**House Hunting**

Sam Jones

**ABSTRACT**

This project analyzes housing prices in India using a dataset containing various features of residential properties. Exploratory data analysis was conducted to understand the distribution and characteristics of the data. Regression models were implemented to predict house prices based on factors such as area, number of bedrooms, and amenities. Data preprocessing techniques, including handling categorical variables, were employed. Model performance was evaluated using appropriate regression metrics. The analysis provides insights into the key determinants of housing prices in the dataset and assesses the effectiveness of the predictive models.

**I. INTRODUCTION**

This project utilizes an Indian housing dataset to explore the factors influencing residential property prices. The dataset includes features such as the area of the property, the number of bedrooms and bathrooms, and the presence of amenities like a main road, guest room, and air conditioning. Understanding the dynamics of housing prices is crucial for buyers, sellers, and real estate stakeholders in the Indian market.

This report details the analysis of the Indian housing dataset, including exploratory data analysis, data preparation, and the implementation and evaluation of regression models to predict housing prices. The goal is to identify the key features that drive property values and assess the predictive power of the models.

**II. BACKGROUND**

The Indian housing market is influenced by a variety of factors, including urbanization, population density, economic growth, and the availability of infrastructure and amenities. These factors can significantly impact the pricing of residential properties. Understanding the relationship between housing features and prices is essential for market analysis and informed decision-making within the real estate sector. The dataset used in this analysis captures several of these relevant features.

**III. EXPLORATORY ANALYSIS**

The Indian housing dataset contains 12 columns of samples with 545 rows of features. The data types of the columns are shown in Table 1.

**Table 1: Indian Dataset Data Types**

|  |  |
| --- | --- |
| **Variable Name** | **Data Type** |
| price | int64 |
| area | int64 |
| bedrooms | int64 |
| bathrooms | int64 |
| stories | object |
| mainroad | object |
| guestroom | object |
| basement | object |
| hotwaterheating | object |
| airconditioning | object |

**IV. METHODS**

**A. Data Preparation**

The data preparation process involved several steps. Categorical variables, such as 'mainroad', 'guestroom', 'basement', 'hotwaterheating', and 'airconditioning', were converted into numerical representations using one-hot encoding. This allows these categorical features to be used in the regression models.

**B. Experimental Design**

The following table outlines the experimental design used to evaluate the models.

|  |  |
| --- | --- |
| Experiment Number | Parameters |
| 1 | All features, 80/20 train/test split, Linear Regression |
| 2 | Selected features (area, bedrooms, bathrooms), 70/30 train/test split, Linear Regression |
| 3 | All features, 80/20 train/test split, Random Forest Regressor |
| 4 | Selected features (area, bedrooms, bathrooms), 70/30 train/test split, Gradient Boosting Regressor |

**C. Tools Used**

The following tools were used for this analysis: Python was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn. Pandas was used for data manipulation and analysis, NumPy for numerical computations, Matplotlib and Seaborn for data visualization, and Scikit-learn for implementing the machine learning models. These tools were chosen for their effectiveness in handling data, performing complex computations, and providing robust machine learning functionalities.

**V. RESULTS**

**A. Classification Measures/Accuracy Measure**

1. Confusion Matrix: This table shows the true positives, true negatives, false positives, and false negatives, providing insight into the model's classification accuracy.
2. ROC Curve: The ROC curve plots the true positive rate against the false positive rate, with the area under the curve (AUC) indicating the model's ability to distinguish between classes.
3. RMSE (Root Mean Square Error): For regression models, RMSE measures the average magnitude of prediction errors. Lower RMSE values indicate better model performance.
4. Accuracy: The proportion of correctly classified instances out of the total instances.
5. Precision and Recall: Precision measures the proportion of true positive predictions out of all positive predictions, while recall measures the proportion of true positive predictions out of all actual positives.

**B. Discussion of Results**

1. Best Model: Identify the model with the highest accuracy or lowest RMSE. Discuss why this model performed well, considering factors like feature selection and data split.
2. Worst Model: Identify the model with the lowest performance metrics. Discuss potential reasons for its poor performance, such as overfitting or inappropriate feature selection.
3. Comparison: Compare the results of different models and experiments, highlighting strengths and weaknesses.

**C. Problems Encountered**

1. Data Collection: Issues with obtaining a complete and clean dataset.
2. Data Preparation: Handling missing values, outliers, and data normalization.
3. Model Implementation: Problems with model convergence, parameter tuning, and computational limitations.
4. Evaluation: Challenges in interpreting results and selecting appropriate metrics.

**D. Limitations of Implementation**

Address the limitations of your models, considering:

1. Model Assumptions: Discuss any assumptions made by the models that may not hold true for the dataset.
2. Feature Selection: Limitations due to the choice of features and their impact on model performance.
3. Data Split: The impact of train-test split ratios on model generalization.
4. Model Complexity: Trade-offs between model complexity and interpretability.

**E. Improvements/Future Work**

1. More Experiments: Conduct additional experiments with different models and parameters.
2. Different Models: Explore other machine learning algorithms that might perform better.
3. Feature Engineering: Create new features or remove irrelevant ones to improve model performance.
4. Alternative Datasets: Use different datasets to validate the model's robustness and generalizability.

**VI. CONCLUSION**

This project analyzed an Indian housing dataset to understand and predict housing prices. Exploratory data analysis provided insights into the characteristics of the data, and regression models were used to predict prices based on various features. The results of the model evaluation highlight the importance of certain features in determining property values. Future work could focus on refining the models and incorporating additional data to improve useful accuracy.