House Rent Prediction System

Sai Teja Patibandla  
 U00876137  
 [sptbndla@memphis.edu](mailto:sptbndla@memphis.edu)

Naveen Mogallapu  
 U00876137  
 [sptbndla@memphis.edu](mailto:sptbndla@memphis.edu)

ABSTRACT

The House Rent Prediction System project is an innovative endeavor in the real estate domain, focusing on the accurate forecasting of house rents using advanced computational models. This initiative employs a variety of predictive algorithms, including Random Forest Regression, Lasso Regression, and Neural Networks, to analyze a dataset rich in real estate features such as location, size, and amenities. The goal is to provide reliable rent predictions, assisting individuals and investors in navigating the complex real estate market. The project showcases the effective use of these algorithms in generating pertinent rent forecasts, a critical tool for financial planning and investment decisions in the ever-changing landscape of real estate.

**INTRODUCTION**

The House Rent Prediction System project is a comprehensive effort to address the escalating issue of housing rent prices, a concern that is becoming increasingly significant in today's economic climate. This system aims to provide accurate forecasting of housing values, which is crucial for both the general public and developers when making investment decisions. The project's central focus is on developing a predictive model that can analyze various factors influencing house rents such as location, size, amenities, and broader economic factors.

The importance of this problem stems from the direct impact of house rent prices on the economy and individual financial planning. Accurate rent prediction aids in better decision-making for potential investors and individuals looking for housing, and it also serves as a barometer for the health of the real estate market.

One of the primary challenges in this project is the complexity of predicting house rents, which involves numerous variables. This complexity necessitates the implementation of multiple prediction models and the rigorous assessment of their accuracy. The project utilizes a variety of models such as Random Forest Regression, Lasso Regression, Extreme Gradient Boosting (XGBoost), Neural Networks, Support Vector Regression (SVR), K-Nearest Neighbors (KNN), and Gaussian Naive Bayes, each evaluated using metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

A comprehensive Graphical User Interface (GUI) is also developed for easy user interaction and rent prediction. The project leverages a dataset from Kaggle, which includes essential features for rent prediction such as BHK, bathrooms, and city. This dataset has no missing values, simplifying the prediction process, though preprocessing is still required for handling outliers and conducting exploratory data analysis.

A diagram of a process

Description automatically generated

**Figure 1: Project Workflow**

**PROPOSED APPROACH**

The House Rent Prediction System incorporates a multifaceted approach by employing a suite of advanced computational models. These include Random Forest Regression, Lasso Regression, Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), Gaussian Naive Bayes, and Support Vector Regression (SVR). Each of these techniques plays a crucial role in the system's ability to accurately predict house rents.

**Random Forest Regression (RFR):**

RFR is employed for its ability to handle a large number of features and its robustness against overfitting. It works by constructing multiple decision trees during training and outputting the average prediction of the individual trees.

**Lasso Regression:**

This technique is useful for feature selection and regularization to improve the predictability of the model.

Lasso regression reduces the complexity of the model by shrinking some coefficients to zero, essentially performing feature elimination.

The lambda value, which determines the amount of shrinkage, is a critical parameter tuned during the model building process.

**XGBoost (Extreme Gradient Boosting):**

XGBoost is a scalable and high-performance gradient boosting framework used for regression tasks. It is chosen for its efficiency, flexibility, and portability.

**K-Nearest Neighbors (KNN):**

KNN is used for its simplicity and effectiveness in regression tasks. It predicts the rent by averaging the rents of the 'K' nearest neighbors in the feature space.

The key parameter, 'K' (the number of neighbors), is tuned for optimal performance. The selection of 'K' is crucial as a small value can lead to high variance, while a large value may smooth out predictions too much.

**Gaussian Naive Bayes:**

This probabilistic classifier assumes independence among predictors. It’s used for its efficiency and simplicity.

Gaussian Naive Bayes is applied with the assumption that the continuous values associated with each feature are distributed according to a Gaussian distribution.

**Support Vector Regression (SVR):**

SVR is included for its ability to minimize errors, making it robust to outliers and effective in capturing complex relationships in the data.

**Assumptions**

The foundation of the House Rent Prediction System is based on several critical assumptions. Firstly, the dataset is believed to be comprehensive and accurately representative of the dynamics of the real estate market. This encompasses a broad range of factors influencing house rent prices. Additionally, it is assumed that the selected features, including BHK, bathrooms, city, and floor size, have a significant impact on house rent prices. Furthermore, the dataset is presumed to be clean, characterized by the absence of missing values and a minimal presence of outliers. These assumptions are pivotal in guiding the subsequent steps in the model's development.

**Feature Selection**

Feature selection is a critical process that involves identifying the most relevant variables to use in model construction. In this project, feature selection was approached through:

1.Exploratory Data Analysis (EDA): Used to understand data distribution and relationships, employing plots and charts to identify key features impacting rent prices.

2.Correlation Analysis: Employed correlation matrices to find features strongly linked to rent prices, prioritizing those with high correlation.

3.Domain Knowledge: Leveraged insights into the real estate market to identify crucial factors like location, size, bedrooms, and amenities.

**Parameter Tuning**

Parameter tuning is a crucial phase in the development of the House Rent Prediction System. This process involves using techniques such as grid search and cross-validation to determine the optimal set of parameters for each model. For instance, The Neural Network, implemented using MLPRegressor, is configured with a hidden layer structure of (64, 32) neurons and a maximum iteration limit of 500. The K-Nearest Neighbors (KNN) model is set up with the number of neighbors parameter (n\_neighbors) fixed at 5.

**Handling Inconsistency/Skewness in Datasets**

To ensure the reliability and accuracy of the House Rent Prediction System, special attention is given to handling inconsistencies and skewness in the dataset. Outliers are identified using statistical methods like box plots and are treated carefully to minimize their impact on the model's performance. Additionally, data normalization techniques are employed to address any skewness present in the dataset. This is complemented by splitting the dataset into training and testing sets, which helps ensure the model's generalizability and guards against overfitting. These practices are essential in maintaining the integrity and quality of the data fed into the models.

**Evaluation Metrics**

The performance of the models in the House Rent Prediction System is rigorously evaluated using several key metrics. These include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE). MAE and RMSE provide insights into the accuracy of the models in terms of the average magnitude of errors in predictions. These metrics are crucial in assessing the effectiveness of the models and in guiding further refinements to enhance their predictive capabilities.

**EXPERIMENTATION AND RESULTS**

The empirical experiments conducted for the House Rent Prediction System not only involve the analysis of models but also the integration of a graphical user interface (GUI) to facilitate user interaction. The GUI, allows users to select parameters such as city, BHK, and bathroom count to predict average rent prices.

From the figure-1, we can observe that the Random Forest model exhibits the highest training score, which suggests that the model has a good fit to the training data. However, the testing score is significantly lower, indicating a potential overfitting issue where the model does not generalize well to unseen data. The Neural Network model also shows a strong training score, but a drop in testing score, again suggesting overfitting. Lasso Regression, on the other hand, maintains a relatively closer performance between training and testing, indicating better generalization.

Figure-2 presents MAE and RMSE values for each model, where lower values indicate better performance. The Random Forest Regressor models show relatively lower errors, which could mean that it is better at handling the variance in the dataset and may provide more reliable predictions.

These insights suggest that while more complex models like Random Forest and Neural Networks can capture the training data well, they may not always provide the best generalization to new data. Simpler models or those with inherent regularization, such as Lasso Regression, could offer a better balance between training performance and generalizability.

A graph of blue rectangular bars

Description automatically generated

**Figure 2: Model comparison on Training and Testing scores**

A graph of blue and black bars

Description automatically generated with medium confidence

**Figure 3: Comparison of models on evaluation metrics**

**Setup and Baseline Method**

The setup for the project is executed within a Jupyter Notebook environment, employing Python's pandas library for data manipulation and preparation. This robust setup allows for an interactive and iterative approach to analyze and visualize the data. The baseline method for the project provides a fundamental comparison point for predictive performance and is established using a simple model, such as a basic machine learning algorithm with default parameters. This baseline serves as the control against which the performance of more sophisticated predictive models can be measured.

**CONCLUSION**

In this project, the challenges associated with accurate house rent prediction were addressed through the development of a comprehensive prediction system. Leveraging a variety of models, including Random Forest, Gradient Boosting, and Neural Networks, among others, the system aimed to predict house rents based on a multitude of factors such as location, size, and amenities.

The models were rigorously tested using a dataset with Python's pandas library, with their performance assessed on key metrics like Mean Absolute Error and Root Mean Squared Error. The insights gained revealed that while complex models could overfit, simpler models or those with regularization techniques tended to generalize better.

**FUTURE WORK**

Future work for this project will focus on refining the models to further enhance their predictive accuracy and generalizability. This includes exploring additional feature engineering techniques, experimenting with model hyperparameter tuning, and extending the GUI for added functionality and user experience improvements. The potential for real-world application remains a strong aspect of this project, aiming to assist individuals and investors in making informed decisions in the real estate market.

REFERENCES

[1] K. Chanasit, E. Chuangsuwanich, A. Suchato and P. Punyabukkana, "A Real Estate Valuation Model Using Boosted Feature Selection," in IEEE Access, vol. 9, pp. 86938-86953, 2021, doi: 10.1109/ACCESS.2021.3089198.

[2] J. Bin et al., "Regression model for appraisal of real estate using recurrent neural network and boosting tree," 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), Beijing, China, 2017, pp. 209-213, doi: 10.1109/CIAPP.2017.8167209.

[3] N. N. Ghosalkar and S. N. Dhage, "Real Estate Value Prediction Using Linear Regression," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697639.

[4] R Manjula, Shubham Jain, Sharad Srivastava and Pranav Rajiv Kher, “Real estate value prediction using multivariate regression models,” IOP Conference Series: Materials Science and Engineering, 2017.

[5] Nihar Bhagat, Ankit Mohorkar and Shreyas Mane, “House Price Forecasting using Data Mining,” International Journal of Computer Applications, 2016.

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00