

### PROJECT BASED LAB REPORT

**On**

**HOUSE PRICE PREDICTION**

**Submitted in partial fulfilment of the Requirements for the award of the Degree of Bachelor of Technology**

**In**

**Computer science and Engineering Under the esteemed guidance of**

P. Ithaya Rani

**By**

|  |  |
| --- | --- |
| **Student Name** | **ID Number** |
| P. Imran Ali Khan | 190031274 |

**(DST-FIST Sponsored Department) K L EDUCATION FOUNDATION**

**Green Fields, Vaddeswaram, Guntur District-522 502 2021-2022**

**K L EDUCATION FOUNDATION DEPARTMENT OF COMPUTER SCIENCE ANDENGINEERING**

**(DST-FIST Sponsored Department)**



**CERTIFICATE**

This is to certify that this project-based lab report entitled **“**House Price Prediction**”** is a bonafide work done by P. Imran Ali Khan (190031274) in the course **Big Data Analytics, 19CS3275P** in partial fulfilment of the requirements for the award of Degree in Bachelor of Technology in **COMPUTER SCIENCE AND ENGNEERING** during the Odd Semester of Academic year 2021-2022.

### Faculty in Charge Head of the Department

P. Ithaya Rani Mr. HARI KIRAN VEGE

**K L EDUCATION FOUNDATION**

**DEPT OF COMPUTER SCINCE AND ENGINEERING**

**(DST-FIST Sponsored Department)**



**DECLARATION**

I here by declare that this project based lab report entitled **“**House Price Prediction**”** has been prepared by me in the course **Big data Analytics**, **19CS3275P** in partial fulfilment of the requirement for the award of degree bachelor of technology in **COMPUTER SCIENCE AND ENGINEERING** during the Odd Semester of the academic year 2021-2022.I also declare that this project-based lab report is of my own effort and it has not been submitted to any other university for the award of any degree.

### Date: 04-05-2022

### Place: KLU

**Signature of the student**

|  |  |
| --- | --- |
| **Student Name** | **ID Number** |
| P. Imran Ali Khan | 190031274 |

**ACKNOWLEDGEMENT**

Our sincere thanks to Dr. K. Vijaya Shree in the Lab for her outstanding support throughout the project for the successful completion of the work.

I express my gratitude to Dr. K. Vijaya Shree the instructor and **Dr. K. Vijaya Sree** Course Co-Ordinator for the course **Big Data Analytics**, **19CS3275P** in the Computer Science and Engineering Department for providing me with adequate planning and support and means by which I can complete this project-based Lab.

I express my gratitude to **Mr. V. HARIKIRAN,** Head of the Department for computer science and Engineering for providing me with adequate facilities, ways and means by which I can complete this project-based Lab.

I would like to place on record the deep sense of gratitude to the honorable Vice Chancellor, K L University for providing the necessary facilities to carry the project-based Lab.

Last but not the least, I thank all Teaching and Non-Teaching Staff of our department and especially our classmates and our friends for their support in the completion of my project-based Lab.

**Name of the student**

|  |  |
| --- | --- |
| **Student Name** | **ID Number** |
| P. Imran Ali Khan | 190031274 |

**TABLE OF CONTENTS**

**CHAPTERS PAGENO**

1: Report and Introduction 6

2: Literature Survey 7

3: Model Building 8

4: Conclusion 16

5: Future Scope 16

**HOUSE PRICE PREDICTION**

**REPORT:**

Now a days there is large increase of Technology, and the lot of job opportunities has been appearing. We know that there is no specific rule to get placed in a specific area. As per the manager requests, he/she must want to migrate to the place where they got a job. Either the place is very new for them, but they must be migrated for the job.

This is not only the case for the job holders, it’s just an example, there are many people who migrates from one place to other place for their livelihood. As when they migrated to the new place, they must want to get clarity of necessary things to be known prior to they move from one place to another place. In that some mentions like: place of living, food availability, neighborhood and etc. In that one of the most important things is place of living. The people who migrate from one place to other place they must choose a particular place to live, and I had chosen this a problem and got it with a solution. I had created a machine learning model to predict the price of the house at a particular place in the banglore city and based on the prices they a have the money in their hand and migrates to that place.

The main aim of the project is to provide a price of the houses in the banglore city. For this purpose, I had used linear regression, lasso regression and ridge regression and built a web application which make the user to feel comfortable to know the particular price of the house.

**INTRODUCTION:**

I had named my House Price Prediction as Housing System. As we see some of our relative, some of our friends migrate to the new place as per their need. And there they must need a place to live. As we know every one has some dreams, in that most of the people who don’t have their own house, they must dreams for the new house. For these purpose Housing System came into the place to help these kinds of peoples who would like to buy a new house at a particular place in banglore. Simply, Housing System is the process of providing data with prices of the house in the banglore city. In this process we will take the inputs from the user with the necessary fields like location, number of bathrooms and number of bedrooms and square feet of the land. With these mentioned data we will provide the price of the house.

Input for our project is the raw data that was extracted from the Kaggle which has nearly 14k of data and it was cleaned for the best purpose of the model prediction. At last, after cleaning the data was reduced to nearly 7K.

House price Prediction comes under Regression technique, this is why because, the data which I had collected if of continuous type. For the continuous type of data, we use regression techniques like Linear Regression, Lasso Regression and Ridge Regression.

In this project I had used flask framework to build a web application for better interaction for getting inputs from the user.

**LITERATURE SURVEY:**

P. DURGANJALI AND M. VANI PUJITHA [1] had collected the data from Kaggle which is nearly 217000 data, which is of categorized data and with that categorized data they had applied different classifications techniques to get the house resale price.

The classification techniques they had used are Logistic Regression, Decision Tree, Naïve bayes, Random Forest and Adaboost algorithms. In that five algorithms adaboost algorithms had got the best prediction accuracy which is 96%.

CH. RAGA MADHURI, ANURADHA AND M. VAANI PUJITHA [2]. In their article, they mainly concentrates on the comparison between different machine learning algorithms (Multiple Linear Regression, Ridge Regression, LASSO Regression, Elastic Net Regression, Ada Boosting Regression, gradient boosting) about House price prediction Analysis. From the above experiment results, gradient boosting algorithm has high accuracy value when compared to all the other algorithms regarding house price predictions. Here the [MSE] Mean Square Error and [RMSE] Root Mean Square Error are used in order to calculate the accuracy value of the algorithm on the King County Dataset which was collected from public dataset. The paper can be extended by applying the above said algorithms to predict House resale value.

And the highest accuracy is for gradient boosting algorithm which was 91% accuracy.

AYUSH VARMA, ABHIJITH SARMA, SAGAR DOSHI AND ROHINI NAIR [3]. In their article they had created a system that aims to provide an accurate prediction of housing prices has been developed. The system makes optimal use of Linear Regression, Forest regression, Boosted regression. The efficiency of the algorithm has been further increased with use of Neural networks. The system will satisfy customers by providing accurate output and preventing the risk of investing in the wrong house. Additional features for the customer’s benefit can also be added to the system without disturbing it’s core functionality. A major future update could be the addition of larger cities to the database, which will allow our users to explore more houses, get more accuracy and thus come to a proper decision.

MANASA J AND RADHA GUPTA [4]. They have taken the data from the Kaggle and applied the regression methods like linear regression, lasso regression, ridge regression and gradient boosting regression algorithm. As a result, they had got very less accuracy score, they got highest 79% for the lasso regression. And finally, they had mentioned the reason for getting very less scores. They mentioned that the score is very less due to the less amount of training dataset and if the training dataset contains more amount of data, then it can be able to predict the data efficiently and correctly with some good score.

**MODEL BUILDING:**

In this project I am going to predict the house price in the Bangalore, for this purpose I had collected the data from the Kaggle ([www.kaggle.com](http://www.kaggle.com)). As the data I had collected was continuous, so for that case we can only be able to use regression analysis to predict the data.

Regression Analysis:

Regression analysis is a predictive modelling technique that analyzes the relation between the target or dependent variable and independent variable in a dataset. It is mainly used to perform operations on the continuous data.

The different types of regression analysis techniques get used when the target and independent variables show a linear or non-linear relationship between each other, and the target variable contains continuous values. The regression technique gets used mainly to determine the predictor strength, forecast trend, time series, and in case of cause & effect relation.

In this project, I am mainly using 3 types of regression analysis techniques to find the best fit prediction data.

1. Linear Regression
2. Lasso Regression
3. Ridge Regression

Linear Regression:

Linear Regression is the regression technique which is used to predict the output of the continuous dependent variable with the help of independent variables.

Simply we can say, Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables.

Chart, scatter chart

Description automatically generated

Formula that linear regression uses to predict the data is y = mx + c

Here, y = dependent variable which needs to be predicted

m = slope of the linear line

c = y-intercept

x = independent variable value.

Lasso Regression:

The word “LASSO” stands for Least Absolute Shrinkage and Selection Operator. It is a statistical formula for the regularization of data models and feature selection. Regularization is an important concept that is used to avoid overfitting of the data, especially when the trained and test data are much varying.

Lasso regression is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

A picture containing watch, clock, gauge

Description automatically generated

Here, lambda denotes the amount of shrinkage.

Ridge Regression:

Ridge regression is a model tuning method that is used to analyses any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

The term multicollinearity also refers to collinearity concept in statistics. In this phenomenon, one predicted value in multiple regression models is linearly predicted with others to attain a certain level of accuracy.

Text, letter

Description automatically generated

**Data pre-procession:**

I had collected nearly 14,000 data from the Kaggle.

As the first step I got the knowledge of the data like what data was present in the data and what are the data types of the data and what are the mean, variance and standard deviation of data.

The data that I had collected contains different types of data types, and they are:

|  |  |  |
| --- | --- | --- |
| **Id** | **Column** | **Data type** |
| 1 | area\_type | Object |
| 2 | Availability | Object |
| 3 | Location | Object |
| 4 | Size | Object |
| 5 | Society | Object |
| 6 | Total\_sqft | Object |
| 7 | Bath | Float |
| 8 | Balcony | Float |
| 9 | Price | Float |

Data Cleaning:

As I had found in my data, the missing values are more which reflects in incorrect prediction of output or which reflect to bad fit of the data to particular model. So, for that purpose I had check which ever data was mostly missing in the data. As a result, I had found location, size, society, balcony and bath data was missing in my data.

|  |  |
| --- | --- |
| **Column** | **No. of missing Data** |
| Location | 1 |
| Size | 16 |
| Society | 5502 |
| Balcony | 669 |
| Bath | 73 |

As in the above table the Society was missing 5502 of total data which is nearly 40% of the society data was missing. So, as the data was missing is in huge amount, so I had deleted the Society data from the total data.

In the above the balcony was missing 669 of total data, as the data was missing is very high so, I had simply deleted the data. And I found the area\_type data and availability data was no need for our model, as these two columns of data has very less correlation with all other data columns. So, for this reason I had deleted the area\_type and availability data from the data frame.

As you can see the above missing data, there is only one missing data for the location columns, for that missing data I had replaced with the most appearing value of the location column.

And in the bath column 73 rows of data was missing, for that data I had replaced with median of the bath column data as it is float data type value.

And at last the missed column was size, as it contains the data of object data type. The data contains data like 2 BHK, 3 BHK, 4 BHK. It contains data with a number followed with some data. So I had extracted the number data and fixed it to the bath column, by that the bath column contains data of number only which is float data type.

I had observed the data in the total\_sqft contains data like 1223-3122 and some data contains directly number only. And to extract the data which are in the form of 1223-3122 I had split the data by ‘-’ and find the average of the two data and fixed the data with one number.

For example, if the data was 1223 – 3122. I had taken the two data 1223 and 3122 and find average like (1223+3122)/2 and fixed the data with that average.

Next, I had found the price per each square feet in the data with the formula price \* 100000 / total square feet. Next, I had done some data preprocessing, let me say with one case study, take a house which has price 120K which is 2BHK house, and now take another house which has price 90Kwhich is 4BHK house. In the above case study, the 4BHk house has less price than the 2BHk house, which is wrong as per our consumption, the 4BHK house must has more price than the 2BHk house. And I had deleted these types of the data in my data frame.

As per now the data has been cleaned and now I had done the modelling phase.

Data Modelling:

The data has been cleaned and now I had separated the independent variable and dependent variable into train and test data.

Let us view the pair plot of the data:

Chart, scatter chart

Description automatically generated

Dist plot of the prices:

Chart, histogram

Description automatically generated

As I had split the data into train and test. The train data contains four columns that is location, size, bathrooms and square feet. And the test data contains price column. Here, I had used OneHotEncoder, StanderScaler and pipeline concepts for fitting the train data into a particular model. Here we are using pipeline because, we have location data which is not numeric type which is string type, for that case first we need to convert the location data from object data type to numeric data type for that purpose I had used OneHotEncoder and as to scale the number from large sclare to small scale I used StandardScaler and used one particular prediction model to fit the data and finally we need to do it sequentially like first we use onehotencoder and then next standardscaler and last fitting into the model, for this purpose we used make pipeline concepts.

Makepipeline systeax was

make\_pipeline(OneHotEncoder([‘location’]),scaler,lir)

We had transformed both train and test dataset from text to numeric form.

Next, we want to find which is the best algorithm to predict the values of test dataset. For that purpose I had split the train dataset into x\_train, x\_test, y\_train, y\_test. Now fit the x\_train and y\_train data to the Linear Regression, Lasso Regression and Ridge Regression and find the r2\_scores of predicted values and y\_test.

As a result,

For linear regression model I had got:

Chart, scatter chart

Description automatically generated

Scatter plot of the actual and predicted data

with linear regression model.

As you see in the above graph, it sows the data was positive fitted to the model.

And I had got 82.51 % of r2\_score with linear regression model.

With lasso regression:

Chart, scatter chart

Description automatically generated

I had got 81.46% of r2 score with lasso regression.

With ridge Regression:

Chart, scatter chart

Description automatically generated

I had got 82.52 % of r2 score with ridge regression.

Prediction Results:

|  |  |
| --- | --- |
| **Model** | **R2 score** |
| Linear Regression | 82.51 |
| Lasso Regression | 81.46 |
| Ridge Regression | 82.52 |

As a result I had got the linear Regression and ridge regression as the best models for house price prediction. So, in that both I had taken Ridge regression models to convert into .pkl file. This is due to with the help of that pkl file we can able to predict one specific inputs output easily.

After that I had created a flask application which take inputs from the user, the inputs contains location, number of bathrooms, number of bedrooms and square feet.

Graphical user interface

Description automatically generated

As a result, we can get the price of the house,

Graphical user interface

Description automatically generated

**CONCLUSION:**

By this project “House Price prediction”, I had done some regression analysis on the data, as the data contains the continuous data. I had used Linear Regression model, Lasso regression model and Ridge Regression model for the prediction. With these three models I had find the r2 scores and in that all the three models have got nearly same value of r2 scores. In that I had choose Ridge Regression model and used it for the flask application development and created a web application where user give inputs and get the house price based on their requirements.

**FUTURE SCOPE:**

As to buy any house, most of the buyers will see from how long the house has been built, that means they will see the strongness of the house, whether it will be capable for these many people or not. And based on the soil of that land some people can say whether it will be strong enough or not. These are something most important factors that must be take care before the buying of the house.

**REFERENCES**

[1]. P. Durganjali and M. vani pujitha on House resale price prediction using classification techniques. M.Tech student, Department of Computer Science and Engineering V R Siddhartha Engineering College, Vijayawada. Presented in IEEE 6th International Conference on smart structures and systems ICSSS 2019.

[2]. Ch. Raga Madhuri, Anuradha and M. Vani pujitha Department of CSE, VRSEC, Vijayawada on house price prediction using regression techniques. Presented on IEEE 6th International Conference on smart structures and systems ICSSS 2019.

[3]. Ayush Varma, Abhijith sarma, Sagar Doshi and rohini nair KJ Somaiya College Of Engineering, Vidyavihar, Mumbai- 400077.

[4]. Manasa J, radha Gupta Dayananda College Of Engineering-RC Visveshwariah Technological University Bengaluru. Presented on IEEE Xplore Part Number: CFP20K58-ART; ISBN: 978-1-7281-4167-1.

[5]. Megha Rathi, Aditya Malik, Daksh Varshney, Rachita Sharma and Sarthak Me on Sentiment analysis of Tweets using Machine Learning Approach. Proceedings of 2018 Eleventh International Conference on Contemporary Computing, 2-4 August, 2018, Nodia, India.