**Acknowledgement**

|  |  |
| --- | --- |
| **pRojECT report on**  **REAl ESTATE**  **pRiCE**  **pREDiCTion** | Big data analytics  **Names**:  Tripathi Shivam Raghavesh(1800268C203)  Jai Kumar (1800223C203)  **Year**: 3rd (6th Sem) **Programme**: SOET  Image result for bml munjal university logo  **Submitted to:** Dr. Yogesh Gupta |

I want to express my utmost gratitude to my senior Shubhi Jain 4th year Computer science student at BML Munjal university for her everlasting support and encouragement towards the project which helped me a lot in completing this project.

I thank my University Professor Dr. Yogesh Gupta, Professor BML Munjal University for his constant encouragement during my project research.

Thanking You

Jai Kumar & Tripathi Shivam Raghavesh

**INDEX**

|  |  |  |
| --- | --- | --- |
| SR NO | TOPIC | PG NO |
| 1 | ABSTRACT | 4 |
| 2 | MOTIVATION | 4 |
| 3 | INTRODUCTION | 5 |
| 4 | PROBLEM STATEMENTS | 5 |
| 5 | LITERATURE SURVEY | 5 |
| 6 | METHODOLOGY | 6 |
| 7 | RESULTS & DISCUSSION | 8 |
| 8 | CONCLUSION & FUTURE WORK | 10 |
| 9 | REFERENCES | 11 |

**Abstract**

In today's world, everybody wants a home that suits their lifestyle and offers the facilities they need. House prices vary significantly a significant amount, indicating that they are often inflated. Many aspects must be considered when estimating house prices, including location, number of rooms, carpet area, age of the home, and other simple local amenities. The relationship between house prices and the economy is a primary factor for house price predictions. In real estate sales, the valuation of a house is important. House price patterns are not only a source of concern for buyers and sellers, but they also reveal the state of the economy. As a result, it is important to forecast home values without biases in order to assist both buyers and sellers in making decisions. In this project, we will build a website where users can enter property information to forecast house prices, enter a date to forecast prices before that date, and enter a budget range to suggest the best spot. This project makes use of two datasets: one contains several attributes and significant entries from Bangalore home transactions, and the other contains the Bangalore house price index. To predict the current house price, we use a combination of feature collection and feature extraction methods combined with Multiple Linear Regression. The dataset is used to forecast house prices using a machine learning algorithm. These are the base models, and with the aid of specialized data mining techniques, algorithms such as a random forest, gradient boosted forests, multi-layer perceptron, and ensemble classification algorithms are used to improve prediction accuracy.

**Keywords:** Random Forest, Multiple Regression, Support Vector Machine, Gradient boosted trees, Multi-layer perceptron, bagging, price prediction, Machine learning, Advanced data mining

**Motivation**

Growing unaffordability of housing has become one of the major challenges for metropolitan cities around the world. In order to gain a better understanding of the commercialized housing market we are currently facing; we want to figure out what are the top influential factors of the housing price. Apart from the more obvious driving forces such as the inflation and the scarcity of land, there are also several variables that are worth looking into. Therefore, we choose to study the house prices predicting problem on Kaggle, which enables us to dig into the variables in depth and to provide a model that could more accurately estimate home prices. In this way, people could make better decisions when it comes to home investment

# 1.INTRODUCTION

The relationship between house prices and the economy is a major motivation for house price forecasting. There is no accurate house calculation. In real estate sales, the valuation of a house is crucial. House price patterns are not only a source of anxiety for buyers and sellers, but they also reveal the state of the economy. As a result, it is important to forecast house prices without biases in order to assist both buyers and sellers in making decisions. House prices are influenced by many factors of variables, including the number of bedrooms and bathrooms. Furthermore, various variations of parameters in Support Vector Regression can have a significant impact on the predictions. Many approaches have been used in price prediction, such as hedonic regression, and in this paper, I am attempting to forecast the future real estate price using machine learning techniques and previous works. To estimate the house price, I used random forest, multiple regression, and other algorithms with various methods. As a result, it would be beneficial for people to be mindful of all current and future circumstances in order to prevent making mistakes.

**1.1 Goals of the Study**

The main objectives of this study are as follows:

* To apply data preprocessing and preparation techniques in order to obtain clean data
* To build machine learning models able to predict house price based on house features
* To analyze and compare model’s performance in order to choose the best model

# 2. PROBLEM STATEMENTS

Prices of real estate properties are sophisticatedly linked with our economy. Despite this, we do not have accurate measures of housing prices based on the vast amount of data available. Therefore, the goal of this project is to use machine learning to predict the selling prices of houses based on many economic factors.

# 3. LITERATURE REVIEW

Housing market trends represent the current economic situation and are a source of concern for both buyers and sellers. House prices are influenced by a variety of variables, including the number of bedrooms and bathrooms. The cost of a house is often determined by its venue. A house with excellent proximity to highways, classrooms, shops, and job opportunities would command a higher price than one without such access. Manually predicting house prices is challenging and infrequently reliable, which is why many systems for house price prediction have been created. Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh [1] had proposed an advanced house prediction system using linear regression. The aim of this method was to create a model that could forecast house prices based on a variety of factors. For the house predection dataset, they used Linear Regression, which provided high precision. The Admin and User modules of the house price prediction project were separate. Locations may be added and viewed by the administrator. Admin has the power to increase density on a per-unit-area basis. Users will look at the place and see what the home price is expected to be in that region.

The Hybrid Regression Approach for House Prices was proposed in this project. The emphasis of the prediction was on the use of innovative feature engineering to identify the best features and their relationship to sales prices. Data normality and linearity were increased thanks to feature engineering. Their method demonstrated that working with the house prediction dataset was easy, and that using Hybrid algorithms (65% Lasso and 35% Gradient Boost) produced better results in forecasting house prices than using lasso, ridge, or gradient boost alone.

## EXISTING STATE OF ART

|  |  |  |  |
| --- | --- | --- | --- |
| S. No. | Existing state of art | Drawbacks in existing state of art | Overcome |
| 1 | Prediction based on number of washrooms | Couldn’t predict on basis of washroom detail | Outlier Removal Using Bathrooms Feature and Balcony Feature |
| 2 | Prediction based on number of balcony | Couldn’t predict on basis of number of balcony present |

# 4. METHODOLOGY

# 4.1. Variable Selection

Since we believe that most variables have a linear relation with price, we fitted a linear regression model prediction of house price with cross validation to find the best shrinkage parameter lambda and variables remaining important with that parameter. We found out that carpet area, no. of room, No. of washroom were shown to be important in model.

**Phase 1: Collection of data**

There are various data processing methods and procedures to choose from. We gathered information about real estate properties in Bangalore from a variety of real estate websites. Place, carpet area, built-up area, property age, zip code, and other attributes will be included in the results. We need to collect structured and categorised quantitative data. Before conducting any machine learning study, it is essential to collect data. Validity of the dataset is essential; otherwise, there is no point in analysing the results.

**Phase 2: Data pre-processing**

Data pre-processing is the process of cleaning our data set. There might be missing values or outliers in the dataset. These can be handled by data cleaning. If there are many missing values in a variable we will drop those values or substitute it with the average value.

**Phase 3: Training the model**

Since the data is broken down into two modules: a Training set and Test set, we must initially train the model. The training set includes the target variable. The decision tree regressor algorithm is applied to the training data set. The Decision tree builds a regression model in the form of a tree structure.

**Phase 4: Testing and Integrating with UI**

The trained model is applied to test dataset and house prices are predicted. The trained model is then integrated with the front-end using Flask in python

Diagram

Description automatically generated

## 4.2 Best Predictive Model Selection

We trained Linear Regression, LASSO, Random Forest, and Decision tree models to evaluate their performance on the validation data. There are various metrics to measure the performance of a regression model, but we used the R-squares and RMSE (Root-mean-square-error) metrics to find the best model.

**Linear Regression**

Linear regression is one of the most well known and the simplest way to predict the outputs, which fits a linear model to minimize the residual sum of squares between the predicted values and the true values. Though the main disadvantage of linear regression is that it assumes the linearity between the predicted and the response variables, but data are rarely linearly separable in the real world.

**LASSO Regression**

LASSO regression is the special type of linear model that adds the constraint to prevent having unnecessary variables in the model. This regularization reduces overfitting and helps in feature selection. LASSO is also well known for its interpretability of the model. We will discuss this later in the article.

**DECISION TREE REGRESSOR**:

The decision tree regressor observes features of an attribute and trains a model in the form of a tree to predict data in the future to produce meaningful output. Decision tree regressor learns from the max depth, min depth of a graph and according to system analyzes the data. Grid Search CV is a way to deal with parameter tuning that will efficiently manufacture and assess a model for every mix of calculation parameters indicated in a grid. Grid Search CV in this algorithm is used to assess the best value for max-depth, using which the decision tree is constructed

**FLASK INTEGRATION**

After building the model and successfully giving the result, the next step is to do the integration with the UI, for this purpose flask is used. Flask is a web framework. This means flask provides you with tools, libraries, and technologies that allow you to build a web application. Flask is easy to put away routes together and this framework is mainly used for integrating python models.

# 5. RESULTS & DISCUSSION

**5.1 Outlier Removal Using Business Logic**

As a data scientist when you have a conversation with your business manager (who has expertise in real estate), he will tell you that normally square ft per bedroom is 300 (i.e. 2 bhk apartment is minimum 600 sqft. If you have for example 400 sqft apartment with 2 bhk than that seems suspicious and can be removed as an outlier. We will remove such outliers by keeping our minimum threshold per bhk to be 300 sqft

A screenshot of a computer

Description automatically generated with medium confidence

**5.2 Outlier Removal Using Standard Deviation and Mean**

Here we find that min price per sqft is 267 rs/sqft whereas max is 12000000, this shows a wide variation in property prices. We should remove outliers per location using mean and one standard deviation

**Based on above charts we can see that data points highlighted in red below are outliers and they are being removed due to remove\_bhk\_outliers function**

### Before and after outlier removal: Rajaji Nagar

Scatter chart

Description automatically generated

**5.3 Use K Fold cross validation to measure accuracy of our LinearRegression model**

**We can see that in 5 iterations we get a score above 85% all the time. This is pretty good but we want to test few other algorithms for regression to see if we can get even better score. We will use GridSearchCV for this purpose.**

**Graphical user interface, text, application, chat or text message

Description automatically generated**

**Based on above results we can say that LinearRegression gives the best score. Hence we will use that.**

**Flask Model**

Graphical user interface, text, application, chat or text message

Description automatically generated

# 6. CONCLUSION & FUTURE WORK

This project is really a great learning experience, which lets us go through the whole entire process of building a machine learning model to solve a practical regression problem in real world, starting from the very beginning of data analysis, cleaning, preparation, etc. till the model stacking, evaluation, and delivery in the very end. We touched upon, thought over, and finally either solved or gained useful knowledge/understandings of many sorts of major or minor practical issues, which is really a very rewarding process overall. In the end, we all agreed and realized that: feature engineering is more often than not one of the most critical part or differentiator on the final model performance.

Due to limited time, we realized but have no time to try many of the interesting/promising directions/ideas, which are summarized as follows for a good reference in the future.

* No need to convert categorical variables to dummy variables for tree-based models.
* Try using different feature selection for different models: i.e. only dropping features for linear models, but not for tree-based non-linear models
* Outlier check and removal
* Clustering analysis to generate new useful categorical features

**6.1 FUTURE WORK**

In the future, we are going to present a comparative study of the systems’ predicted price and the price from real estate websites such as Housing.com for the same user input. Also, to simplify it for the user, we are going to recommend real estate properties to the user based on the predicted price. The current dataset only includes cities of Bangalore, expanding it to other cities and states of India is the future goal. To make the system even more informative and user-friendly, we will be including Gmap. This will show the neighborhood amenities such as hospitals, schools surrounding a region of 1 km from the given location. This can also be included in making predictions since the presence of such factors increases the valuation of real estate property

**7. References**

1. Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh, “A Hybrid Regression Technique for House Price Prediction”, December 2017
2. M. Risdal, “Predicting House Prices Playground Competition: Winning Kernels”, 2017. [Online]. Available: http://blog.kaggle.com/2017/03/29/predicting-house-prices-playgroundcompetition-winning-kernels/ [Accessed: 25-Mar-2019].
3. Byeonghwa Park , Jae Kwon Bae (2015). Using machine learning algorithms for housing price prediction , Volume 42, Pages 2928-2934.
4. Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, 2015. Introduction to Linear Regression Analysis.