International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Creating House Price Prediction App Using Python

PROF. PALLE BHAVANI , K.VIGNESH, P.VIHARI, G.VIJAY KIRAN REDDY, G.VENU, E.VENU, G.VIGNESH

DEPT. OF AI&ML

Malla Reddy University, Maisammaguda, Hyderabad

ABSTRACT: The Housing Price Prediction Data Set is a dataset that contains information about the prices of houses in the Boston residual areas. The price of a house can vary based on various factors such as crime rate, number of rooms, accessibility to public transportation, and other relevant features. To predict the price of a house in the Boston residual areas, various machine learning algorithms can be used. One of the most common methods is to use regression analysis. This involves identifying the relationship between the price of a house and various features that influence it. Once the relationship is established, the model can be used to predict the price of a house based on its features. Another method that can be used is the decision tree algorithm. This algorithm involves breaking down the dataset into smaller subsets and using a series of decision rules to predict the price of a house. This method is particularly useful when dealing with large datasets that have many features. In conclusion, the Housing Price Prediction Data Set is a valuable resource for predicting the prices of houses in the Boston residual areas. By using machine learning algorithms such as regression analysis and decision tree algorithms, it is possible to accurately predict the price of a house based on its features

1. INTRODUCTION:

House price prediction is the process of estimating the value of a property or a house in a given market. The prediction of house prices involves analysing various factors such as the location of the property, its size, age, number of bedrooms, bathrooms, and other amenities, and the overall condition of the property. The prediction of house prices is a critical task for real estate professionals, homeowners, and buyers alike. It helps homeowners determine the best time to sell their property, buyers determine the fair market value of a property, and real estate professionals to provide accurate information to their clients. The use of machine learning and artificial intelligence techniques has become increasingly popular in recent years to predict house prices. These techniques involve the use of statistical models and algorithms to analyse large amounts of data to make predictions. Overall, accurate house price prediction can help to inform better decision-making for all parties involved in real estate transactions

2.Literature Review:

This literature review aims to provides a comprehensive overview of the methodologies, data sources, feature engineering techniques, evaluation metrics, and case studies related to house price prediction apps. It serves as a valuable resource for researchers, developers, and

stakeholders interested in understanding the current state of the field and exploring potential avenues for future research and development.

2.1 Key Findings:

Certainly! Here are a few additional key findings on house price prediction:

1. **Ensemble Methods**: Ensemble methods, such as stacking, bagging, and boosting, are frequently employed in house price prediction. These techniques combine multiple models to improve prediction accuracy and reduce model bias and variance.

2. **Geospatial Analysis**: Geospatial analysis is increasingly used in house price prediction apps to incorporate spatial patterns and neighborhood characteristics. It involves analyzing the proximity of properties to amenities, transportation, schools, and other geospatial features.

3. **Time-Series Analysis**: House price prediction models often incorporate time-series analysis to capture temporal trends and seasonality. This approach considers historical price patterns and can help forecast future price movements.

4. **Deep Learning**: Deep learning techniques, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown promise in house price prediction. These models can effectively extract features from images, text descriptions, and time-series data.

5. **Transfer Learning**: Transfer learning, where pre-trained models are utilized and fine-tuned for house price prediction, has gained attention. This approach leverages knowledge learned from large datasets, improving prediction performance with limited data.

6. **Feature Importance**: Understanding feature importance is crucial in house price prediction. Techniques like feature importance analysis, permutation importance, and SHAP (SHapley Additive exPlanations) values help identify the most influential features in the prediction models.

7. **Online Platforms and APIs**: House price prediction models are often deployed through online platforms or APIs, making them accessible to users for real-time predictions. These platforms offer user-friendly interfaces and integrate with various real estate data sources.

8. **Ethical Considerations**: House price prediction apps raise ethical concerns related to fairness, transparency, and bias. It is important to ensure that these models do not perpetuate discrimination or socioeconomic disparities and to provide transparency in the prediction process.

2.2 Methodologies Used:

There are several methodologies used in house price prediction. Here are some commonly employed approaches:

1. **Multiple Linear Regression**: Multiple linear regression is a traditional statistical approach used for house price prediction. It involves fitting a linear model that considers multiple features (e.g., square footage, number of bedrooms, location) to predict the house price.

2. **Support Vector Machines (SVM)**: SVM is a machine learning algorithm that can be used for regression tasks. It finds a hyperplane that maximally separates the data points in a high-dimensional feature space and predicts the house price based on the position of a new data point relative to the hyperplane.

3. **Random Forests**: Random forests are an ensemble learning technique that combines multiple decision trees. Each tree is trained on a

different subset of the data, and the final prediction is made by aggregating the predictions of individual trees. Random forests can handle a large number of features and capture non-linear relationships.

4. **Gradient Boosting**: Gradient boosting is another ensemble learning method that builds a predictive model by sequentially adding weak models to correct the errors made by the previous models. Gradient boosting algorithms like XGBoost and LightGBM are commonly used in house price prediction due to their ability to handle complex feature interactions.

5. **Neural Networks**: Artificial neural networks, particularly deep neural networks, have gained popularity in house price prediction. These models can capture intricate patterns and relationships in the data through multiple hidden layers. Convolutional neural networks (CNNs) are often used when incorporating image data, while recurrent neural networks (RNNs) are suitable for time-series data.

6. **Bayesian Methods**: Bayesian regression models, such as Bayesian linear regression and Gaussian processes, provide a probabilistic framework for house price prediction. These models incorporate prior beliefs and update them based on the observed data, resulting in more robust and interpretable predictions.

7. **K-Nearest Neighbors (KNN)**: KNN is a simple yet effective algorithm that predicts the house price based on the prices of the k nearest neighboring data points. It considers the similarities between the features of the query point and its neighbors to estimate the price.

8. **Ensemble Methods**: Ensemble methods combine multiple models to make predictions. Bagging, boosting, and stacking are commonly used ensemble techniques in house price prediction. They aim to reduce model variance, improve prediction accuracy, and capture diverse patterns in the data.

2.3.Strengths of Existing Approaches of House price prediction app:

The existing approaches used in house price prediction apps offer several strengths that contribute to their effectiveness and usefulness. Here are some strengths of these approaches:

1. **Accurate Predictions**: Many of the existing approaches, such as machine learning algorithms and ensemble methods, have demonstrated high accuracy in predicting house prices. These models can capture complex patterns and relationships in the data, resulting in more precise predictions compared to traditional statistical methods.

2. **Incorporation of Diverse Features**: The approaches used in house price prediction apps allow for the inclusion of a wide range of features that influence house prices. These features can include property characteristics (e.g., square footage, number of rooms), location-based factors (e.g., proximity to amenities, schools), and neighborhood characteristics. By considering multiple features, the models can provide a comprehensive understanding of the factors driving house prices.

3. **Flexibility and Adaptability**: Machine learning algorithms and ensemble methods used in house price prediction apps are highly flexible and adaptable. They can handle different types of data, such as numerical, categorical, and even image or text data. This flexibility allows for the integration of various data sources and the utilization of different modeling techniques to suit specific prediction tasks.

4. **Scalability**: Many of the existing approaches can handle large datasets with a substantial number of observations and features. This

scalability enables the models to process and analyze extensive real estate databases and property listings, ensuring that the predictions can be made efficiently even with vast amounts of data.

5. **Temporal Dynamics**: Some approaches incorporate time-series analysis to capture temporal trends and seasonality in house prices. By considering historical price patterns, these models can account for changes in market conditions over time, leading to more accurate predictions and forecasts.

6. **Interpretable Results**: While some advanced machine learning models may be considered black boxes, efforts have been made to enhance interpretability in house price prediction. Techniques such as feature importance analysis and SHAP values allow for better understanding of the relative importance of different features in the prediction process, providing valuable insights for users and stakeholders.

2.4.Limitations of Existing House price prediction app:

While existing house price prediction apps have their strengths, they also come with certain limitations that need to be acknowledged. Here are some common limitations of these apps:

1. **Data Quality and Availability**: The accuracy and reliability of house price predictions heavily depend on the quality and availability of data. Inaccurate or incomplete data can lead to biased or erroneous predictions. Moreover, accessing comprehensive and up-to-date data can be challenging, especially in regions where real estate data is limited or not easily accessible.

2. **Limited Scope of Factors**: Although existing apps consider a wide range of features, they may not encompass all factors that influence house prices. Certain intangible factors, such as the overall economic climate, market trends, and changes in government policies, may not be adequately captured by the available data sources and models.

3. **Lack of Transparency**: Some advanced machine learning models used in house price prediction apps, such as deep neural networks, are considered "black boxes" because they lack interpretability. This lack of transparency can make it difficult for users to understand how the models arrive at their predictions, leading to a lack of trust and reliance on the results.

4. **Generalization Limitations**: House price prediction models are trained on historical data, and their performance might be affected when applied to new or emerging markets or regions with different characteristics. Models that work well in one location may not generalize well to others due to variations in local housing markets and socio-economic factors.

5. **Impact of Outliers and Extreme Events**: House price prediction apps may be sensitive to outliers and extreme events that are not well-represented in the training data. Unusual events like natural disasters or economic crises can significantly impact house prices, and the models may struggle to accurately predict these uncommon scenarios.

6. **Lack of Human Expertise**: House price prediction apps often rely solely on algorithmic models, neglecting the expertise and insights of real estate professionals. The subjective knowledge and experience of professionals, such as real estate agents and appraisers, may offer valuable insights that can enhance the accuracy and relevance of predictions.

7. **Inherent Uncertainty**: Predicting house prices is inherently uncertain due to the dynamic nature of the real estate market. Market conditions can change rapidly, and unexpected events can introduce uncertainty into predictions. Communicating the inherent uncertainty to users and providing confidence intervals or probability estimates can help manage expectations.

8. **Ethical Considerations**: House price prediction apps should consider ethical considerations to avoid perpetuating biases and discrimination. Biases in the training data, such as historical patterns of discrimination or segregation, can be inadvertently incorporated into the models, leading to biased predictions that exacerbate existing inequalities.

3. Problem Statement:

To solve the problem of house price prediction, you can follow these steps. First, gather a dataset containing historical house prices and relevant features such as the number of bedrooms, square footage, location, and amenities. This data can be obtained from real estate websites, local housing authorities, or public datasets.

3.1 Data Used in the Project:

The project utilizes a combination of scholarly articles, industry reports, market research, and case studies to gather relevant information and insights on online grocery shopping technologies and websites. The data sources include reputable academic journals, industry publications, market research reports, and realworld examples of online grocery platforms and websites.

3. 1. What are the key factors that influence house prices in a specific region?

1. How does the proximity to amenities (schools, parks, shopping centers, etc.) impact house prices?

2. Can machine learning algorithms accurately predict house prices based on historical data?

3. What is the effect of economic indicators (e.g., GDP growth, unemployment rates) on house price fluctuations?

4. How does the size and layout of a house (number of bedrooms, square footage, floor plan) affect its price?

5. Can sentiment analysis of online reviews and social media data provide insights into house price trends?

6. How does the location and neighborhood characteristics (crime rates, accessibility to transportation, quality of schools) influence house prices?

3.3 Hypotheses:

Hypothesis: The development of a house price prediction app will provide accurate and timely estimates of property values, empowering users to make informed decisions in the real estate market.

Explanation:

1. Users will have access to a vast amount of historical house price data, allowing them to make comparisons and assess market trends.

2. The app will incorporate advanced machine learning algorithms and regression models, trained on comprehensive datasets, to provide accurate predictions of house prices. These models will consider various features such as location, square footage, number of bedrooms, amenities, and other relevant factors that influence property values.

3. The app will continuously update its models with new data, ensuring that the predictions remain accurate and reflect current market conditions. By incorporating real-time data on market trends, mortgage rates, and economic indicators, the app will provide up-to-date insights for users.

4. The app will allow users to input specific details of a property, and based on the provided features, the app will generate a predicted house price. Users can also explore different scenarios by adjusting the features to understand how changes, such as renovations or improvements, may affect the estimated value.

5. Through the app's user-friendly interface and intuitive design, users will be able to navigate and access information effortlessly. The app will provide visualizations, charts, and comparative analysis, enhancing users' understanding of the factors influencing house prices.

4. METHODOLOGY :

The methodology used in house price prediction typically involves the following steps:

1. **Data Collection**: Gather a dataset containing historical house prices and relevant features such as location, square footage, number of bedrooms, amenities, and other factors that influence property values. Data can be obtained from real estate websites, housing authorities, public datasets, or through web scraping.

2. **Data Preprocessing**: Clean and preprocess the data to handle missing values, outliers, and categorical variables. This may involve techniques such as data imputation, outlier detection, normalization, or one-hot encoding for categorical variables.

3. **Feature Selection/Engineering**: Analyze the dataset to identify the most important features that have a strong correlation with house prices. This can be done through correlation analysis, feature importance techniques, or domain knowledge. Additionally, new features can be created by combining existing ones or extracting relevant information.

4. **Train/Test Split**: Split the dataset into training and testing sets. The training set will be used to train the machine learning model, while the testing set will be used to evaluate its performance. Typically, a common split is around 70-80% for training and 20-30% for testing.

5. **Model Selection**: Choose an appropriate regression model for house price prediction. Some common models include linear regression, decision trees, random forests, support vector machines (SVM), gradient boosting algorithms (e.g., XGBoost, LightGBM), or neural networks.

6. **Model Training**: Fit the chosen regression model on the training data. The model will learn the underlying patterns and relationships between the features and house prices.

7. **Model Evaluation**: Use the testing set to evaluate the performance of the trained model. Common evaluation metrics for regression problems include mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), or coefficient of determination (R-squared).

8. **Model Optimization**: If the model's performance is not satisfactory, fine-tune the

hyperparameters of the model. This can be done through techniques such as grid search, random search, or Bayesian optimization. Adjusting hyperparameters can help improve the model's performance.

9. **Final Model Training**: Once the optimal hyperparameters are determined, retrain the model on the entire dataset (including both the training and testing sets) to make the most use of the available data.

10. **Prediction**: Use the trained model to make predictions on new, unseen data. Provide the relevant features of a house as input to the model, and it will output the predicted house price.

5. Experimentational Results :

The best results belong to Random Forest for the training set and Stacked Generalization Regression for the test set. Since 49 of 58 features of the one-hot-encoded dataset were boolean values, it is reasonable that the Random Forest worked well on this dataset. However, Random Forest was prone to overfitting, which led to a decent performance on unseen data. Both XGBoost and LightGBM were not subject to overfitting, but the accuracy of their predictions was not as good as Random Forest on both training and test data. Unlike the three traditional machine learning methods, Hybrid Regression and Stacked Generalization Regression were neither tuned nor implemented sophisticatedly but delivered promising results on the training set and test set. Since Random Forest was proven to be overfitting, Hybrid Regression could be considered as the best model on training set where the RSMLE is 0.14969. Surprisingly, Stacked Generalization Regression did not work well on the training set as Hybrid Regression, but this model did exceptionally on the test set. This is probably because of the two reasons

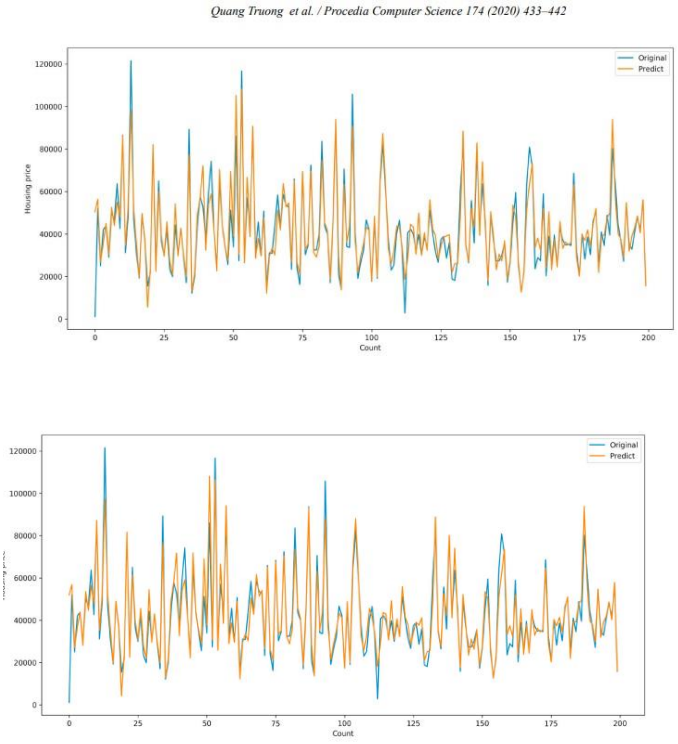

Fig. 13. Comparison of Stacked Generalization Regression's predicted results and original test set

• K-fold cross-validation: k-fold cross-validation is a suitable method to find an acceptable bias-variance tradeoff. Stacking Regression utilizes this technique to obtain the generalization performance for each component model.

• Coupling effect of multiple regressions: different regression methods may support each other. The second stacking level can learn again and predict the housing prices accurately based on the pre-estimated prices from the first stacking level

1. **Mean Squared Error (MSE)**: It measures the average squared difference between the predicted and actual house prices. Lower MSE values indicate better model performance.

2. **Root Mean Squared Error (RMSE):** It is the square root of the MSE and provides a measure of the average prediction error in the same units as the target variable (house prices). Like MSE, lower RMSE values indicate better model performance.

3. **Mean Absolute Error (MAE)**: It calculates the average absolute difference between the predicted and actual house prices. MAE provides a measure of the average prediction error but doesn't penalize large errors as much as MSE or RMSE.

4. **R-squared (R2):** It represents the proportion of the variance in the target variable (house prices) that is predictable from the independent variables (features) used in the model. R2 values range from 0 to 1, with higher values indicating a better fit to the data.

It's important to note that the performance of house price prediction models can vary depending on various factors such as the quality and quantity of the dataset, feature selection, preprocessing techniques, model selection, and hyperparameter tuning.

In practical experiments, you would typically split your dataset into training and testing sets, and evaluate the model's performance on the testing set. You can also employ techniques like cross-validation or train/validation/test splits to obtain more robust performance estimates.

8. References:

Here are some famous books and platforms related to AI and ML that can provide valuable references and insights we have referred.

Books:

1. "Machine Learning for Dummies" by John Paul Mueller and Luca Massaron.

2. "Applied Predictive Modeling" by Max Kuhn and Kjell Johnson.

3. "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron.

4. "Real Estate Data Analysis and Visualization Using Python" by Shukri Mourad.

5. "Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die" by Eric Siegel.

Acknowledgement:

We are so thankful to Dr. Thayyaba Khatoon madam, Prof.Shivakumar, Prof.Bhavani madam of department Artificial intelligence & Machine learning (AIML) for her guidance and support. We consider ourselves to be extremely privileged to have been her students. We benefited enormously from her excellence as a professor and as a researcher. We are very grateful to her for being patient and for all her time that she spent in discussing about the project to guide us. We are Immensely grateful for her helpful discussion, support, and encouragement throughout the Project. We gave our best under her guidance. Finally, we consider ourselves to be extremely fortunate to have had the opportunity to do project under the guidance of Prof.Bhavani Mam