INTERNNCRAFT
NAVIGATING SUCCESS THROUGH EXPERIENCE

House Price Analysis and Prediction

Project Overview:

This project aims to analyze a dataset of house prices to understand the factors affecting pricing, identify outliers, and develop a predictive model for future house prices. The dataset includes various features such as property type, location, price, number of bedrooms, and other relevant details.

Objectives

- **Data Cleaning and Exploration:** Clean the data to handle missing values, inconsistencies, and outliers. Explore the data to understand the distribution of house prices and other features, and identify potential relationships using visualizations.
- **Feature Engineering:** Create new features that might be relevant for price prediction, and encode categorical features into numerical values suitable for modeling.
- **Outlier Analysis:** Identify houses with significantly higher or lower prices than the average to understand the factors contributing to these outliers.
- **Predictive Modeling:** Develop a predictive model for future house prices using machine learning techniques.
- **Evaluation:** Evaluate the model's performance using appropriate metrics and improve it through hyperparameter tuning.

Methodology:

Data Cleaning and Exploration

1. Loading Data:

- Loaded the dataset and parsed the `date_added` column as datetime.
- Displayed the first few rows, a statistical summary, and information about the dataset.

2. Handling Missing Values:

- Dropped the 'page_url' column as it was not useful for analysis.
- Identified numeric and non-numeric columns, and checked for missing values.
- Handled missing values by creating new columns indicating missing values and converting categories to numeric codes.

3. Feature Engineering:

- Extracted new features from the `date_added` column such as year, month, day, day of the week, and day of the year.
- Dropped the original `date_added` column.

4. Data Visualization:

- Plotted the distribution of bedrooms, baths, and cities using bar plots.
- Plotted pairwise relationships of selected features against 'price' using scatter plots.
- Created a heatmap of the correlation matrix of numeric features.
- Plotted a scatter plot of 'Area Size' vs 'price'.

Feature Engineering

- Created new features like `date_added_year`, `date_added_month`, `date_added_day`, `date_added_day_of_week`, and `date_added_day_of_year`.
- Converted non-numeric columns to ordered categorical types.

Outlier Analysis

- Identified outliers in the `price` column using box plots.
- Analyzed these outliers to understand the factors contributing to the significantly higher or lower prices.

Predictive Modeling

1. Data Preparation

- Separated features (X) and target (y).
- Standardized the features using `StandardScaler`.
- Split the data into training and test sets.

2. Model Training:

- Initialized and fit a `RandomForestRegressor` model.
- Evaluated the model on the test set.

3. Hyperparameter Tuning:

- Defined a parameter grid for hyperparameter tuning.
- Used `RandomizedSearchCV` to find the best hyperparameters.
- Evaluated the model with the best hyperparameters on the test set.

Challenges

- **Missing Values:** Some columns had missing values which were handled by creating new columns indicating the presence of missing values and encoding categorical features.
- **Outliers:** Identifying and understanding the reasons for outliers was challenging but crucial for improving the model's performance.
- **Feature Engineering:** Creating meaningful features and encoding categorical features appropriately required careful consideration and domain knowledge.
- **Model Tuning:** Hyperparameter tuning was time-consuming but necessary to improve the model's performance.
-

Conclusion

The final Random Forest model achieved an R-squared value of 0.8737, indicating that it explains a significant portion of the variance in house prices. The model's performance metrics are as follows:

Mean Absolute Error (MAE): 3,090,620.42

Mean Squared Error (MSE): 142,333,050,317,583.88

Root Mean Squared Error (RMSE): 11,930,341.58

R-squared (R2): 0.8737

Explained Variance Score: 0.8737

Recommendations and Future Steps

1. Feature Engineering: Investigate additional features that might impact house prices, such as proximity to amenities, age of the property, and neighborhood safety.

2. Model Selection: Experiment with other machine learning models like Gradient Boosting, XGBoost, or neural networks to see if they can provide better performance.

3. Data Quality: Ensure higher data quality by reducing missing values and inconsistencies in future data collection.

4. Deployment: Consider deploying the model using a web application to make predictions on new data in real-time.

5. Regular Updates: Regularly update the model with new data to maintain its accuracy over time.