Data-Driven Decisions: A Comprehensive Display of Machine Learning Models for Indian House Price Prediction

INTRODUCTION

This paper explores the application of machine learning in predicting house prices in various parts of India using specifically Linear Regression, decision trees, random forest, svm. To improve model performance data has been pre-processed before fitting data into model. The report aims to provide individuals with accurate price based on current trends to help take informed prices.

LTERATURE REVIEW

The paper aims to predict the average house price for each district in Beijing, Shanghai, Guangzhou, and Shenzhen using house price data from January 2004 to October 2016. The authors compare the performance of Autoregressive Integrated Moving Average (ARIMA) model and LSTM networks in terms of Mean Squared Error (MSE) for time series prediction. The LSTM model shows excellent properties in predicting time series data, and stateful LSTM networks and stacked LSTM networks are employed to further improve the accuracy of the house prediction model. The stacked LSTM has similar accuracy to the basic LSTM, but the authors note the need for further exploration to find better structures and parameters for stacked LSTM. The paper mentions the use of a single hidden layer RNN with 4 hidden units and ARIMA as a baseline system for house price prediction. The authors conducted experiments to determine the best number of hidden units for the prediction model. G Satish, Ch Raghavendran, Ch Rao, and Srinivasulu: Lasso Regression for Informed Decision-Making This paper focuses on predicting future housing prices using lasso regression, outperforming other models in terms of accuracy. The aim is to support house sellers or real estate agents in making informed decisions about house valuation. The use of machine learning algorithms is seen as contributing to the advancement of real estate policies and schemes. Consideration of Multiple Attributes The study by M Thamarai and S Malarvizhi concentrates on modelling house price prediction using techniques such as decision tree classification, decision tree regression, and multiple linear regression. The model considers attributes like the number of bedrooms, age of the house, and nearby facilities to predict house availability and prices.

ARCHITECTURE

We endeavour to forecast house prices in India through a comprehensive exploration of various machine learning models. Our investigation relies on a dataset comprising over 3000+ real estate records from all over India. Employing a diverse array of algorithms, such as Decision Trees, Linear Regression, Random Forest, XGBoost, and Support Vector Machines, we aim to ascertain the most effective model for predicting house prices in the Indian real estate market.

The initial phase of our research involved the acquisition of a robust dataset, comprising diverse real estate entries. Subsequently, we undertook a meticulous pre-processing of the data—a critical step in preparing and refining raw data for meaningful analysis. This entailed addressing issues such as handling missing values, encoding categorical variables, scaling features, and normalizing data. By systematically attending to these pre-processing tasks, we ensured that the input data was well-

formatted, thereby enhancing the performance of our machine learning models and facilitating accurate predictions.

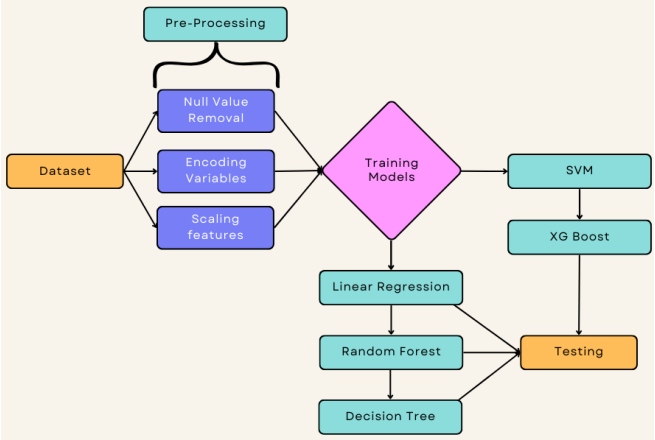
The first pre-processing step involved the removal of null values from the dataset, mitigating the potential interference with prediction accuracy. Following this, we performed encoding on the data, a process that entails converting qualitative information into numerical format. This conversion is pivotal for enabling machine learning algorithms to interpret and process the data effectively.

Scaling of features was the subsequent pre-processing step. This involved normalizing or standardizing numerical values to a consistent range, ensuring equitable contributions in the models. This step is crucial for preventing features with larger scales from dominating the learning process.

Having successfully completed the pre-processing phase, the study progressed to the model-building stage. We employed a suite of machine learning models, namely Random Forest, Decision Tree, Linear Regression, Support Vector Machines (SVM), and XGBoost. These models were selected for their proven efficacy in handling diverse datasets and predicting housing prices.

Post-model construction, we transitioned to the testing phase, where the predictive capabilities of each model were rigorously evaluated. Performance metrics such as accuracy, Mean Squared Error (MSE), and R2 score were employed to gauge the effectiveness of each model. This comprehensive testing phase aimed to discern which model yielded the most accurate and reliable predictions for house prices in the Indian real estate market.

In summary, this research represents a systematic and rigorous exploration of machine learning models for predicting house prices in India. The methodology encompassed robust data pre-processing, judicious model selection, and thorough testing, culminating in a comprehensive evaluation of the models' predictive capabilities. The findings of this study hold valuable insights for stakeholders in the real estate domain, aiding in informed decision-making and strategic planning.



METHODOLOGY

In this study, we used several well-known machine learning methods. Support vector machines (SVM), random forest, XGBoost, Decision Tree, and linear regression were some of the methods used in our investigation.

Algorithms: In the process of developing this model, various machine learning algorithms were studied. The model is trained on Support vector machines (SVM), random forest, XGBoost, Lasso regression, and linear regression. Out of this Decision Tree gives highest accuracy in prediction of housing prices and the next highest accuracy achieved is by Random Forest.

Implementation

- ## Data Collection

Gather a dataset either from Kaggle, that includes relevant features of houses such as location, number of rooms, square feet, and sale prices. Ensure the dataset that has the features that you are about to consider.

```
#exploration of training data
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29451 entries, 0 to 29450
Data columns (total 12 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   POSTED_BY              29451 non-null  object
 1   UNDER_CONSTRUCTION     29451 non-null  int64
 2   RERA                   29451 non-null  int64
 3   BHK_NO.                29451 non-null  int64
 4   BHK_OR_RK              29451 non-null  object
 5   SQUARE_FT              29451 non-null  float64
 6   READY_TO_MOVE          29451 non-null  int64
 7   RESALE                 29451 non-null  int64
 8   ADDRESS                29451 non-null  object
 9   LONGITUDE              29451 non-null  float64
 10  LATITUDE               29451 non-null  float64
 11  TARGET(PRICE_IN_LACS)  29451 non-null  float64
dtypes: float64(4), int64(5), object(3)
memory usage: 2.7+ MB
```

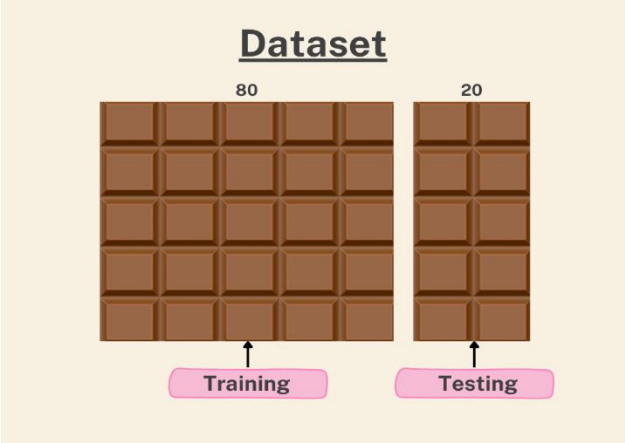
- ## Data Pre-processing

Clean and prepare the collected data for model training. Handle missing values, perform feature scaling to bring features to a similar range, encode categorical variables, and address outliers. Additionally, one can explore feature engineering techniques to create new meaningful features.

- ## Splitting the Dataset

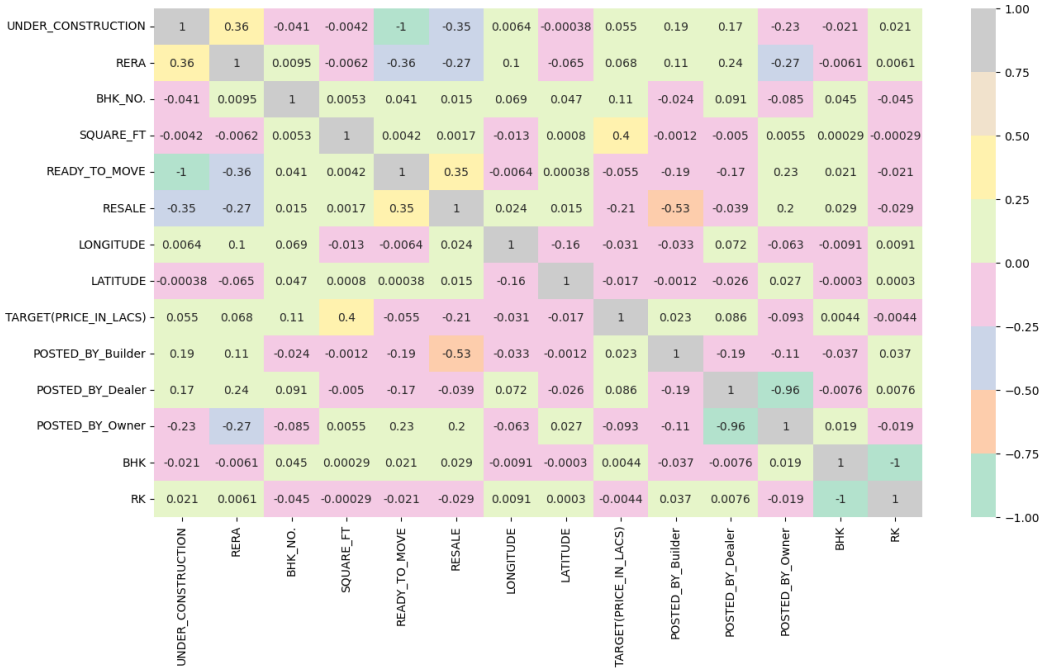
Now we perform splitting operation where we split the data in two parts in the ratio of 80:20

Where 80% of data is used for training the model and 20% is used for testing the model.



- ## Heatmap Correlation

  In our exploratory data analysis (EDA) for house price prediction, we created a correlation heatmap to examine the relationships between the variables. The correlation heatmap visually represents the strength and direction of correlations between these variables.



- ## Model Training

  - ### Decision tree

```
from sklearn.tree import DecisionTreeRegressor
model=DecisionTreeRegressor(criterion='squared_error',splitter='best')

model_fit=model.fit(x_train,y_train)
y_pred_dec=model.predict(x_test)

print(r2_score(y_test, y_pred_dec))
print(mean_absolute_error(y_test, y_pred_dec))
```

  - ### Randon Forest

```
from sklearn.ensemble import RandomForestRegressor
regressor = RandomForestRegressor(n_estimators=100, random_state=42)

regressor.fit(x_train, y_train)

y_pred_rf = regressor.predict(x_test)

print(r2_score(y_test, y_pred_rf))
print(mean_absolute_error(y_test, y_pred_rf))
```

  - ### Linear Regression

```
from sklearn.linear_model import LinearRegression
```

```
regressor = LinearRegression()
regressor.fit(x_train, y_train)
y_pred_lr=regressor.predict(x_test)

from sklearn.metrics import r2_score
print(r2_score(y_test, y_pred_lr))
from sklearn.metrics import mean_absolute_error
print(mean_absolute_error(y_test, y_pred_lr))
```

- SVM

```
from sklearn.svm import SVR
svm_reg = SVR(kernel='rbf', C=100, gamma='auto', epsilon=0.5)
svm_reg.fit(x_train, y_train)

y_pred_svr = svm_reg.predict(x_test)

print(r2_score(y_test, y_pred_svr))
print(mean_absolute_error(y_test, y_pred_svr))
```

- XG Boost

```
from xgboost import XGBRegressor

my_model = XGBRegressor(max_depth=5, learning_rate=0.1, gamma=0.2, n_estimators=500,
min_child_weight=5, subsample=0.8,colsample_bytree=0.8)
my_model.fit(x_train, y_train)
y_pred_xg = my_model.predict(x_test)

print(r2_score(y_test, y_pred_xg))
print(mean_absolute_error(y_test, y_pred_xg))
```

RESULTS AND ANALYSIS

After our testing on the dataset which we acquired and pre-processed we came to the following results

Sl. No	Model	Score	MSE
1	Radom forest	0.742024709	35.21625083
2	Decision tree	0.780241652	38.07502034
3	Linear regression	0.40622028	131.349713
4	XG Boost	0.725058858	45.16874333
5	SVM	0.366275146	88.88291751

CONCLUSION

The "House Price Prediction Using Machine Learning" project aims to predict house prices with a 78% accuracy which our best predicted value, leveraging diverse features.

To distinguish our model, we prioritize incorporating additional parameters, addressing budget constraints, and minimizing financial risks for prospective homebuyers. Utilizing a range of algorithms for comprehensive valuation, the model considers socio-economic factors and market trends.

This approach aims to benefit users significantly by providing nuanced insights into housing prices, empowering informed decision-making in real estate transactions. The project stands out for its holistic

consideration of numerous influencing elements, ensuring a distinctive and effective predictive tool for navigating the dynamic housing market.

REFERENCES

1. Lancaster, K. J. 'A new approach to consumer theory', The Journal of Political Economy, 1966, Vol. 74, No. 2, pp. 132- 157.
2. T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), 2018, pp. 35-42, doi: 10.1109/iCMLDE.2018.00017.
3. Chengke Zou, The House Price Prediction Using Machine Learning Algorithm: The Case of Jinan, China, University California Santa Barbara, Jinan, China.
4. M. Jagan Chowhaan, D. Nitish, G. Akash, Nelli Sreevidya and Subhani Shaik, Machine Learning Approach for House Price Prediction, Asian Journal of Research in Computer Science.
5. Cesar Vasquez, Vinodh Chellamuthu, PhD, House Price Prediction with Statistical Analysis in Support Vector Machine Learning for Regression Estimation, Mathematics, Dixie State University.
6. Dr. M. Thamarai and Dr. S P. Malarvizhi, House Price Prediction Modelling Using Machine Learning.
7. G. Naga Satish, Ch. V. Raghavendran, M.D. Sugnana Rao, Ch. Srinivasulu, House Price Prediction Using Machine Learning.
8. Chen, Xiaochen, et al. House Price Prediction Using LSTM. Sept. 2017.