SVKM's NMIMS
Mukesh Patel School of Technology Management & Engineering
A.Y. 2023 - 24
Course: Machine Learning

Project Report

Program	Btech Artificial Intelligence	
Semester	4	
Name of the Project:	Real estate price prediction	
Details of Project Members		
Batch	Roll No.	Name
B1	IO14	Atharva Chavan
B1	I015	Avnish Chitrigi
B1	I016	Ayush Kadam
B1	I026	Dipto Sen
Date of Submission: 4-4-24		

Contribution of each project Members:

Roll No.	Name:	Contribution
14	Atharva Chavan	Code, backend and Report
15	Avnish Chitrigi	Code, ppt and backend
16	Ayush Kadam	Code, backend and ppt
26	Dipto Sen	Code, report and backend

Github link of your project:

Note:

1. Create a readme file if you have multiple files
2. All files must be properly named (I004_MLProject)
3. Submit all relevant files of your work
4. **Plagiarism is highly discouraged (Your report will be checked for plagiarism)**

Rubrics for the Project evaluation:

- Evaluation of project will be based on following rubrics
- Domain knowledge and literature review in the selected topic (5 marks)
- EDA, Implementation, and performance metrics used (10 marks)
- Beyond classroom knowledge gained and implemented (5 marks)

Project Report

Real Estate Price Prediction using multiple linear regression

by

Atharva Chavan, I014

Avnish Chitrigi, I015

Ayush Kadam, I016

Dipto Sen, I026

Course: Machine Learning

AY: 2023-24

Table of Contents

I. Storyline or Applications of Project

Real estate price prediction is a vital aspect of the real estate industry, facilitating various stakeholders such as buyers, sellers, investors, and financial institutions in making informed decisions. Leveraging machine learning techniques like multiple linear regression, this project aims to predict real estate prices based on various features such as location, number of bedrooms, and property type. Financial institutions utilize such models for risk assessment during mortgage approvals.

II. Literature Review

Title of the Paper	Published Year	Major Contributions	Algorithms Used	Performance
Housing Price Prediction Based on Multiple Linear Regression	2021	Analyzes the effectiveness of multiple linear regression for real estate valuation.	Multiple Linear Regression	Achieved an R-squared value of 0.78
Applying Multiple Linear Regression in house price prediction	2019	Demonstrates the application of multiple linear regression for house price prediction.	Multiple Linear Regression	Achieved a Mean Absolute Error (MAE) of $17,000
Incorporating Multiple Linear Regression in Predicting the House Prices Using a Big Real Estate Dataset with 80 Independent Variables	2014	Investigates the impact of many features on model performance.	Multiple Linear Regression with Feature Selection	Achieved higher accuracy with a reduced feature set

III. Data Preprocessing and Exploratory data Analysis with Visualization

Perform all data cleaning and preprocessing steps. Perform data visualization using different charts/graph. Make sure to write your own inferences

1. Data Cleaning:

- Missing Values: You replaced missing values denoted by "NA" with np.nan (Not a Number) to ensure proper handling during data processing.
- Inconsistencies: The code removes columns starting with "Unnamed" which might indicate inconsistencies in column naming during data collection.

2. Data Transformation:

- Log Transformation: For features with positive skew, applying a logarithmic transformation can help normalize the distribution.

3. Feature Engineering:

- Combining Features: You could have combined features like "living area" and "number of bedrooms" into a new feature representing "living space per bedroom."

4. Exploratory Data Analysis (EDA):

- Distribution of Price: Include a visualization (histogram or kernel density estimation plot) to illustrate the distribution of the target variable (price).
- Relationship between Features and Price: Create scatter plots or heatmaps to explore the relationships between features (e.g., square footage, number of bedrooms) and the price variable.
- Data Selection: The code snippet merged = merged[(merged['Latitude'] > 0) & (merged['Longitude'] > 0)] filters the data to exclude entries with invalid latitude or longitude values (0 or negative values). This ensures the model is trained on data with valid location information.

5. Visualizations:

- Integrate visualizations (replace with your actual plots) to explore the distribution of the target variable (price) and its relationship with independent variables (e.g., square footage, number of bedrooms).

```
# Sample code to generate a histogram of Price
plt.figure(figsize=(8, 6))
plt.hist(merged["Price"], color='blue', edgecolor='black')
plt.xlabel("Price")
plt.ylabel("Number of Properties")
plt.title("Distribution of Price")
plt.grid(True)
plt.show()

# Sample code to generate a scatter plot
plt.figure(figsize=(8, 6))
plt.scatter(merged["Living Area"], merged["Price"], color='blue')
plt.xlabel("Living Area (sq ft)")
plt.ylabel("Price")
plt.title("Living Area vs. Price")
plt.grid(True)
plt.show()
```

IV. Machine learning models with hyper parameter tuning

- Several machine learning models are employed for real estate price prediction, including multiple linear regression, decision tree regressor, random forest regressor, and Boost regressor. Hyperparameter tuning techniques such as grid search or random search are applied to optimize the models' performance.

V. Performance Evaluation

- Interpretation: Based on the R-squared value, the model explains a sizable portion of the variance in price. However, there's still room for improvement as evidenced by the RMSE and MAE values.

VI. Comparison of different techniques used

- **Model Comparison:**

1. Multiple Linear Regression:

Pros: Simple to interpret, computationally efficient, provides insights into feature importance.
Cons: Assumes linear relationships between features and target variable, might underperform for complex relationships.

2. XGBoostRegressor:

Pros: Powerful for capturing complex non-linear relationships, often achieves high accuracy.
Cons: More complex to tune, less interpretable compared to linear models.

3. RandomForestRegressor:

Pros: Handles non-linearity and high dimensionality well, robust to outliers.
Cons: Can be computationally expensive for large datasets, feature importance interpretation can be less straightforward compared to linear models.

Interpretation: Based on the results, XGBoostRegressor might have achieved the best performance with the lowest MSE and MAE, followed by Random ForestRegressor and then multiple linear regression. This suggests that for your data, the more complex models were able to capture non-linear relationships and improve prediction accuracy compared to the linear model.

VII. Deployment/GUI/ Learning beyond classroom

- Tools/software/ libraries used:

1) Python
2) Scikit-learn
3) Matplotlib
4) Seaborn
5) Folium
6) Plotly
7) Eli5
8) Graphviz
9) NetworkX
10) Geopy

- Screenshot and Description of the Demonstration of project (If GUI is made)

- During the project, I gained hands-on experience in data preprocessing, exploratory data analysis, machine learning model development, hyperparameter tuning, and performance evaluation. Additionally, I explored various Python libraries and visualization techniques, enhancing my skills beyond the classroom curriculum.

VIII. Learnings and challenges you faced while doing the Project

:

- 50% project was based on data preprocessing and data cleaning ,10% included data visualization 40% model creation.
- Due to data cleaning and preprocessing we learnt about how data we need should be clean and accurate, slight change in dataset can affect the results of prediction and model easily.
- Model training was bit tricky as the results were under average.
- Most of our time was spent on data preprocessing and one hot encoding.

IX. Conclusion

- In conclusion, real estate price prediction using multiple linear regression and other machine learning techniques offers significant value to the real estate industry. By leveraging advanced analytics and predictive modeling, stakeholders can make informed decisions, mitigate risks, and capitalize on lucrative opportunities in the dynamic real estate market.