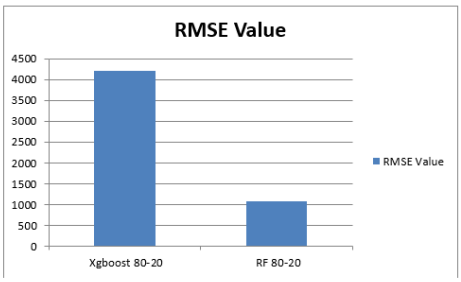


Fig 5.9 RMSE Value representation in bar graph for 80-20 split of RF and Xgboost

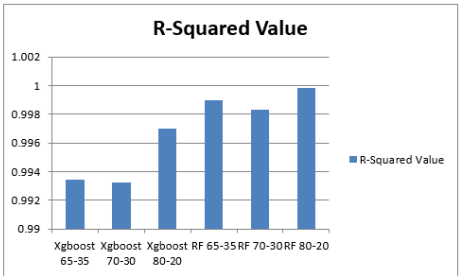


Fig 5.10 RMSE and R-Squared Value representation in bar graph for all splits of RF and

7. CONCLUSION

The goal of this statistical analysis is to help us understand the relationship between house features and how these variables are used to predict house price. We observed that this dataset has heterogeneous data.

In this project, we implemented the most fundamental machine learning algorithms like Random forest, XG boost algorithms. Work is implemented using Scikit-Learn machine learning tool. This work helps the users to predict the prices of the houses. Two algorithms which come under decision tree regression and linear regression were used in predicting the prices of the houses. The performance metrics used were RMSE (Root Mean Square Error) and R-Squared value. The model with better accuracy has greater performance and the accuracy is analyzed with RMSE value. Lesser the RMSE value, it is a more accurate model. On the other side, greater the R-Squared value, better is the model.The R-Squared value is highest in Random Forsest 80-20 split compared to rest of the splits of both algorithms. The 80-20 Train test split gave lesser RMSE Value compared to 65-35 and 70-30 split. Thus, the performance of random forest algorithm is found to be better than the XG boost algorithm in predicting the house prices.

Future Enhancement

Future work on this study could be divided into seven main areas to improve the result even further. Which can be done by:

The used pre-processing methods do help in the prediction accuracy. However, experimenting with different combinations of pre-processing methods to achieve better prediction accuracy. Make use of the available features and if they could be combined as binning features has shown that the data got improved.

Training the datasets with different regression methods such as Elastic net regression that combines both L1 and 12 norms. In order to expand the comparison and check the performance.

The correlation has shown the association in the local data. Thus, attempting to enhance the local data is required to make rich with features that vary and can provide a strong correlation relationship.

In future the dataset can be prepared with more features and advanced machine learning techniques can be used for constructing the house price prediction model.

In future more real time data can be considered in the website for predicting the house prices and GPS can be added to find housing prices in different locality.

8. References

[1] David E. Rapach , Jack K. Strauss " Forecasting real housing price growth in the Eighth District states"

[2] Vasilios Plakandaras+ and Theophilos♦, Rangan Gupta*, PeriklisGogas "Forecasting the U.S. Real House Price Index"

[3] Gupta and Das (2010) Forecasting the US Real House Price Index: Structural and Non-Structural Models with and without Fundamentals

[4] Rangan Gupta "Forecasting US real house price returns over 1831– 2013: evidence from copula models"

[5] R. Gupta, A. Kabundi and S. M. Miller, "Forecasting the US real house price index: Structural and non-structural models with and without fundamentals", *Economic Modelling*, vol. 28, no. 4, pp. 2013-2021, 2011.

[6] R. Gupta, A. Kabundi and S. M. Miller, "Forecasting the US real house price index: Structural and non-structural models with and without

fundamentals", *Economic Modelling*, vol. 28, no. 4, pp. 2013-2021, 2011.

[7]J. Mu, F. Wu and A. Zhang, "Housing Value Forecasting Based on Machine Learning Methods", *Abstract and Applied Analysis*, pp. 7, 2014.

[8] L. Bork and S. Moller, "Forecasting house prices in the 50 states using Dynamic Model Averaging and Dynamic Model Selection", *International Journal of Forecasting*, vol. 31, no. 1, pp. 63-78, 2015.

[9] M. Balcilar, R. Gupta and S. M. Miller, The out-of-sample forecasting performance of nonlinear models of regional housing prices in the US. Applied Economics, vol. 47, no. 22, pp. 2259-2277, 2015.

[10] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County", *Virginia housing data. Expert Systems with Applications*, vol. 42, no. 6, pp. 2928-2934, 2015.

[11] V. Plakandaras, R. Gupta, P. Gogas and T. Papadimitriou, "Forecasting the US real house price index", *Economic Modelling*, vol. 45, pp. 259-267, 2015.

[12] A. Ng and M. Deisenroth, "Machine learning for a London housing price prediction mobile application" in Technical Report, London, UK:Imperial College, June 2015