## Artificial Intelligence

## and

## Machine Learning

Project Report

Semester-IV (Batch-2022)

House Rental Price Predictor

A red and white sign

Description automatically generated with low confidence

**Supervised By: Submitted By:**

Faculty Name : Shubham Singhal Group: G11

Manan Jain 2210990551

Manraj Singh 2210990560

Mayank Bisht 2210990566

Mayank Gupta 2210990567

**Department of Computer Science and Engineering**

## Chitkara+ University Institute of Engineering & Technology,

## Chitkara University, Punjab

**Abstract**

The rapid growth of the real estate market has led to surge in house rental prices, making it challenging for potential tenants to estimate fair rental prices. To address this issue, we propose a machine learning based House Rental Price Predictor model. This model leverages historical rental price data and various house features such as the number of bedrooms, bathrooms, total area, location, proximity to amenities and more. Through rigorous preprocessing and feature extraction, we extract meaningful features from the rental data. We then train and evaluate several machine learning models to classify houses based on their rental prices. Our experimental results demonstrate the effectiveness of our approach in accurately predicting house rental prices, achieving high levels of precision, recall, and overall regression accuracy. This project showcases the potential of machine learning in revolutionizing the real estate industry by providing quick, accurate, and automated rental price.

1. **Introduction**

In the contemporary era of rapid urbanization and digital connectivity, the challenge of accurately predicting house rental prices has emerged as a significant concern in the real estate industry. House rental prices, determined by a multitude of factors such as location, size, proximity to amenities, and market trends, have a substantial impact on both tenants and landlords. Tenants often grapple with finding accommodations that fit their budget and meet their needs, while landlords strive to set competitive rental prices that also yield a reasonable return on their investment. The advent and subsequent proliferation of online real estate platforms have both simplified and complicated this process. While these platforms have made it easier for users to find rental properties, they have also led to information overload, making it difficult for users to discern a fair rental price. In this context, the development of effective mechanisms to predict and analyze house rental prices has become crucial. Our project aims to address this issue by leveraging machine learning techniques to develop a House Rental Price Predictor, providing users with accurate and instant rental price estimates and thereby aiding in informed decision-making.

**1.1 Background:**

The advent of the internet and digital platforms has transformed the real estate industry, making it easier for individuals to find rental properties. However, determining a fair rental price remains a complex task due to the multitude of factors influencing it, such as location, size, proximity to amenities, and market trends. Traditional methods of price estimation often rely on real estate agents or manual comparisons, which can be time-consuming and subjective.

Rental prices significantly impact both tenants and landlords. Tenants seek to find accommodations that fit their budget and meet their needs, while landlords aim to set competitive rental prices that yield a reasonable return on their investment. Mispriced rentals can lead to prolonged vacancies or financial losses for landlords, and financial strain or missed opportunities for tenants.

**1.2 Objectives:**

The primary objective of the House Rental Price Predictor project is to develop an effective rental price prediction system that can accurately estimate the rental price of a house based on various influencing factors. By leveraging machine learning algorithms and data analysis techniques, the system aims to analyze the features of rental properties and identify patterns indicative of rental prices.

* 1. **Significance:**

The House Rental Price Predictor model holds significant value in the contemporary real estate landscape, characterized by fluctuating market trends and diverse property options. The significance of this model lies in its potential to transform the rental market by providing accurate and instant rental price estimates, thereby fostering transparency, fairness, and efficiency in rental transactions.

Access to accurate and reliable rental price estimates is crucial for both tenants and landlords. For tenants, it aids in budget planning, ensures they are getting a fair deal, and helps them make informed decisions when choosing a rental property. For landlords, it assists in setting competitive rental prices that yield a reasonable return on their investment, ensuring their properties do not remain vacant for extended periods due to overpricing.

Moreover, the House Rental Price Predictor model plays a vital role in promoting financial literacy and informed decision-making in the real estate market. By providing insights into the factors influencing rental prices and their relative importance, we empower individuals to navigate the complex real estate landscape more effectively.

Furthermore, this model can contribute to the broader economy by aiding in policy-making and urban planning. Understanding rental price trends can provide valuable insights into housing affordability, population distribution, and urban development patterns, informing policies and strategies in these areas.

Problem Definition:

The rapid urbanization and fluctuating real estate markets pose a significant challenge to individuals and families looking for rental properties. The problem is to develop an effective house rental price prediction system capable of accurately estimating the rental price of houses based on various factors such as location, size, number of rooms, proximity to amenities, and other relevant features. This will aid potential tenants in making informed decisions, ensuring fair pricing, and enhancing the transparency of the rental market.

1. **Requirements:**

**3.1 Software:**

1. Programming Language: Python (version 3.x)
2. Machine Learning Libraries:
   * Scikit-learn
   * Pandas and NumPy for data manipulation
3. Development Environment: Jupyter Notebook or any IDE (PyCharm, VS Code, etc.)
4. Version Control: Git and GitHub (or any other version control system)
5. Visualization Libraries:
   * Matplotlib and Seaborn for data visualization
   * Plotly or Bokeh for interactive visualizations

**3.2 Hardware Requirements:**

1. Processor: Multi-core processor (Intel Core i5 or equivalent)
2. RAM: 8 GB or higher recommended
3. Storage: At least 100 GB of free disk space for storing datasets, models, and related files

**3.3 Dataset:**

https://www.kaggle.com/datasets/amitabhajoy/bengaluru-house-price-data?resource=download

**4.Methodology:**

1. **Data Collection:**
   * Gather a diverse dataset containing both genuine news articles and fake news articles from reputable sources and datasets.
   * Organize the dataset into separate directories for genuine and fake news articles.
2. **Feature Extraction:**
   * Optionally, include metadata features such as publication date, author information, and user engagement metrics.
3. **Model Selection:**
   * Experiment with various machine learning and deep learning algorithms available in scikit-learn.
   * Choose the model that achieves the best performance on validation data.
4. **Model Training:**
   * Split the dataset into training, validation, and test sets using scikit-learn.
   * Train the selected model using the training data.
   * Tune hyperparameters using techniques like grid search or random search provided by scikit-learn.
5. **Result:**

Some metrics on which Machine Learning project is tested:

1. **Model Performance Metrics:**

Accuracy: the proportion of correctly classified instances.

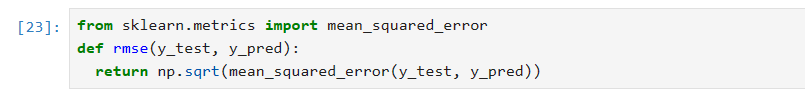
Precision: the proportion of true positive predictions among all positive predictions.

Recall: the proportion of true positive predictions among all actual positives.

F1-score: the harmonic mean of precision and recall.

ROC-AUC (Receiver Operating Characteristic - Area Under the Curve): measures the ability of the model to distinguish between classes.

Provide the values of these metrics for each model evaluated on the testing dataset.



1. **Comparison of Models:**

Compare the performance of different models based on the evaluation metrics.

Discuss which model(s) performed the best and why.

Highlight any trade-offs between models, such as computational complexity, interpretability, and predictive power.

1. **Feature Importance:**

If applicable, discuss the importance of features in predicting the target variable.

Present the feature importance scores generated by the model(s).

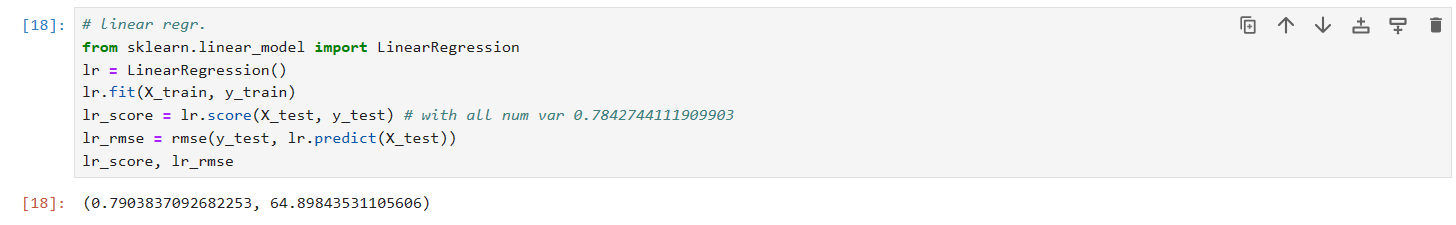
Interpret the significance of each feature in the context of the problem domain.

Identify any features that have a strong impact on the model's predictions.

After cleaning of model different models were used on data and using metrics, we tried to decide which works best on our data.

1. **Linear Regression:**

Linear regression is a statistical method used to model the relationship between a dependent variable (target variable) and one or more independent variables (predictor variables). It is a simple yet powerful technique commonly employed in various fields such as economics, finance, social sciences, and machine learning.



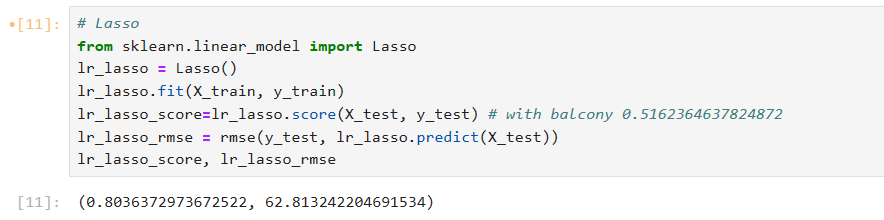
Here we can clearly see this model has: -

**Accuracy**: 79%

**RMSE**: 64.89

1. **Lasso:**

Lasso, which stands for Least Absolute Shrinkage and Selection Operator, is a regression technique used for variable selection and regularization in linear models. It is particularly useful when dealing with datasets that have a large number of features (high-dimensional data) and potentially collinear predictors.



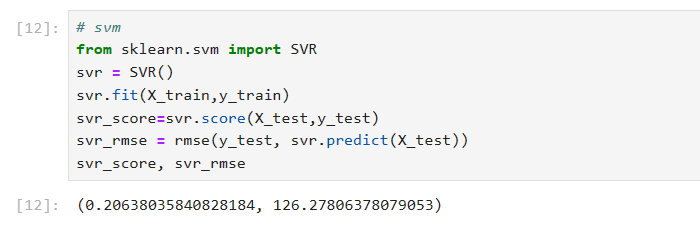
Here we can clearly see this model has: -

**Accuracy**: 80%

**RMSE**: 62.81

1. **Support Vector Machine (SVM):**

Support Vector Machines (SVM) is a supervised learning algorithm used for classification, regression, and outlier detection tasks. It is particularly effective in high-dimensional spaces and is widely used in machine learning for various applications, including text classification, image recognition, and bioinformatics.



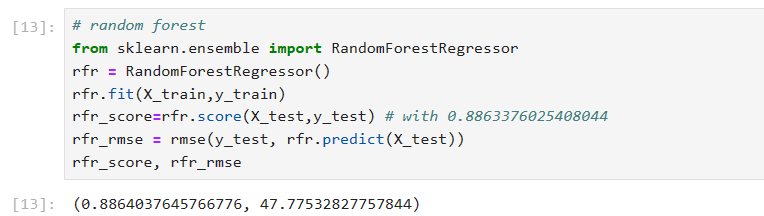
Here we can clearly see this model has: -

**Accuracy**: 20%

**RMSE**: 126.27

1. **Random Forest Regressor:**

Random Forest Regressor is an ensemble learning method used for regression tasks. It belongs to the family of decision tree-based algorithms and is widely used due to its flexibility, robustness, and ability to handle complex data

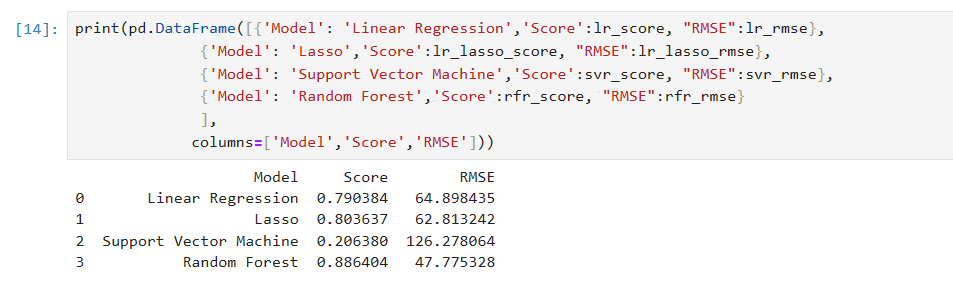


Here we can clearly see this model has: -

**Accuracy**: 88%

**RMSE**: 47.77

**Results of all Models:**

****

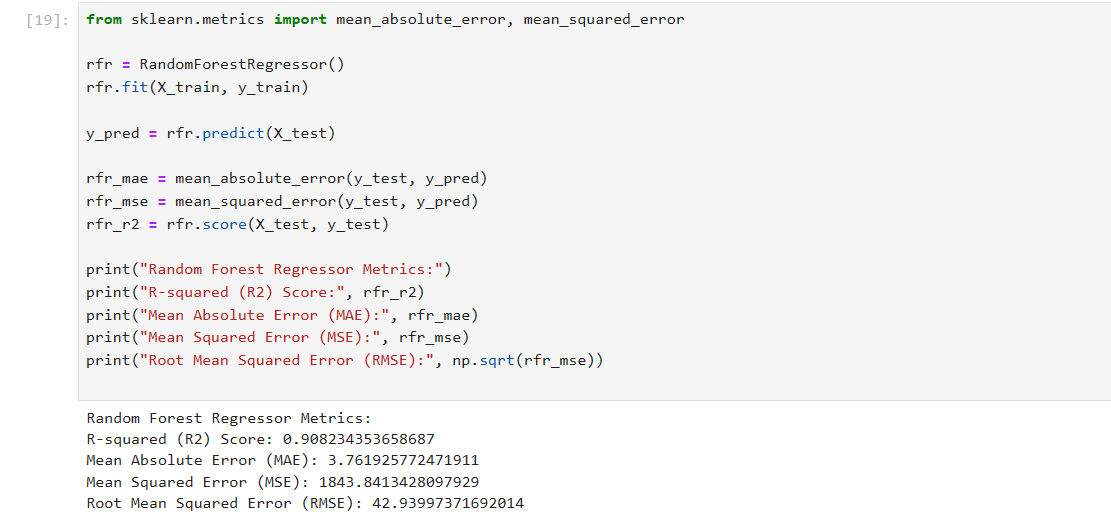
Based on the results obtained from the evaluation of different regression models, including Linear Regression, Lasso Regression, Support Vector Machine (SVM), and Random Forest, it is evident that Random Forest outperforms the other models in terms of both predictive performance and model accuracy

**Random Forest** achieves the highest score among all models, with a score of **0.886404**, indicating its superior ability to capture the underlying patterns and relationships in the data. Additionally, it has the lowest **Root Mean Squared Error (RMSE) of 47.775328**, signifying its effectiveness in making accurate predictions.

Random Forest provides valuable insights into feature importance, allowing us to identify the most influential variables driving the predictions. This can aid in feature selection, model interpretation, and identifying key factors impacting the target variable.

Based on these factors, **Random Forest** emerges as the preferred choice for the regression task at hand. Its combination of high predictive performance, robustness, and ability to handlecomplex data makes it a suitable and reliable model for making accurate predictions in real-world scenarios.

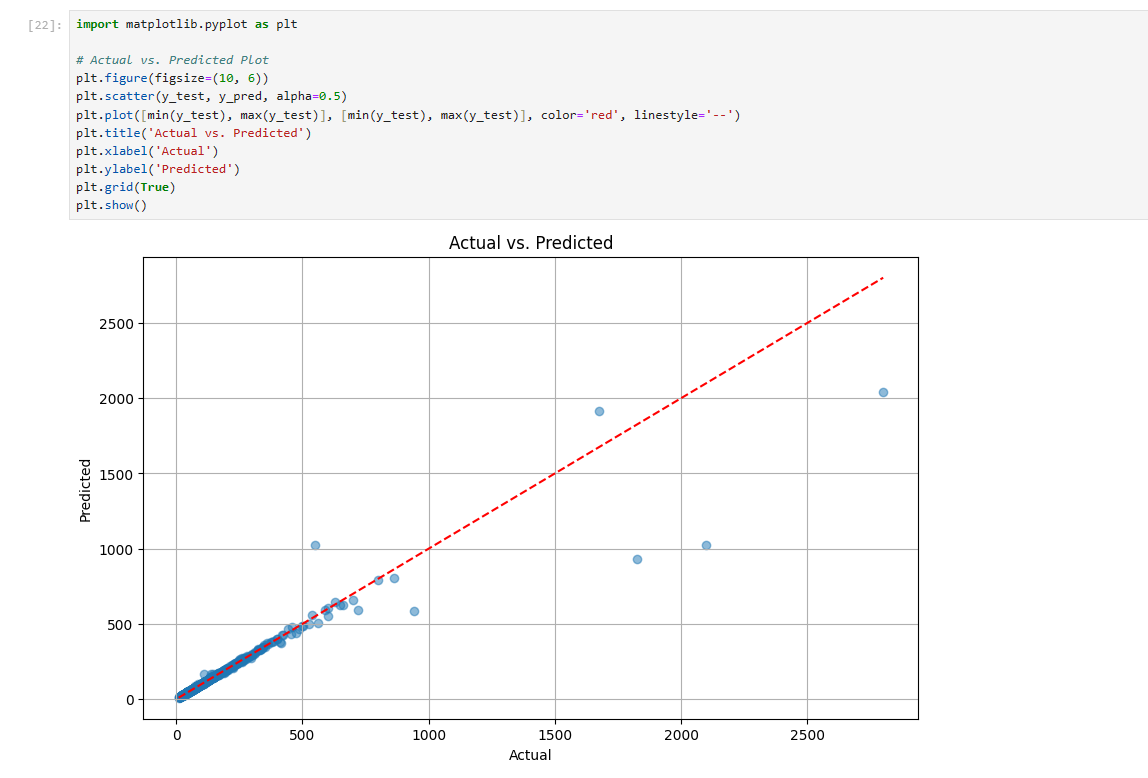
**Some more metrics evaluated on Random Forest:**

****

**Visual Representation of Random Forest Model**

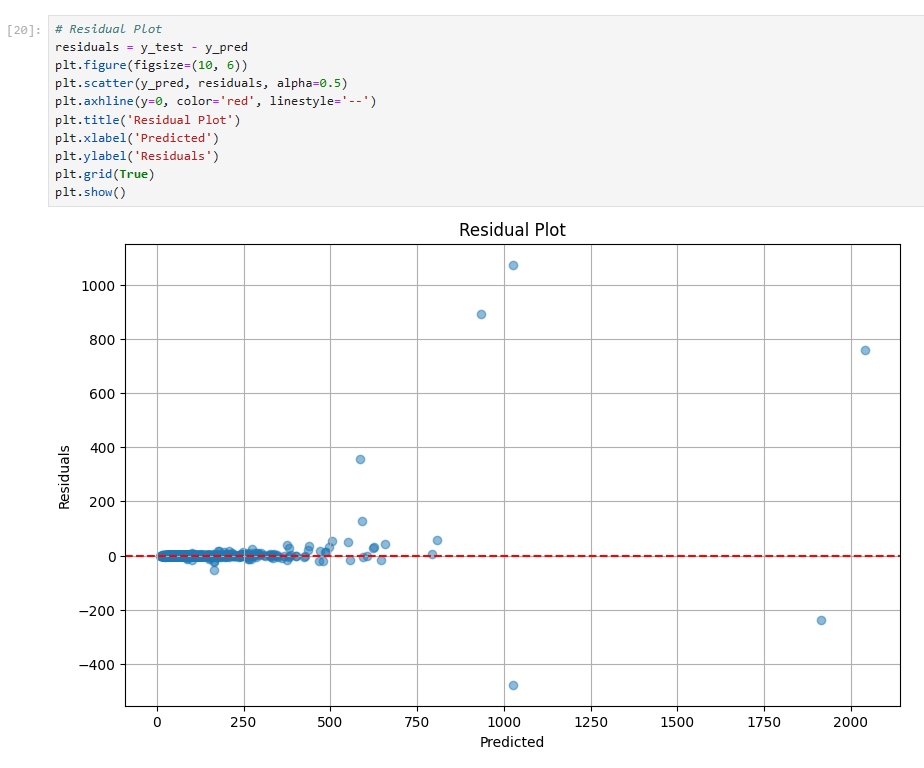
* **Actual Vs Predicted Plot:**

An Actual vs. Predicted Plot is a visualization used to compare the actual values of the target variable with the predicted values obtained from a machine learning model. It provides a visual assessment of how well the model's predictions align with the ground truth.

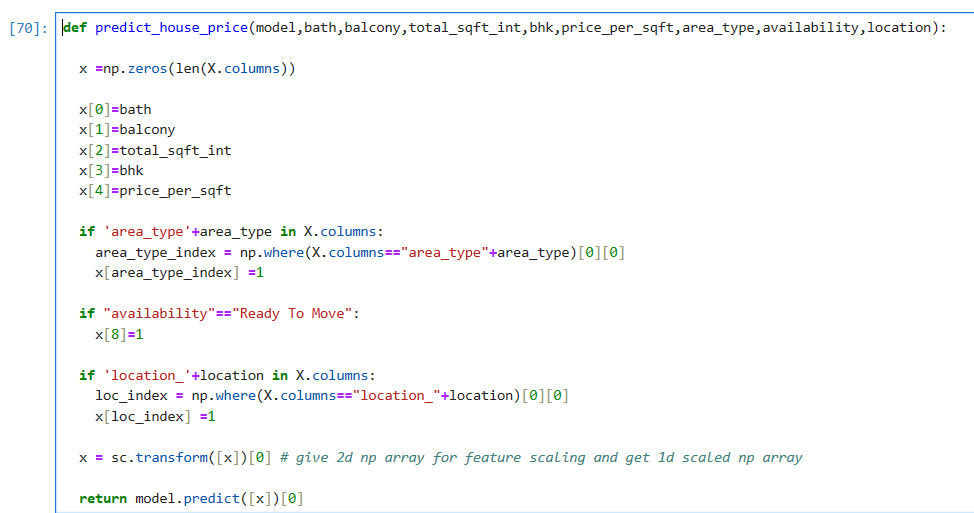
****

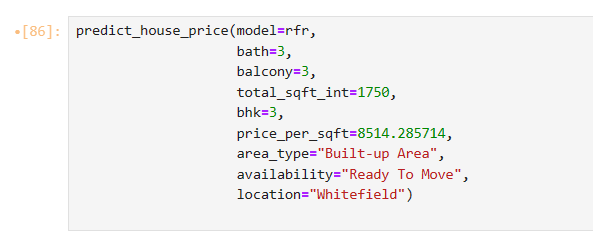
* **Residual Plot:**

A Residual Plot is a graphical tool used to analyze the residuals, or the differences between the observed (actual) values and the predicted values from a regression model. It helps in assessing the goodness of fit of the model and identifying patterns or trends in the residuals.



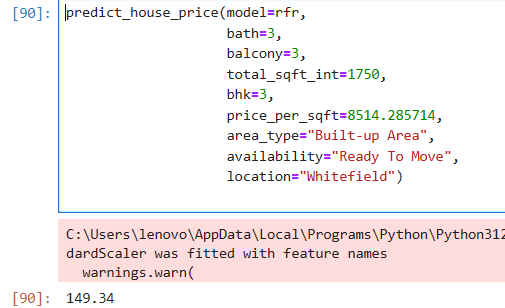
**Final Prediction**

****

****

The provided code defines a function called (predict\_house\_price) to predict the price of a house using a Random Forest Regressor model. This function takes several input parameters representing features of the house, such as the number of bathrooms, balconies, total square footage, number of bedrooms, price per square foot, area type, availability status, and location. These features are used to create a feature vector, which is then scaled using a scaler. The scaled feature vector is passed to the model to obtain the predicted house price, which is returned by the function. The function can be called with specific feature values to predict the price of a house based on its characteristics.

Here we can choose features and their input according to our need and we get a number which shows the price of property in lacs.



Here 149.34 shows the price for selected features in lacs.