Comparative Analysis of Regression Models for California Housing Price Prediction

Ahmad Alsharif Student ID: 4110623

Hassan Alharbi Student ID: 4210831

Khaldoun Alkhoja Student ID: 4111908

Abstract

This paper compares the performance of three regression models in predicting housing prices in California. The models evaluated are Linear Regression, Decision Tree, and Random Forest, each employing sophisticated preprocessing techniques and model construction strategies. The goal is to understand how different approaches perform on a common dataset. The paper discusses intensive experiments and evaluates the models using various metrics, including Mean Squared Error, Root Mean Squared Error, Mean Absolute Error, and R-squared. The findings reveal insights into the strengths and weaknesses of each model, providing guidance for future applications and research in housing price prediction.

1 Introduction

Housing prices play a crucial role in the economic health and social stability of regions, particularly in dynamic markets such as California. The ability to accurately predict housing prices is valuable for a variety of stakeholders, including real estate professionals, potential homeowners, investors, and policymakers. This predictive capacity aids in decision-making, investment strategies, and policy formulation.

Predicting housing prices is a complex task influenced by numerous factors, including geographic location, economic trends, demographic characteristics, and property-

specific features. Machine learning offers robust methodologies for addressing this complexity, providing advanced tools to capture and model the intricate relationships between these factors and housing prices.

Various machine learning regression models have been employed for housing price prediction, ranging from linear models that assume a direct proportional relationship between features and target, to more sophisticated non-linear and ensemble models that can capture complex interactions and patterns. Each model type presents unique advantages and challenges, necessitating careful evaluation to determine the most suitable approach for specific datasets and objectives.

In this study, we aim to investigate the predictive performance of different regression models in predicting housing prices using a dataset of California housing. The models compared include Linear Regression, Decision Tree, and Random Forest, each representing a different approach to regression analysis. We evaluate these models based on several key metrics, such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2), to provide a comprehensive assessment of their accuracy and robustness.

By exploring the strengths and limitations of each model, this research contributes to the understanding of effective strategies for housing price prediction and offers insights that can guide future research and practical applications in the field.

2 Literature Review (Related Work)

Previous research in housing price prediction has explored diverse methodologies and approaches. Two notable studies in this field are:

2.1 Study 1: Literature Review on House Price Prediction

In a study by Ritu [1], the author reviewed the application of machine learning techniques for house price prediction. The study analyzed the effectiveness of regression models, including Linear Regression and Random Forest, in predicting housing prices based on socio-economic factors. The review emphasized the importance of feature selection and model evaluation for improving prediction accuracy.

2.2 Study 2: Housing Price Prediction

In a study by Truong et al. [2], the authors investigated housing price prediction using improved machine learning techniques. The researchers examined the performance of traditional and advanced machine learning models, including ensemble methods like Stacked Generalization, in predicting housing prices based on diverse features. The study highlighted the significance of comprehensive model validation and the potential of advanced models for accurate housing price prediction.

These studies underscore the importance of employing advanced regression techniques and rigorous evaluation methodologies in predicting housing prices accurately.

3 Experiments

3.1 Data Description

The dataset on housing in California includes details like income, age of housing, number of rooms and bedrooms, population, households, and geographic coordinates. These aspects represent socio-geographical factors that impact housing costs.

Figure 1 illustrates the correlation matrix among the features and the target variable, highlighting the relationships between various factors affecting housing prices.

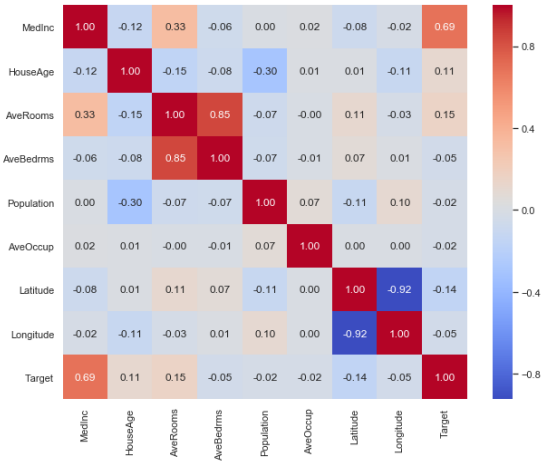

Figure 1: Correlation matrix of the features and target variable.

Figure 2 shows the distribution of each feature and the target variable, providing insights into their frequency distributions.

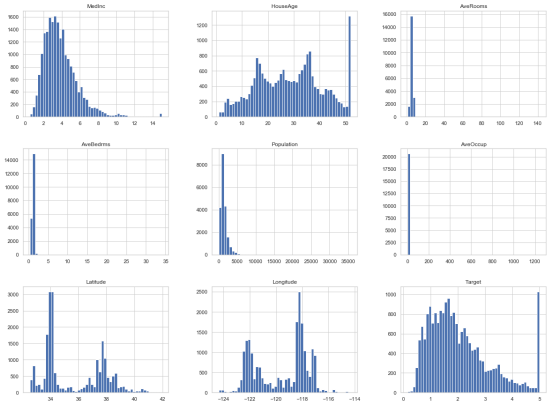

Figure 2: Histograms for each feature and the target variable.

3.2 Data Pre-processing

Before the model can be constructed the data is rigorously pre-processed making sure the data is clean and the solution is robust. This includes missing values handling, feature Standard-Scaling to scale the values in order to have the same standard deviation and same mean, and splitting the data into the training set and the testing set which is needed to evaluave the model in a robust manner.

3.3 Models Built

Three regression models are constructed, each with distinct complexities and capabilities tailored to the task of housing price prediction:

- **Linear Regression**: A foundational regression model that assumes a linear relationship between input features and target variable.
- **Decision Tree**: A non-linear regression model that recursively partitions the feature space into hierarchical decision nodes, capturing complex interactions among features.
- **Random Forest**: An ensemble regression model comprising multiple decision trees, which aggregates predictions from individual trees to improve generalization performance and mitigate overfitting.

3.4 Evaluation Metrics

The performance of each regression model is assessed using a suite of evaluation metrics, including:

- Mean Squared Error (MSE): Measures the average squared difference between predicted and actual housing prices, providing a comprehensive assessment of model accuracy.
- Root Mean Squared Error (RMSE): The square root of MSE, offering an interpretable measure of prediction error in the same units as the target variable.
- Mean Absolute Error (MAE): Computes the average absolute difference between predicted and actual housing prices, offering robustness against outliers.
- R-squared (R2): Also known as the coefficient of determination, R2 quantifies

the size of variance in the target variable explained by the regression model, with higher values indicating better model fit.

3.5 Results

The performance of each regression model is summarized based on the evaluation metrics outlined above.

- **Linear Regression**: MSE: 0.556, RMSE: 0.746, MAE: 0.533, R2: 0.576
- **Decision Tree**: MSE: 0.408, RMSE: 0.639, MAE: 0.431, R2: 0.689
- **Random Forest**: MSE: 0.296, RMSE: 0.544, MAE: 0.367, R2: 0.774

Figure 3 shows a comparison of the performance of the three regression models based on the evaluation metrics. The plot highlights the strengths and weaknesses of each model, demonstrating the effectiveness of Random Forest in achieving lower error metrics and higher R-squared values.

4 Discussion

The results analysis informs on the advantage and disadvantage of each regression model.

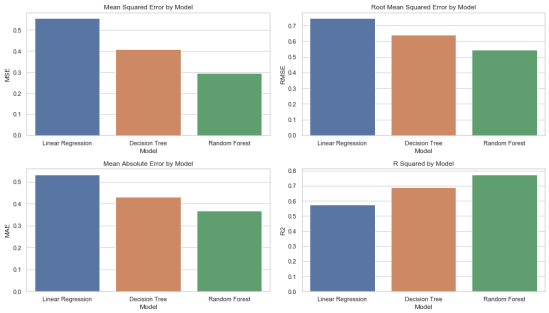

Figure 3: Models Performance.

For example, linear regression is easy to interpret but suffers from data overfitting. On the other hand, decision tree is flexible in capturing nonlinear relationship but vulnerable to noise. Random forest, on the other hand, is robust against linear regression overfitting. Furthermore, we discuss the real-world applications of the commonly used regression models and forward some ideas for future research.

5 Conclusion

This study concludes that selecting an appropriate regression mode is very crucial in housing price prediction task. Linear Regression is the basic model but more advance models like Decision Tree and Random Forest models provide better prediction and handle complex data pattern more robustly. By utilising advance regression techniques and validating

the model using a proper test procedure, we can make better policy decisions in California's highly dynamic housing market.

6 References

References

[1] Ritu, "Machine Learning Techniques for House Price Prediction: A Literature Review," *ResearchGate*, 2023. Retrieved from https://doi.org/10.13140/RG.2.2.31544.52480

[2] Q. Truong, M. Nguyen, H. Dang, and B. Mei, "Housing Price Prediction via Improved Machine Learning Techniques," *Procedia Computer Science*, vol. 174, pp. 433-442, 2020. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050920316318