REASONABLE HOUSE PRICE PREDICTION

EG/2020/4212 , Sewwandi L.L.C. EG/2020/4214 , Shamali A.V

Abstract

The application of machine learning algorithms which are Linear Regression and Decision Trees, to predict rental house prices based on relevant features are explored by this project. In response to the growing demand for accurate real estate market predictions, the project is evaluated the performance and accuracy of these algorithms. Various housing features, including location, bedrooms, bathrooms and amenities are in data set. After preprocessing the data, the models are trained and tested, and their predictive capabilities are assessed using metrics such as Mean Absolute Error and Root Mean Squared Error. The results offer valuable insights into the strengths and weaknesses of each algorithm chosen the most suitable approach for rental house price prediction.

I. INTRODUCTION

The problem of finding the rental price for houses has been difficult for landlords and tenants in practical scenario. So, given the challenges inherent in this process, there is a smart computer program to build and improve the decision-making process related to house rental price. The main goal is creating a prediction model that can determine the perfect rent price for a home based on a range of characteristics. The goal is developing a system that uses machine learning techniques to guaranteed fairness for both tenants and landlords and streamlining the price determination process.

According to our project, the main object is predict house rental price by using machine learning algorithms. For that there are two regression type algorithms are used. Because this problem is supervised regression type problem happening in the world. Creating the important model for predict the rent is very important thing in the real market. So, from this project mainly address to the huge issues and problem which faced by the landlords. In addition, it provides good and clarity in the rental pricing system. The machine learning approach identifies patterns from the past experience data to provide accurate predictions for unseen data.

The experiential component of the project requires the processing of a large data set containing various information about several features caused for the rent and their related rental costs. The machine learning model is trained using historical data to gain insight into the complex aspects that influence rental pricing. As a result of this learning process, the model is expected to improve its ability to accurately predict houses that are not in the training dataset. Performance is measured using some matrixes, such as mean squared error (MSE), which quantifies the degree of agreement between predicted and actual rental prices, will be used to assess the effectiveness of the prediction model. A lower MSE value indicates a higher accuracy of house rent prediction. The program demonstrates success in producing accurate and fair rental price predictions. The project essentially uses machine learning to improve and simplify the complex process for find the rental prices. Actually, by implementing this result it can be used as a solution for real-world problem.

II. METHODOLOGY

A. *Data:*

Source of Data: The dataset for "House Rent Prediction" was obtained from the Kaggle website and below one is the web link for dataset.

Web Link:
House Rent Prediction Dataset (kaggle.com)

Type of Data and Features: The dataset consist of several information which are related with available rental houses. There are 12 features are included in the dataset. They are "Posted On," “number of bedrooms,” "Rent," "Size," "Floor," "Area Type," "Area Locality," "City," "Furnishing Status," "Tenant Preferred," "Bathroom," and "Point of Contact."

In this dataset there are 4746 observations are collected and all they have unique feature variable set.

Output Variable: The target variable is Rental price for the houses.

B. *Pre-processing:*

Pre-processing steps are applied to ensure the efficiency and accuracy of the dataset. At first null values and duplicates were checked but there were no any null values and duplicates. Two columns were removed by assuming those are not cause to change the prediction price. To convert the categorical features into numerical features the label encoding is suitable for machine learning algorithms. Then the outliers were identified

and removed to enhance model robustness. The dataset was split into training and testing data, The aim is this pre-processing is prepare a practicable data model to train and by using that model predict the reasonable rental price from this machine learning project.

From the full dataset there is 80% of the data was got for the training stage. Then the other 20% is for testing.

By applying these methods the data can be more accruable and they avoid the overfitting and under fitting. Finally all the data can be reduced the sensitivity to outliers and system is stable.

C. Algorithm:

Both Linear Regression and Decision Tree algorithms are supervised learning methods. Linear Regression is a parametric method, features for predict rental house price dataset has linear relationship with the output variable. So, this linear regression algorithm is used. As well as, the decision trees is a non-parametric method. In the dataset, there are eight features are categorical features out of twelve. So, decision tree is used because it can handle both categorical and numerical variables at a time. However, both these algorithms has power to create a best model to predict the rental price reasonably.

Linear Regression:
Final prediction of this project is rent price. So, this Linear Regression is important to use for this machine learning project. As well as there is a linear relationship between all the features with rent price. Therefore by applying this algorithm prediction can be more accruable.

Decision Trees:
Mainly this decision tree algorithm is used for dataset which has non-linear relationships and complex decision boundaries. In the dataset there are more categorical features other than numerical features. Applying decision tree algorithm rental price can be predict according to selected features.

D. Implementation:

The main goal of this project is implement the ideal house rental price by importing some libraries. Some of them are numpy, pandas, Scikit-Learn library etc. The numpy library is provided some mathematical functions and those are used for normalization and standardization processers under the pre-processing stage. The panda library is mainly used for data handling and data cleaning tasks.
Using the Decision Trees technique, which was selected for its ability to accurately capture complex relationships in the dataset, was an essential aspect of our strategy.
We used Grid Search Cross Validation to perform a thorough hyper parameter tuning process during the implementation. The objective of this methodical investigation of several configurations was to improve the models' forecast accuracy and fine-tune them. Significantly, a reduced mean absolute error and root mean square error showed that the Decision Tree model performed better than the Linear Regression model. These metrics show the Decision Tree's effectiveness in predicting property rental prices and act as quantitative indications of the models' correctness.

III. RESULTS

- Linear Regression Algorithm:

Absolute Error: 24185.536
Root Mean Square Error: 43666.1206

- Decision Tree Algorithm:

Absolute Error: 14451.898
Root Mean Square Error: 37354.549

To evaluate predictive performance the results were obtained using Linear Regression and Decision Tree algorithms. The absolute error is measured at 24,185.536, it is indicating with the average magnitude of differences between predicted and actual prices. The Root Mean Square Error for Linear Regression is measured overall prediction accuracy and the value for that is 43,666.1206. By comparing these two error values for both Linear Regression and Decision Tree model, the lower errors are predicted from Decision Tree model. The calculated MAE is 14,451.898 and RMSE is 37,354.549. So, Decision tree model has super predictive capability to predict rent house. Because of the lower error values, the outcome of the Decision Tree model is more accurate and effective for get the reasonable house rent price.

- Models of Evaluation:

In summary, the Decision tree model is performed very reasonably to predict the house rent price by considering the complexity of the data set with more accurately.

Problems with Overfitting:

In machine learning, overfitting is caused for high variance, poor generalization, and noise sensitivity. When applied to new data, too complicated models may perform worse since they capture noise rather than underlying trends. Standardization, cross-validation, feature selection, gathering more data model architecture are useful technique to deal with overfitting. These steps are intended to build more resilient models that perform well in generalization and generate trustworthy predictions on unknown data.

- Tuning the Models :

Grid Search Cross-validation for hyper parameter Tuning:

To summarize, a thorough grid search using the `param_grid` was part of the Decision Tree model's hyperparameter tuning procedure. Important hyperparameters including {max_depth}, `min_samples_split}, and `min_samples_leaf} were explored. The goal of applying grid search cross validation into decision tree model is improve the training model accuracy and predict the rent price reasonably.
When the tree is too deep, it leads to overfitting the system. When this model creation there are some hyperparameters were implemented within some range.

They are,
maximum depth of tree
minimum number of samples,
minimum number of samples required to be at a leaf node.

The main purpose is evaluate the model's performance across different subsets of training data and select the best hyper parameters for generate the unseen data very accurately.

- Conclusions drawn from the Data

The inferences used for highlight the data which improve the results attained by utilizing the Decision Tree and Linear Regression algorithms in the machine learning model. Both these algorithms are implemented the usefulness for real world applications by successfully projecting rental values for houses depending on many features. As well as, the there are some features should be improved. Some experimental datasets are lead to improved generalization and performance on unseen data during the train and testing, the necessity for a larger and more balanced dataset is highlighted. Overfitting issues are prioritized, demonstrating an understanding of the significance of model resilience. Furthermore, the recommendation to optimize the support vector analysis raises the possibility of another direction for improvement and begs investigation.
In the whole dataset there are some data sets are in out of range that we consider for create the model? So, to get the beneficial outcome remove those is a good step. However, the final conclusion drawn from data is that

- Data Visualization

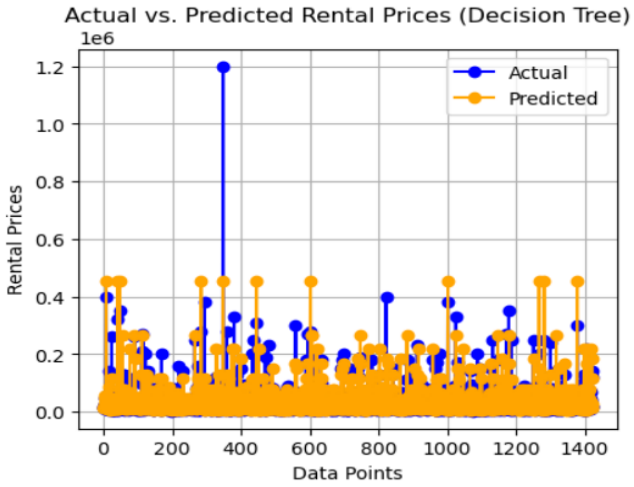

Figure 1.Actual price vs Predicted price for linear regression model

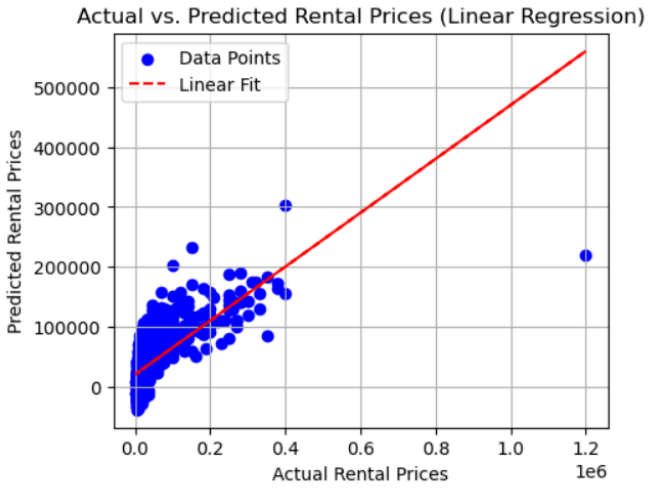

Figure 2. Actual price vs Predicted price for decision tree model

IV. Discussion

Ethical Considerations:

Ethics and accuracy measurement is more considerable fact in the machine learning projects. Because implementing a new thing for the society is more dangerous and responsible process. So, this responsibility is depending on this ethical background. When consider the rental house prediction project, it has a high priority on ethical issues. They are data privacy or protection problem of data, transparency of the crated model, and consideration of the user feedback and improvement of those. To predict the good outcome there should give some effort to keep privacy and data security to

protect the data of individuals. To improve the transparency of the collected data, there should be used storage and established through the informed consent of contributors to the dataset. Then the project's commitment to fairness and reduction of bias is an essential ethical element.

Discussion:

To estimate the rental price of the houses the best model created by Decision tree algorithm. The outcome of the final demonstration is that the Decision Tree algorithm has lower Mean Squared Error and Mean Absolute Error values. It means that Decision trees are naturally good at identifying non-linear patterns and complex correlations in the data. The ability of the model can prioritize different criteria offers a useful insight into the primary factors affecting rental estimates. When making real estate decisions, this feature is especially helpful. Furthermore, the interpretation of Decision Trees facilitates a transparent decision-making process and facilitates the dissemination of insights to stakeholders.

In real world examples, Decision trees can have an ability to stand with outliers and they are frequently seen in datasets. It helps to produce solid forecasts and raises the model's overall reliability. Decision trees' ability to adapt to heterogeneous datasets is in line with the variety of factors affecting home values. The decision tree is the recommended method for estimating house rent in this project because of its overall performance which includes accuracy, interpretability.

V. Conclusion

In conclusion, the Decision Tree algorithm is a best algorithm for predicting rental house price in this project, because it has better performance than linear regression algorithm and there is low error values for MSE and MAE. Its capacity can be handle complicated relationships, feature importance transparency, and remove outliers. To make sure the selected model can be used to a wide range of scenario. To increase the performance there can be applied some additional methods and processors to the system.
So, by considering all the above measurements and ethics, the final development model for the dataset is very responsible and very accurate machine learning.