A Project Report

On

“House Price Prediction”

Submitted in partial fulfillment of the

Requirement for the award of the degree of

B. TECH (C.S.E)

Under The Supervision of

Ms. Himanshi Sharma

Assistant Professor

Submitted By

Ayushi Srivastava

20SCSE1010078

Smriti Kumari Gupta

20SCSE1010424

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the project, entitled "HOUSE PRICE PREDICTION" in partial fulfilment of the requirements for the award of the BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING submitted in the School of Computer Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of SEPTEMBER-2021 to DECEMBER-2021, under the supervision of Mr. K Suresh, Assistant Professor, Department of Computing Science and Engineering, Galgotias University, Greater Noida.

The matter presented in the project has not been submitted by me for the award of any other degree of this or any other places.

Smriti Kumari Gupta (20SCSE1010424)

Ayushi Srivastava(20SCSE1010078)

Supervisor

Mr. K Suresh, Assistant Professor

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of Smriti Kumari Gupta(20SCSE1010424) and Ayushi Srivastava (20SCSE1010078) has been held on _____ and his/her work is recommended for the award of B. TECH

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING GALGOTIAS UNIVERSITY, GREATER NOIDA

INDIA

OCTOBER, 2021

ABSTRACT

Machine Learning plays a major role from past years in image detection, spam reorganization, normal speech command, product recommendation and medical diagnosis. Present machine learning algorithm helps us in enhancing security alerts, ensuring public safety and improve medical enhancements. Machine Learning systems also provides better customer service and safer automobile systems. In the present project we discuss about the prediction of future housing prices that is generated by machine learning algorithms. This system helps find a starting price for a property based on the geographical variables. By breaking down past market patterns and value ranges, and coming advancements future costs will be anticipated. This examination means to predict house prices with decision tree regressor. It will help clients to put resources into a bequest without moving towards a broker. The result of this project proved that the decision tree regressor gives an accuracy of 89%.

List of Figures

List of tables

Table no. **Table Name**

List of Figures

Table

Chapter 1: Introduction

1.1 Introduction

What is house price prediction? Need of house price prediction.

House prices increases every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help customer to arrange the right time to purchase a house.

What house price prediction will do?

. The main objectives of this study are as follows:

- To apply data pre-processing and preparation techniques in order to obtain clean data
- To build machine learning models able to predict house price based on house features
- To analyse and compare models' performance in order to choose the best model

1.2 Formulation of Problem

How the technique will work?

Basically, at first it will compare all the algorithms of machine learning. After that, the algorithm which will be most suitable and accurate, will be used.

In our project, decision tree regressor is the best algorithm. It will use all the data which are key factors for pricing of a house, after that it will perform the activity for predicting the house price.

1.2.1Tool and Technology Used

What tools used in House price prediction?

Notepad or IDE can be used to program all the functions that are required.

What technology used in house price prediction?

The technologies used are machine learning and python.

Chapter 2: Literature Survey/Project Design

Machine learning is a form of artificial intelligence which compose available computers with the efficiency to be trained without being veraciously programmed. Machine learning interest on the extensions of computer programs which is capable enough to modify when unprotected to new-fangled data. Machine learning algorithms are broadly classified into three divisions, namely; Supervised learning, Unsupervised learning and Reinforcement learning. Supervised learning is a learning in which we teach or train the machine using data which is well labelled that means some data is already tagged with correct answer. After that, machine is provided with new set of examples so that supervised learning algorithm analyses the training data and produces a correct outcome from labelled data. Unsupervised learning is the training of machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data. Unlike, supervised learning, no teacher is provided that means no training will be given to the machine. Therefore, machine is restricted to find the hidden structure in unlabelled data by our-self.

Reinforcement learning is an area of Machine Learning. Reinforcement. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behaviour or path it should take in a specific situation. Reinforcement learning differs from the supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of training dataset, it is bound to learn from its experience. Machine learning has many application's out of which one of the applications is prediction of real estate. The real estate market is one of the most competitive in terms of pricing and same tends to be vary significantly based on lots of factor, forecasting property price is an important modules in decision making for both the buyers and investors in supporting budget allocation, finding property finding stratagems and determining suitable policies hence it becomes one of the prime fields to apply the concepts of machine learning to optimize and predict the prices with high accuracy. The study on land price trend is felt important to support the decisions in urban planning. The real estate system is an unstable stochastic process. Investors decisions are based on the market trends to reap maximum returns. Developers are interested to know the future trends for their decision making. To accurately estimate property prices and future trends, large amount of data that influences land price is required for analysis, modelling and forecasting. The factors that affect the land price have to be studied and their impact on price has also to be modelled. An analysis of the past data is to be considered. It is inferred that establishing a simple linear mathematical relationship for these time-series data is found not viable for forecasting. Hence it became imperative to establish a non-linear model which can well fit the data characteristic to analyse and forecast future trends. As the real estate is fast developing sector, the analysis and forecast of land prices using mathematical modelling and other scientific techniques is an immediate urgent need for decision making by all those concerned. The increase in population as well as the industrial activity is attributed to various factors, the most prominent

being the recent spurt in the knowledge sector viz. Information Technology (IT) and Information technology enabled services. Demand for land started of showing an upward trend and housing and the real estate activity started booming. All barren lands and paddy fields ceased their existence to pave way for multistore and highrise buildings. Investments started pouring in Real estate Industry and there was no uniform pattern in the land price over the years. The need for predicting the trend in land prices was felt by all in the industry viz. the Government, the regulating bodies, lending institutions, the developers and the investors. Therefore, in this paper, we present various important features to use while predicting housing prices with good accuracy. We can use regression models, using various features to have lower Residual Sum of Squares error. While using features in a regression model some feature engineering is required for better prediction. Often a set of features multiple regressions or polynomial regression (applying a various set of powers in the features) is used for making better model fit. For these models are expected to be susceptible towards over fitting ridge regression is used to reduce it. So, it directs to the best application of regression models in addition to other techniques to optimize the result.

Functionality/Working of Project

We will present a program that will attempt to predict house prices based upon some features that are related to home prices. The following steps will be performed using machine learning and Python.

1. Import the required software libraries.

2. Access and import the dataset.

3. Data Analysis and Exploration.

4. Data Cleansing

4. Split the data into training and test data sets.

5. Train the model on the training data.

6. Make predictions on the test data.

7. Evaluate the model’s performance.

8. Draw conclusions from evaluations.

CODE IMPLEMENTATION

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import ensemble
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn import metrics
```

Access and Import the Data Set

```
# Load the data into Google Colab.
from google.colab import files
uploaded = files.upload() #Store the data into a data frame.
houses = pd.read_csv('housing.csv')
```

Data Analysis and Exploration

```
# Print the first 5 rows.
houses.head(5)
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

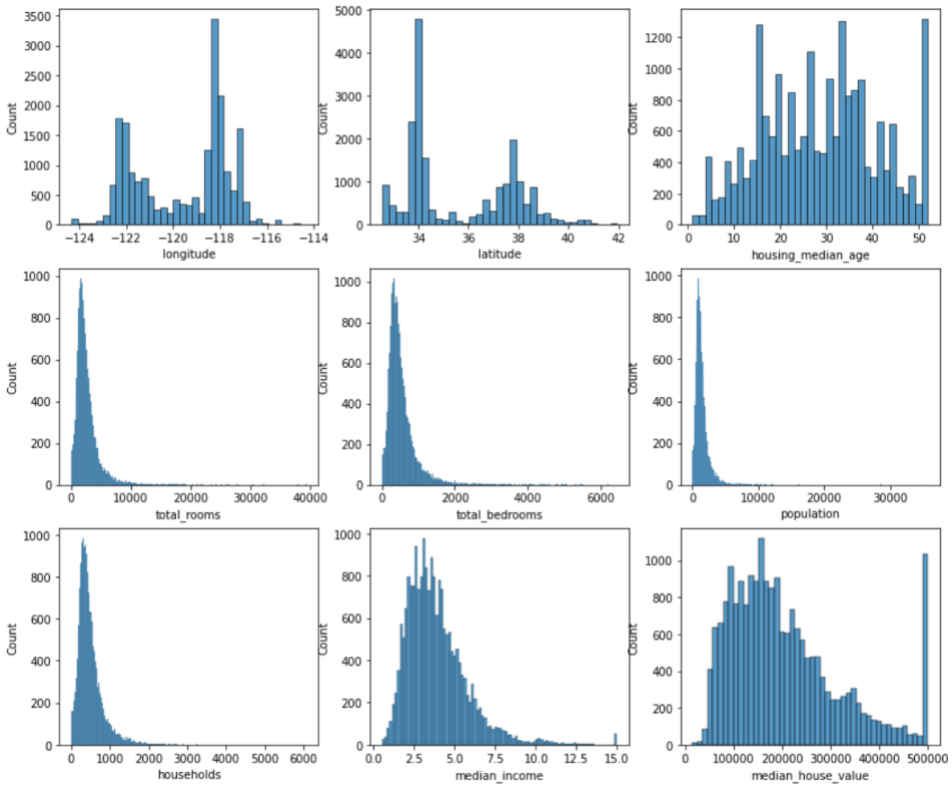
Visualize the distribution of all the numerical variables.

```
import warnings
warnings.filterwarnings("ignore")

plt.figure(figsize=(14, 12))

plt.subplot(3, 3, 1)
```

```
sns.histplot(houses['longitude'])
plt.subplot(3,                    3,                    2)
sns.histplot(houses['latitude'])
plt.subplot(3,                    3,                    3)
sns.histplot(houses['housing_median_age'])
plt.subplot(3,                    3,                    4)
sns.histplot(houses['total_rooms'])
plt.subplot(3,                    3,                    5)
sns.histplot(houses['total_bedrooms'])
plt.subplot(3,                    3,                    6)
sns.histplot(houses['population'])
plt.subplot(3,                    3,                    7)
sns.histplot(houses['households'])
plt.subplot(3,                    3,                    8)
sns.histplot(houses['median_income'])
plt.subplot(3,                    3,                    9)
sns.histplot(houses['median_house_value'])
```



Linear Regression Model

```
# Create model object.
model_lr = LinearRegression()
# Train the model on the training data.
model_lr.fit(x_training_data, y_training_data)
# Make predictions on the model using the test data.
# The predictions variable holds the predicted values of the        #
features stored in x_test_data.
predictions_lr = model_lr.predict(x_test_data)
```

Calculate the Root Mean Squared Error (RMSE) to measure the performance of the linear regression model.

```
from sklearn.metrics import mean_squared_error
lin_mse = mean_squared_error(y_test_data, predictions_lr)
lin_rmse = np.sqrt(lin_mse)
lin_rmse
```

74045.82617684643

The RMSE shows that the linear regression model has a typical prediction error of $74,045.

Decision Tree Model

```
# Create model object.
model_dt = DecisionTreeRegressor(random_state=42)
# Train the model on the training data.
```

```
model_dt.fit(x_training_data, y_training_data)
# Make predictions on the model using the test data.
predictions_dt = model_dt.predict(x_test_data)
```

Calculate the Root Mean Squared Error (RMSE) to measure the performance of the decision tree model.

```
from sklearn.metrics import mean_squared_error
lin_mse = mean_squared_error(y_test_data, predictions_dt)
lin_rmse = np.sqrt(lin_mse)
lin_rmse
```

78839.60896660246

The RMSE shows that the decision tree model has a typical prediction error of $78,839.

CONCLUSION

In this project, the decision tree machine learning algorithm is used to construct a prediction model to predict potential selling prices for any real estate property. This system provides 89% accuracy while predicting the prices for the real estate prices. By using this we can add many features in our project which will be helpful in predicting price more accurately.