House Rent Prediction Using Linear Regression and Random Forest

Samhitha Alapati[1], Gopinath Perumalla[2]
Dept. of Data Science, University of Memphis, USA

ABSTRACT

In many real-world applications, predicting a price variance rather than a fixed figure is more plausible and appealing. In this context, price prediction can be seen as a classification problem. The House Price Index (HPI), however, is a popular method for determining the irregularities of housing prices. Estimating individual house prices requires information other than HPI because housing prices are tightly connected with other characteristics including location, city, and population.

The HPI, which measures average price changes in repeat sales or refinancing of the same assets, is a repeat sale index. Because HPI is a rough predictor based on all transactions, it is unsuccessful at forecasting the price of a particular house. The use of the Random Forest, Linear Regression machine learning technique for predicting housing prices is explored in this study.

Using a machine learning repository, Ames Housing Dataset from Kaggle were employed to assess the effectiveness of the suggested prediction model.

INTRODUCTION

There has been a significant body of research on text mining for information retrieval during the past 35 years (IR). There are several obstacles involved in using automated text mining algorithms to extract knowledge from natural language documents, but they also present special opportunities. Texts written in natural language are one of the most organic ways to store data. Although a human can understand this with ease, computers still find it quite difficult to evaluate this data. In contrast to human abilities, computers do have one significant advantage: computing capacity. This indicates that computer systems are more efficient and precise than humans at identifying patterns, which are non-trivial recurrences, inside data, but this is only possible if the structure of the data is known. Although hidden grammatical structures exist in natural language, they are extremely complicated and differ greatly between languages.

As the demand increases the house prices also increases. It is also depending on various factors. These factors, according to authors in [1], are place, concept, and physical condition like size of the house, number of rooms, kitchen area, garage area etc., When the values get changed, it is difficult to predict the price of the house as well. Sometimes it also difficult to calculate the house prices accurately because the agents get biased if they are interested in the property. Physical characters such as year built, interior features etc. also show impact on the house according to the author in [3].

To solve the problem, we are employing the most popular machine learning algorithms called linear regression and Random Forest regression algorithms. To work with these algorithms, we are using the pre-existing real dataset 'Ames Housing Dataset' from Kaggle.

AIM AND OBJECTIVE

The objective is to forecast the effective home pricing for real estate clients considering their priorities and finances. By evaluating recent market trends and price ranges, and forthcoming developments future prices will be projected.

The current real estate market is frenetic and pricey. Since the client must travel and pay the real estate agent's commission. Additionally, the customer/buyer is unsure of the property's future profitability. Due to conflicts of interest with buyers, sellers, or mortgages, the evaluators may occasionally be biased.

EXISTING SYSTEM

The current system has some flaws and is not idiot proof. The current system has several potential flaws and limitations because it is manual. Among them are:

Human Resources: The laborious process involved in the existing system—from filling out forms to completing paperwork to delivering manifestos—is too onerous. This adds to the workload for employees but does not produce the desired outcomes.

Thorny Work - If any changes are made to the current system, there will be an increase in manual labor and risk of error.

Error: Since the system is run and operated by people, mistakes are a possibility.

PROPOSED SYSTEM

E-learning and e-education are heavily affected today. Automation is replacing manual systems everywhere. This project's goal is to forecast house prices to lessen the difficulties the consumer would experience. The customer currently approaches a real estate agent to handle his or her investments and recommend appropriate estates for his investments. However, this approach carries some risk because the agent could make a mistaken estate prediction and lose the customers' capital as a result. The manual approach that is still prevalent in the market is risky and out-of-date. There is a need for an updated, automated system to address this flaw. Data mining algorithms can be used to guide investors toward making a suitable real estate investment based on their stated needs. The new system will also save money and time. Its operations will be straightforward. The suggested system utilizes RANDFOREST and the LINEAR REGRESSION classification technique.

The new system will also save money and time. Its operations will be straightforward. The classification algorithm used by the suggested system is Nave Bayes. According to the information entered by the administrator, the system will estimate the cost of the hotel.

RANDOM FOREST

Random Forest The products of the type are the choice tree's choices, while random forests are hierarchical categories produced by a collection of choice trees. The "Beijing" and "Beiman" ideas of random option selection are combined in random forests. Leo Bremen and Adele Bargainer developed the rule to support Random Forest.

LINEAR REGRESSION

A supervised machine learning model called "linear regression" aims to simulate a linear connection between dependent variables (Y) and independent variables (X). Every observation that is subjected to model evaluation has its target (Yactual)'s value and anticipated value compared; the largest discrepancies between these values are referred to as residuals. The total squared residuals are what the linear regression model seeks to reduce.

PROCEDURE

Data Preprocessing:

For the models to learn the patterns more rapidly, the data that will be used for model training and testing should be carefully examined before being turned into models. While categorical data were encoded one at a time, numerical values were standardized. The next step is pre-processing the data of the chosen features that will be used after exploring the data and choosing the best feature with the help of a heatmap. The datasets obtained for the training and testing task typically have several attributes. Scaling was done to make sure that the features are on a scale that is roughly similar because it is highly likely that the values of various characteristics are on different scales, which may reduce the performance of the model. This task was handled by the Standard Scaler function found in the Phyton Skitlearn module. The Standard Scaler adjusts your data so that it is now clustered around 0 and has a standard deviation of 1, presuming that your data is naturally distributed within each function. After measuring the feature's mean and standard deviation, the feature is scaled using:

Xi – mean(x) / (stdev(x))

We collect data from Kaggle which is 'Ames Housing Dataset'. Once we collect the dataset we read and load the data. After loading the dataset, we look for the dummy data if any possible. We drop the dummy data from the dataset.

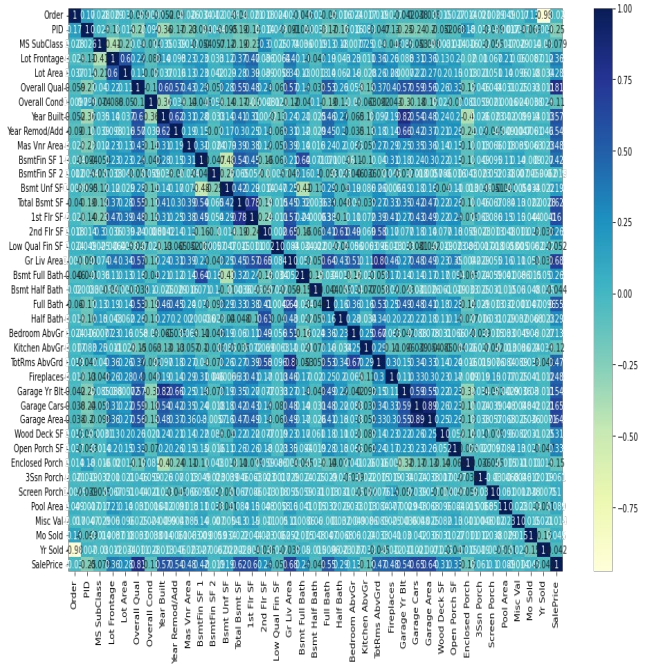
Splitting data:

Using sklearn train test split, we will now divide our dataset into a training set and a testing set (). The training set will be used to train the model, and the testing set to evaluate it. The training model's data are in the files X train and y train. The testing model's data are in the files X test and y test. Names for the characteristics and target

variables are X and y.

Find Correlation Coefficient:

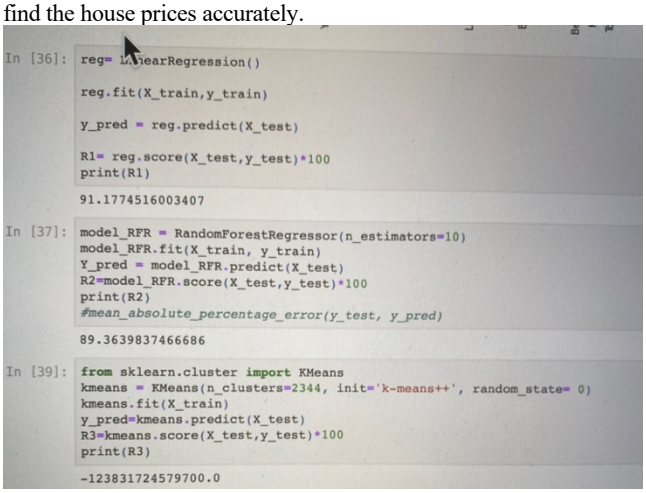
To determine which variables are strongly associated, let's look at the correlation coefficients. A heatmap is a type of graph where the values of the data are represented by colors. In other words, it uses color to tell the reader something. When there is a lot of data, using a heatmap is an excellent approach to direct the reader to the key areas. Seaborn heatmaps have an appealing visual appearance and seem to convey simple data messages instantaneously. Therefore, this method of correlation matrix visualization is used by both data scientists and data analysts. The heatmap generated is shown in the below figure1.



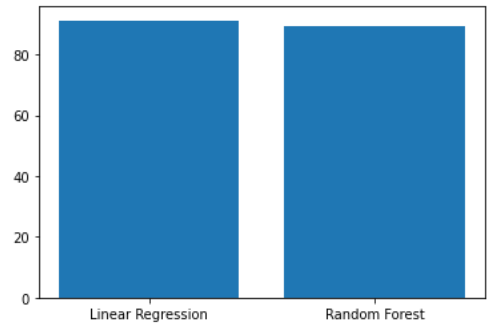
We find mean absolute errors using both the algorithms, Linear Regression and Random Forest and K-means. We will import the training dataset, build a sklearn linearmodel LinearRegression object, RandomForestRegressor, KMeans and fit it to it.

RESULTS

We find the house price predictions using the three machine learning algorithms. Here, we can clearly say which algorithm is best suitable to find the house prices accurately.



```
In [36]: reg= LinearRegression()

         reg.fit(X_train,y_train)

         y_pred = reg.predict(X_test)

         R1= reg.score(X_test,y_test)*100
         print(R1)

         91.1774516003407

In [37]: model_RFR = RandomForestRegressor(n_estimators=10)
         model_RFR.fit(X_train, y_train)
         Y_pred = model_RFR.predict(X_test)
         R2=model_RFR.score(X_test,y_test)*100
         print(R2)
         #mean_absolute_percentage_error(y_test, y_pred)

         89.3639837466686

In [39]: from sklearn.cluster import KMeans
         kmeans = KMeans(n_clusters=2344, init='k-means++', random_state= 0)
         kmeans.fit(X_train)
         y_pred=kmeans.predict(X_test)
         R3=kmeans.score(X_test,y_test)*100
         print(R3)

         -123831724579700.0
```



CONCLUSION

It is now difficult to store such vast amounts of data and then retrieve them for one's own needs in the real estate industry. The data that was extracted should also be helpful. The system utilizes the data mining algorithm to its full potential. The system makes the best use of such data possible. By improving estate selection accuracy and lowering the danger of estate investment, the data mining algorithm aids in client satisfaction. There are several features that might be added to the system to increase acceptance. One of the main future goals is to expand the estate database to include new cities, which will enable users to investigate more estates and make informed decisions.

There are other additional models that can be used for prediction. Data provided as input to such a model should be compatible with the process's operators and the tool being utilized. Additionally, extra data sets can be employed to improve the model's accuracy. The major goal of using a different model should be to simplify the procedure and shorten the calculation time. From the above results and the graph, Linear Regression has 91% accuracy rate, Random Forest has 89% and K-Means has negative value. SO, we can clearly say that Linear Regression performs well to predict the house prices.

FUTURE SCOPE

Numerous things could be improved or added in the upcoming development. However, most of the residential areas could be located. There may be a few more locations with apartment buildings or housing complexes with multiple floors that are situated in business areas. Such flats could be counted in the future to provide a more accurate conclusion because they weren't included in this article. There is a noticeable rise in the number of private builders who add extra facilities to real estate to entice more buyers as the demand for housing in metropolitan areas grows. There are other additional models that can be used for prediction. Data provided as input to such a model should be compatible with the process's operators and the tool being utilized. Additionally, extra data sets can be employed to improve the model's accuracy. The major goal of using a different model should be to simplify the procedure and shorten the calculation time.

REFERENCES

[1] Malang, C. S., Java, E., & Febrita, R. E. (2017). Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization. International Journal of Advanced Computer Science and Applications, 8(10), 323–326.

[2] Chogle, A., khaire, priyanka, gaud, A., & Jain, J. (2017). House Price Forecasting using Data Mining Techniques. *International Journal of Advanced Research in Computer and Communication Engineering*, *6*(12), 82–83. https://doi.org/10.17148/IJARCCE.2017.61216

[3] Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2020). Land Use Policy Understanding house price appreciation using multi-source big geo-data and machine learning. Land Use Policy, July, 104919. https://doi.org/10.1016/j.landusepol.2020.104919

[4] Byeonghwa Park , Jae Kwon Bae (2015). Using machine learning algorithms for housing price prediction, Volume 42, Pages 2928-2934

[5] Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House Price Prediction using Random Forest Machine Learning Technique. *Procedia Computer Science*, *199*, 806–813. https://doi.org/10.1016/j.procs.2022.01.100

[6] Burse, S., Anjaria, D., & Balaji, H. (2021). Housing Price Prediction Using Linear Regression. *Journal of Emerging Technologies and Innovative Research (JETIR)*, *8*(10), d11. https://www.jetir.org/papers/JETIR2110302.pdf