Pratibodh
A Journal for Engineering
A free and Open Access Journal
Homepage: https://pratibodh.org
RATNA & BM RANVEER Foundation

Real Estate House Price Predictor Model

Sherly Sharma, Sarthak Kumar, Sarthak Jain, Ruchi Sharma
Department of Artificial Intelligence & Data Science
Jaipur Engineering College & Research Centre
sherlysharma.ai25@jecrc.ac.in

Abstract

Training Project Description: REAL ESTATE HOUSE PRICE PREDICTOR MODEL. This research delves into the development of a Real Estate House Price Predictor Model using data science techniques. It all begins with the collection of historical property data from various sources, followed by rigorous data cleaning and feature engineering..

Keywords: *Python, Pandas, Numpy, Matplotlib, Sklearn, HTML, CSS, Javascript.*

Article Status

Available online :

1. Introduction

This data science project series walks through step by step process of how to build a real estate price prediction website. We will first build a model using sklearn and linear regression using banglore home prices dataset from kaggle.com. Second step would be to write a python flask server that uses the saved model to serve http requests. Third component is the website built in html, css and javascript that allows user to enter home square ft area, bedrooms etc and it will call python flask server to retrieve the predicted price. During model building we will cover almost all data science concepts such as data load and cleaning, outlier detection and removal, feature engineering, dimensionality reduction, gridsearchcv for hyperparameter tunning, k fold cross validation etc.

A Real Estate House Price Predictor Model in data science is a sophisticated and data-driven solution designed to forecast property prices in the ever-fluctuating real estate market. Leveraging data collection, preprocessing, and advanced machine learning techniques, this model helps prospective buyers and sellers make informed decisions by providing accurate predictions of house prices.

2. Research Methodology

Creating a house price predictor model in real estate using data science involves several steps in the research methodology. Here's a simplified overview:

Problem Definition:
1.Define the objective clearly. Are you predicting prices based on location, size, amenities, etc.?
Data Collection:
1.Gather relevant data: housing prices, location, size, number of rooms, nearby amenities, historical data, etc.

2.Utilize public datasets, APIs, web scraping, or collaborate with real estate agencies.

Data Preprocessing:
1.Handle missing values, outliers, and inconsistencies.
2.Normalize or scale numerical features.
3.Encode categorical variables (one-hot encoding, label encoding).

Feature Selection/Engineering:
1,Identify important features affecting house prices.
2.Engineer new features if necessary (e.g., calculating price per square foot).

Model Selection:
1.Choose appropriate models (linear regression, decision trees, random forests, neural networks, etc.).
2.Split data into training and testing sets for evaluation.

Model Training:
1.Train the selected models on the training dataset.
2.Tune hyperparameters to improve model performance (cross-validation, grid search, etc.).

Model Evaluation:
1.Evaluate models using appropriate metrics (RMSE, MAE, R^2 score, etc.) on the test dataset.
2.Compare different models to select the best-performing one.

Model Deployment:
1.Deploy the model to predict house prices.
2.Create a user-friendly interface (web app, API, etc.) for users to input data and get predictions.

Monitoring and Maintenance:
1.Monitor the model's performance over time.
2.Update the model periodically with new data or retraining if necessary.

Ethical Considerations:
1.Ensure fairness, transparency, and ethical use of the model.
2,Avoid biases in data and model predictions

3.Proposed Approach

The proposed approach involves collecting diverse housing data from reliable sources, followed by meticulous data preprocessing to clean, normalize, and encode features. Exploratory Data Analysis (EDA) helps uncover correlations and crucial features influencing house prices, guiding the selection and engineering of impactful features. Experimentation with various regression models, validation, and optimization through hyperparameter tuning leads to the selection of the best-performing model. Evaluation metrics on a validation set validate the model's accuracy before deployment. Once deployed, the model's performance is continuously monitored, updated with new data, and maintained ethically. Comprehensive documentation, including methodologies and model performance, supports transparency and stakeholder understanding throughout the process.

4. Backend :

The backend in developing a house price predictor model in real estate involves a robust technical infrastructure that orchestrates various stages of data processing, model creation, deployment, and ongoing maintenance.

Firstly, it manages the Data Pipeline, encompassing processes for Data Collection from diverse sources like real estate databases, public records, and APIs. This involves scripting or automated procedures to gather relevant housing data. Subsequently, the backend handles Data Preprocessing tasks, which include cleaning the data, dealing with missing values or outliers, and preparing the data for analysis and model development. This stage typically involves normalization of numerical data and encoding categorical variables.

Once the data is prepared, the backend facilitates the Model Development phase. It involves setting up mechanisms for Feature Engineering to create or select pertinent features affecting house prices. Additionally, the backend handles the computational load for Model Training and Tuning, allowing for the testing of various models, their hyperparameter optimization, and selection of the most suitable model for deployment.

The backend infrastructure also plays a critical role in Model Deployment. This includes establishing an API or Service that hosts the trained model, enabling user interaction and input for predictions. Scalability and performance are key considerations here to ensure the system can efficiently handle multiple requests.Furthermore, it's essential for the backend to manage Monitoring and Maintenance aspects. This involves setting up mechanisms to continually Monitor Performance metrics of the deployed model, allowing for timely identification of any degradation or issues. An Update Mechanism is integral, enabling seamless incorporation of new data or improved algorithms into the model without disrupting services.

Ethical considerations such as Data Governance and Bias Mitigation are embedded within the backend infrastructure. This involves establishing robust systems to ensure data privacy, security, and compliance with regulations. Additionally, measures are put in place to identify and mitigate biases in the model, promoting fairness and ethical use.Lastly, the backend incorporates mechanisms for Documentation and Reporting, storing logs of model changes, data sources, and performance metrics. It also provides platforms to house comprehensive documentation detailing the entire process, methodologies applied, and specific details about the model, facilitating transparency and understanding for stakeholders involved in the project.

5. Sklearn:

In a real estate house price prediction project, scikit-learn serves as a fundamental tool for numerous stages, starting with Data Preparation and Preprocessing. Initially, the library assists in managing the collected housing data, cleaning it, and preprocessing it to ensure its quality and suitability for analysis. Utilizing modules like SimpleImputer, it handles missing values, while MinMaxScaler or StandardScaler can standardize numerical features. Categorical variables can be effectively encoded using tools such as OneHotEncoder, making the data ready for model training.

Moving to the Model Development phase, scikit-learn facilitates the creation and optimization of predictive models. It provides a range of algorithms such as LinearRegression, RandomForestRegressor, and tools for feature selection like SelectKBest or PolynomialFeatures. Through modules like GridSearchCV or RandomizedSearchCV, it enables thorough exploration of hyperparameters, aiding in the selection and fine-tuning of models that best suit the prediction task.

The subsequent Model Evaluation and Deployment stages rely on scikit-learn for assessing model performance and integrating it into real-world applications. Metrics like mean_squared_error or r2_score assist in evaluating model accuracy and reliability. Additionally, the library offers joblib for saving trained models, ensuring their seamless deployment in production environments or for future use.An illustrative workflow exemplifies this process: after splitting data into training and testing sets, scikit-learn is used for scaling features, training a model (such as RandomForestRegressor), and evaluating its performance by predicting house prices and calculating mean squared error. Finally, the trained model is saved using joblib for deployment or integration into systems handling house price predictions.

In summary, scikit-learn serves as a versatile and comprehensive toolkit throughout the various stages of a real estate house price prediction project, offering an array of functionalities for data handling, model development, evaluation, and deployment. Its flexibility and extensive library of algorithms make it a go-to choice for implementing machine learning-based solutions in real estate analytics and prediction tasks.

6. Matplotlib:

In a real estate house price prediction project, matplotlib serves as a powerful tool for visualizing data, model performance, and predictions, complementing the functionalities provided by scikit-learn. Starting with Exploratory Data Analysis (EDA), matplotlib facilitates data exploration by creating various plots like histograms, scatter plots, and box plots. These visualizations help in understanding the distributions of features, identifying relationships between variables, and spotting potential outliers or patterns within the housing dataset.

Moving to Model Evaluation and Performance Visualization, matplotlib plays a crucial role in assessing model accuracy and effectiveness. It aids in generating visual comparisons between actual and predicted house prices, typically through scatter plots. Additionally, residual plots created using matplotlib showcase the differences between predicted and actual values, providing insights into the model's performance and any patterns in prediction errors.

As the model gets deployed, matplotlib continues to be valuable in Visualizing Predictions. It enables the representation of predicted house prices geographically on maps, showcasing trends in predicted prices over time, or illustrating the influence of different features on house prices through graphical presentations. These visualizations offer stakeholders a clearer understanding of the model's predictions and underlying trends, aiding in decision-making processes related to real estate transactions or investments.

An example usage demonstrates how matplotlib can create scatter plots for exploring relationships between features and house prices, showcasing the comparison between actual and predicted prices, and generating residual plots for assessing the model's accuracy. Through these visualizations, stakeholders can gain insights into the dataset, evaluate model performance, and interpret predictions effectively, enhancing the project's comprehensibility and facilitating informed decision-making in the real estate domain.

7. Conclusion:

In conclusion, a Real Estate House Price Predictor Model in data science is a valuable tool for both consumers and professionals in the real estate market. It leverages data and machine learning to offer accurate predictions, enhancing transparency and efficiency in property transactions. However, to ensure its long-term relevance and effectiveness, ongoing enhancements and improvements are crucial. Future developments may include advanced algorithms, geospatial analysis, personalization, market insights, and a focus on user experience, among others. Staying responsive to evolving market needs and emerging technologies is key to maintaining the model's value and impact in the dynamic real estate sector.

8. Future Enhancement:

In the future, consider enhancing your Real Estate House Price Predictor Model by incorporating advanced algorithms, geospatial and time-series analysis, automated feature selection, and natural language processing. Offer customization and personalization, market insights, and mobile applications. Explore blockchain integration, automate data updates, provide predictive analytics, and improve the user experience. Foster community model retraining to adapt to changing market conditions and maintain prediction accuracy. Implement mechanisms for monitoring system performance and addressing any issues that arise.

9. References:

1.About Company. (n.d.). Retrieved from http://www.tutorialspoint.com/.

2.The Physics Classroom. (n.d.). Retrieved from https://www.physicsclassroom.com/.

3.Kelly, L., & Breault, K. (2006). Developing Educational Websites: Investigating Internet Use by Students and Teachers. In Proceedings of Thinking, Evaluating, Rethinking, ICOM-CECA Conference, Rome.

4.AglaSem Admission. (n.d.). Retrieved from https://admission.aglasem.com/

5.Sachan, N. (2019, February 20). Welcome to BHU Student Club, BHU Student Club. Retrieved from http://bhustudentclub.in/.
6.Jalote, P. (2003). An Integrated Approach towards Software Engineering. Narosa Publishing House.
7.Musciano, C., & Kennedy, B. (1996). HTML, The Definitive Guide. O'Reilly & Associates.
8.Powell, T. A. (2010). HTML & CSS: The Complete Reference. The McGraw-Hill Companies.
9.Flanagan, D. (2006). JavaScript: The Definitive Guide. O'Reilly Media, Inc.
10.Shenoy, A., & Sossou, U. (2014). Learning Bootstrap. Packt Publishing Ltd.
11.Kuhlman, D. (2011). A Python Book: Beginning Python, Advanced Python, and Python Exercises. Platypus Global Media.
12.Holovaty, A., & Kaplan-Moss, J. (2008). The Definitive Guide to Django: Web Development done right. Apress.
13. Lokhande, P. S., Aslam, F., Hawa, N., Munir, J., & Gulamgaus, M. (2015). Efficient way of Web Development using Python and Flask. International Journal of Advanced Research in Computer Science, 54-57.
14.Thakur, M. S. (2017). Review on Structural Software Testing Coverage Approaches. International Journal of Advance Research, Ideas and Innovations in Technology, 281-286.
15.MKdos,Encode -OS LTD 2011- presenthttps://www.django-rest-framework.org/
16.Read the docs -sphinx using theme https://channels.readthedocs.io/en/latest/
17.SpringGuide -Github : https://github.com/spring- guides/gs-messaging-stomp-websocket
18.Akash Shrivastava (2022)-Published in ScaleReal Medium - https://medium.com/scalereal/push- notifications-through-django-db528c303b92evaluation Metrics at ACC, PRE, REC, and F1 at 98.87, 98, 98.87, and 98.89, respectively. To ensure model trust, efficiency, and effectiveness, our research also provides model explanations both at the model level (global explanation) and at the instance level (local explanation)