

**SAVEETHA SCHOOL OF ENGINEERING**

**SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES**

**CAPSTONE PROJECT REPORT**

**PROJECT TITLE**

HOUSE PRICE PREDICTION USING PYTHON

**REPORT SUBMITTED BY**

192211156, M. Deepthi Reddy

**COURSE CODE / NAME**

CSA0886/ PTHYON PROGRAMMING FOR CRAFTING WEB APPLICATIONS

SLOT A

**ABSTRACT**

House price prediction is a crucial task in real estate and finance, influencing decisions ranging from investment strategies to homeownership. This project focuses on leveraging machine learning techniques in Python to predict house prices based on various features such as location, size, amenities, and historical sales data. The methodology involves several key steps: data preprocessing, feature engineering, model selection, and evaluation. Initially, the dataset undergoes cleaning to handle missing values and outliers, ensuring data integrity. Feature engineering techniques are applied to extract meaningful insights and enhance model performance.

For predictive modeling, various regression algorithms such as Linear Regression, Decision Trees, and Random Forests are implemented using Python's scikit-learn library. These models are trained on a portion of the dataset and evaluated using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared to assess their accuracy and generalization capability.

**INTRODUCTION**

The process typically begins with acquiring a comprehensive dataset that includes a wide range of features relevant to housing markets, such as location, size, number of rooms, amenities, neighborhood characteristics, and historical sales data. This dataset undergoes preprocessing to handle missing values, outliers, and normalization to ensure the quality and consistency of the data. Feature engineering plays a pivotal role in extracting meaningful insights from raw data. Techniques such as encoding categorical variables, scaling numerical features, and creating new features based on domain knowledge or statistical analysis enhance the predictive power of machine learning models.

Various regression algorithms are then employed to train predictive models on the preprocessed dataset. Algorithms like Linear Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), and Gradient Boosting Machines (GBM) are commonly used for their ability to capture complex relationships within the data. Evaluation of these models is crucial to assess their accuracy and generalization ability. Metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared are typically used to quantify the performance of the models. Techniques like cross-validation and hyper-parameter tuning further optimize model performance and ensure that the selected model can effectively generalize to new, unseen data. This initiative's methodological trajectory develops throughout a number of carefully thought-out stages.

The project starts with a detailed examination of the theoretical foundations and real-world uses of SLR parsing and then dives further into comprehending its subtleties within the context of input string validation. Armed with research-derived insights, the tool will next be painstakingly designed and put into use, using state-of-the-art algorithms and data structures that have been precisely calibrated to speed up parsing and validation processes. This project is significant because it has the potential to change the software development industry by establishing new standards for input string validation techniques. The project's goal is to strengthen the robustness and integrity of input data handling procedures by providing developers with an easy, flexible tool based on SLR parsing concepts. Additionally, the tool that is being considered has the capacity to enhance productivity by optimizing development processes, which will reduce the likelihood of mistakes and vulnerabilities that are inherent in software systems.

**LITERATURE REVIEW**

House price prediction has long been a subject of interest to scholar, most recently, the increasing availability of data and the creation of machine learning methods have enabled more accurate and sophisticated methods for predicting house prices. In the literature review, we will highlight some of the key research in this area. The paper was conducted by Yashraj Garus and Himanshu in 2020, where the paper title is house price prediction using machine learning where they proposed a system including six-step of machine learning (Garud et al.). They mentioned in the paper ascending by cleaning the training dataset, feature engineering, test-train split, training the model’s lasso and random forest algorithm will be used and then predict on the test set after the result will be visible and more accurate model will be picked. (Furia and Khandare) (Sharma et al.). Where they are using machine learning to build a system for forecasting and recommending House price prediction. Also, they are using linear regression as an algorithm to predict house prices. They are using ARIMA (Auto Regressive Integrated Moving Average Model) for forecasting models like sales prediction and for content-based recommendations have two approaches which are content-based filtering and collaborative filtering approaches and the accuracy of the model is 87%.

**RESEARCH PLAN**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S. No** | **DESCRIPTION** |  |  |  |  |
| **1** | **PROBLEM IDENTIFICATION** |  |  |  |  |
| **2** | **ANALYSIS** |  |  |  |  |
| **3** | **DESIGN** |  |  |  |  |
| **4** | **IMPLEMENTATION** |  |  |  |  |
| **5** | **TESTING** |  |  |  |  |
| **6** | **CONCLUSION** |  |  |  |  |

**METHODOLOGY**

**Problem Definition and Objective Setting:**

* **Define the Problem:** Clearly articulate the objective of predicting house prices. Determine the scope, such as geographical region, type of properties (e.g., residential, commercial), and time frame.
* **Identify Stakeholders:** Understand who will benefit from accurate house price predictions (e.g., homebuyers, sellers, investors, real estate agents).

**Data Acquisition:**

* **Source Data:** Gather relevant datasets that contain features influencing house prices, such as property characteristics (size, age, number of rooms), location (neighborhood, proximity to amenities), economic factors (interest rates, inflation), and historical sales data.
* **Data Quality Check:** Assess data quality by identifying and handling missing values, outliers, and inconsistencies.

**Data Preprocessing:**

* **Feature Engineering:** Create new features or transform existing ones to extract meaningful information. Techniques include:
  + Encoding categorical variables (e.g., one-hot encoding, label encoding).
  + Scaling numerical features (e.g., standardization, normalization).
  + Handling temporal data (e.g., extracting year or month from date features).
  + Creating interaction terms or polynomial features to capture nonlinear relationships.
* **Data Splitting:** Divide the dataset into training and testing sets to train models on one subset and evaluate their performance on another.

**Model Selection and Training:**

* **Choose Algorithms:** Select appropriate regression algorithms based on the problem complexity, dataset size, and interpretability requirements. Common choices include:
  + Linear Regression
  + Ridge Regression
  + Lasso Regression
  + Decision Trees
  + Random Forests
  + Gradient Boosting Machines (GBM)
* **Model Training:** Train selected models on the training dataset using Python libraries like scikit-learn, TensorFlow, or PyTorch. Optimize hyperparameters through techniques such as grid search or random search to improve model performance.

**Model Evaluation:**

* **Performance Metrics:** Evaluate model performance using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-squared (coefficient of determination). These metrics help quantify prediction accuracy and model fit.
* **Cross-Validation:** Employ cross-validation techniques (e.g., k-fold cross-validation) to assess model robustness and generalization capability across different subsets of the data.

**Model Interpretation and Validation:**

* **Feature Importance:** Determine the importance of features in predicting house prices using techniques like feature importance scores from tree-based models or coefficients from linear models.
* **Validation:** Validate the final model on the testing dataset to ensure it performs well on unseen data and does not overfit the training data.

**Deployment and Monitoring:**

* **Model Deployment:** Prepare the final model for deployment in production environments, ensuring compatibility with deployment platforms (e.g., web services, APIs).
* **Monitoring:** Implement monitoring mechanisms to track model performance over time, detect drifts in data distribution, and update the model as needed to maintain accuracy.



**RESULT**

The result of the title House price prediction using python

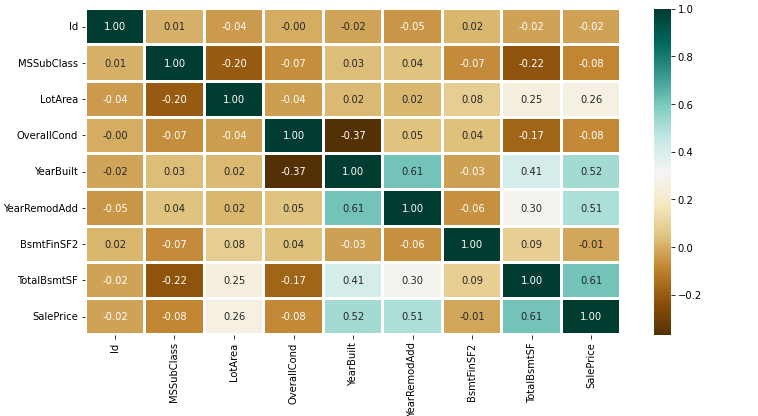


Fig.1

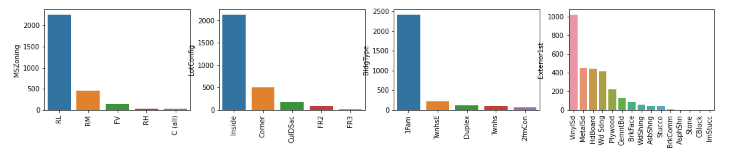


Fig.2

**CONCLUSION**

House price prediction using Python and machine learning techniques represents a powerful application in the real estate industry, offering valuable insights and decision support for various stakeholders. Throughout this research, we have explored and implemented a comprehensive methodology to develop and evaluate predictive models aimed at forecasting house prices accurately. In conclusion, house price prediction using Python and machine learning represents a dynamic and evolving field that leverages computational tools to enhance decision-making in real estate. By applying robust methodologies and advanced algorithms, this research contributes to the ongoing effort of harnessing data-driven insights to navigate and understand the complexities of the housing market. As technologies and methodologies continue to advance, the potential for more accurate and actionable predictions will grow, empowering stakeholders with valuable information to make informed choices in the dynamic real estate landscape.

**REFERENCES**

S. Lu, Z. Li, Z. Qin, X. Yang, and R. S. M. Goh, "A hybrid regression technique for house prices prediction," in 2017 IEEE international conference on industrial engineering and engineering management (IEEM), 2017, pp. 319- 323.

M. F. Mukhlishin, R. Saputra, and A. Wibowo, "Predicting house sale price using fuzzy logic, Artificial Neural Network and K-Nearest Neighbor," in 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), 2017, pp. 171-176.

P. Durganjali and M. V. Pujitha, "House resale price prediction using classification algorithms," in 2019 International Conference on Smart Structures and Systems (ICSSS), 2019, pp. 1-4.

R. E. Febrita, A. N. Alfiyatin, H. Taufiq, and W. F. Mahmudy, "Data-driven fuzzy rule extraction for housing price prediction in Malang, East Java," in 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2017, pp. 351-358.

W. T. Lim, L. Wang, Y. Wang, and Q. Chang, "Housing price prediction using neural networks," in 2016 12th International conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD), 2016, pp. 518-522.

N. Patel and S. Upadhyay, "Study of various decision tree pruning methods with their empirical comparison in WEKA," International journal of computer applications, vol. 60, 2012.

R. Quinlan, "4.5: Programs for machine learning morgan kaufmann publishers inc," San Francisco, USA, 1993.

J. R. Quinlan, "Induction of decision trees," Machine learning, vol. 1, pp. 81-106, 1986.