DAT 554 Final Project: California House Price Prediction

Matta Ramyasravani

***Abstract* - This project focuses on the development of a predictive model for housing prices using the California housing dataset. The dataset comprises various features such as geographical coordinates, housing median age, population, and median income, with the objective of predicting median house values. The project encompasses data preprocessing, exploratory data analysis (EDA), and model development. In the data preprocessing phase, missing values in the 'total_bedrooms' column were addressed by imputing them with the mean value. EDA unveiled valuable insights into the dataset, including feature distributions, correlations, and the creation of new features like 'rooms_per_household' and bedrooms_per_room.' Further analysis includes the evaluation of different regression models to predict housing prices. This report highlights the implementation of hyperparameter tuning using GridSearchCV to optimize the model's performance. Results demonstrate the model's capability in explaining a significant portion of the variance in house prices. The research showcases the importance of machine learning in real estate forecasting and provides a framework for improving predictions in the housing market. The project contributes to the field of predictive modeling by enhancing housing price estimation accuracy**.

***Keywords:* Predictive Modeling, Housing Prices, California Housing Dataset, Hyperparameter Tuning, Machine Learning.**

I. INTRODUCTION

Background:

The housing market is a critical sector of any economy and predicting housing prices accurately is of paramount importance for various stakeholders, including homebuyers, sellers, real estate agents, and investors. The California housing market, in particular, is known for its complexity due to its diverse geographical and socioeconomic factors [5]. Accurate price predictions are essential for informed decision-making in this dynamic environment. This project aims to develop a predictive model for housing prices in California using machine learning techniques. The dataset used for this project is obtained from Kaggle website. The dataset provides a rich set of features, including geographical coordinates, housing characteristics, and socioeconomic factors, which make it a suitable candidate for housing price prediction [5].

The Problem Statement:

The primary problem addressed in this project is the prediction of median house values in various districts of California. Given a set of features such as location, housing median age, population, and median income, the goal is to build a model that can accurately estimate the median house value in a district.

This project will explore data preprocessing techniques, conduct exploratory data analysis (EDA) to gain insights into the dataset, and develop predictive models to tackle this problem [1]. Additionally, hyperparameter tuning will be employed to optimize the model's performance.

Project Evolution:

Throughout the course of this project, our focus evolved as we delved deeper into the dataset and gained a better understanding of its characteristics. We initially set out to predict housing prices solely based on geographical features but realized that incorporating additional socioeconomic features significantly improved model accuracy. This evolution reflects the iterative and adaptive nature of data science projects, where insights gained during the process can lead to refinements in problem definition and methodology.

II. DATASET

(1) Source of the Dataset:

The dataset used for this project was sourced from Kaggle can be accessed via the following

URL:

https://www.kaggle.com/datasets/camnugent/california-housing-prices.

(2) Relevance to the Problem:

The dataset is highly relevant to the problem of predicting housing prices in California. It provides a comprehensive collection of features that influence housing prices, such as geographical coordinates, housing median age, total rooms, total bedrooms, population, households, and median income [5]. Understanding the impact of these factors on housing prices is crucial for various stakeholders in the real estate industry and for homebuyers and sellers.

Characteristics of the Dataset:

Size: The dataset consists of a total of 20,640 entries, each representing a district in California.

Features: The dataset contains both numerical and categorical features, including:

Numerical Features: Longitude, Latitude, Housing Median Age, Total Rooms, Total Bedrooms, Population, Households, Median Income.

Categorical Feature: Ocean Proximity.

Target Variable: The target variable is "Median House Value," which represents the median value of owner-occupied homes in a district.

Preprocessing and Cleaning:

The dataset required several preprocessing steps to ensure its suitability for machine learning modeling:

Handling Missing Values: The 'total_bedrooms' column had missing values, which were addressed by replacing them with the mean value of that column. This ensured that the dataset was complete and ready for analysis.

Exploratory Data Analysis (EDA):

The EDA phase involved a comprehensive analysis of the dataset to gain insights and understand patterns. Key findings from EDA include:

Feature Distributions: Histograms and kernel density estimates were used to visualize the distributions of numerical features. This revealed the spread and central tendencies of each feature.

Feature Correlations: A correlation matrix was calculated to identify relationships between numerical attributes. Notable correlations were observed between median house value and median income.

Feature Engineering: New features such as 'rooms_per_household,' 'bedrooms_per_room,' and 'population_per_household' were created to capture potentially informative relationships.

Overall, EDA helped identify patterns, correlations, and outliers in the dataset. These insights guided feature selection and engineering, laying the foundation for building predictive models for housing prices in California.

III. SOLUTION

In this section, we describe the techniques and methodologies employed to address the problem of predicting housing prices in California, emphasizing the use of Linear Regression and Hyperparameter tuning.

Machine Learning Techniques:

Linear Regression: Linear Regression was the primary modeling technique utilized in this project. It is a straightforward yet effective method for predicting continuous target variables. Linear Regression establishes a linear relationship between the input features and the target variable and was employed as a baseline model for performance comparison.

Hyperparameter Tuning:

Hyperparameter tuning was a crucial aspect of the project to optimize the Linear Regression model's performance. The hyperparameters considered for tuning may include regularization parameters (e.g., alpha), feature scaling methods, and other parameters that influence the model's behavior. Techniques like cross-validation were employed to find the best combination of hyperparameters that minimized the prediction error.

Challenges and Alternative Approaches:

During the course of the project, several challenges were encountered, and alternative approaches were considered:

Feature Engineering: A significant challenge was feature engineering. We experimented with various feature combinations and transformations to capture meaningful relationships. Some features, such as 'rooms_per_household' and 'bedrooms_per_room,' were introduced to enhance model performance.

Model Selection: Model selection played a crucial role in achieving accurate predictions. While linear regression was straightforward, it had limitations in capturing complex non-linear relationships.

Hyperparameter Tuning: Tuning hyperparameters for complex models was computationally intensive. We conducted grid search and cross-validation to optimize hyperparameters for the chosen models.

Data Preprocessing: Data preprocessing, including handling missing values and encoding categorical features, required careful consideration to ensure that the dataset was prepared adequately for modeling.

Interpretability vs. Performance: Balancing model interpretability with predictive performance was a key consideration. Random Forest and Gradient Boosting models provided improved performance but were less interpretable compared to linear regression.

Overfitting: Overfitting was a concern when using complex models. Regularization techniques and hyperparameter tuning were explored to mitigate overfitting.

IV. EVALUATION METRICS

In this section, we elaborate on the evaluation methods and metrics used to assess the performance of the Linear Regression model and present the corresponding results.

Evaluation Metrics:

Mean Squared Error (MSE): MSE is a widely used metric for regression problems. It measures the average squared difference between the predicted values and the actual target values. Lower MSE values indicate better model performance.

Mean Absolute Error (MAE): MAE calculates the average absolute difference between the predicted and actual values. It provides a measure of the magnitude of errors and is easier to interpret than MSE.

Root Mean Squared Error (RMSE): RMSE is the square root of the MSE and is expressed in the same units as the target variable. It gives an idea of the average prediction error and is often used for its interpretability.

R-squared (R2) Score: R2 measures the proportion of variance in the target variable that can be explained by the model. It ranges from 0 to 1, with higher values indicating better model fit. It is a useful metric for understanding how well the model captures variance in the data.

Results:

After training and evaluating the Linear Regression model on the California housing dataset, the following results were obtained:

Mean Squared Error (MSE): Approximately 5569121311.08

Mean Absolute Error (MAE): Approximately 51511.95

Root Mean Squared Error (RMSE): Approximately 74626.55

R-squared (R2) Score: Approximately 0.58.

V. CONCLUSION

In conclusion, this project aimed to predict housing prices in California using a machine learning approach, primarily focusing on Linear Regression. The project encompassed data preprocessing, exploratory data analysis (EDA), model development, and the evaluation of model performance.

The key findings and outcomes of the project are summarized as follows:

Model Performance Before Hyperparameter Tuning:

Mean Squared Error (MSE): Approximately 5,569,121,311.08

Mean Absolute Error (MAE): Approximately 51,511.95

Root Mean Squared Error (RMSE): Approximately 74,626.55

R-squared (R2) Score: Approximately 0.58

These initial results provided a baseline for model performance. The Linear Regression model demonstrated its capability to make predictions but also indicated room for improvement.

Challenges and Considerations:

Extensive feature engineering was performed, including the creation of new features to capture relationships within the data. Missing values in the dataset were addressed by imputing them with appropriate strategies, ensuring a complete dataset [1]. The choice of Linear Regression balanced interpretability with performance. More complex models were considered but selected based on the interpretability requirement.

REFERENCES

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani, "An Introduction to Statistical Learning with Applications in R," Springer, 2013.

[2] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer, 2009.

[3] S. Raschka and V. Mirjalili, "Python Machine Learning," Packt Publishing, 2019.

[4] W. McKinney, "Python for Data Analysis," O'Reilly Media, 2017.

[5] "Kaggle Datasets - California Housing Prices," [Online]. Available: https://www.kaggle.com/datasets/camnugent/california-housing-prices.

[6] "Scikit-Learn Documentation," [Online]. Available: https://scikit-learn.org/stable/documentation.html.

[7] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, Oct. 2011, pp. 2825-2830. [Online]. Available: https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

[8] H. McKinney, "Exploratory Data Analysis," [Online]. Available: https://www.jstor.org/stable/2682899.

[9] L. Torgo, "Data Mining with R: Learning with Case Studies," Chapman and Hall/CRC, 2017.

[10] "California Department of Housing and Community Development," [Online]. Available: https://www.hcd.ca.gov/.