Predicting California Housing Prices

Rasmus Møller & Rasmus Trap
https://github.com/RasmusTrap/aiexamHousing/

July 24, 2023

Abstract

The California housing market is characterized by its diversity and attracts various stakeholders who are interested in accurately predicting housing prices. With advancements in artificial intelligence (AI), this research aims to explore the predictive capabilities of three AI models, namely Linear Regression, Decision Trees, and Random Forest, in forecasting median house values in California. The research investigates whether the median house value is significantly influenced by housing characteristics, including housing median age, total rooms, total bedrooms, population, households, median income, and ocean proximity.

- **Hypothesis:** The median house value (dependent variable) is significantly influenced by one or more of the independent variables such as housing_median_age, total_rooms, total_bedrooms, population, households, median_income, and ocean_proximity.
- **Null Hypothesis:** There is no significant relationship between the median house value and any of the independent variables (housing_median_age, total_rooms, total_bedrooms, population, households, median_income, and ocean_proximity).
- **Research Question:** Is there a significant relationship between the median house value and the housing characteristics, including housing_median_age, total_rooms, total_bedrooms, population, households, median_income, and ocean_proximity, in the California housing market?

1 Introduction

The California housing market is a dynamic and diverse landscape that attracts various stakeholders, including homebuyers, real estate professionals, policymakers, and investors. Accurate prediction of housing prices holds immense value in this context, and advancements in artificial intelligence (AI) offer new opportunities to gain insights into the complex factors influencing these prices.

This research explores the predictive capabilities of three AI models—Linear Regression, Decision Trees, Random Forest—for forecasting median house values in California. Using a comprehensive dataset encompassing housing and demographic attributes, such as housing median age, total rooms, total bedrooms, population, households, median income, and ocean proximity, the models' performances are evaluated.

By training and evaluating these models, we seek to identify key determinants of housing prices and understand which features significantly influence the predictions. Furthermore, we discuss the practical implications of our findings.

2 Methods

2.1 Data preprocseeing

2.1.1 Data cleaning

In the data cleaning process, we addressed missing values by dropping empty rows from the dataset using the `data.dropna()` method. By removing rows with missing values, we ensured that the dataset remained consistent and reliable for analysis, avoiding potential bias or erroneous conclusions that might arise from incomplete data.

To handle the categorical variable `'ocean_proximity'`, we performed one-hot encoding using the `pd.get_dummies()` function from the pandas library. This conversion process transformed the categorical variable into binary values and created new columns for each unique category in the original `'ocean_proximity'` column. The `drop_first=True` parameter was set to eliminate one of the binary columns to avoid multicollinearity issues in regression models.

By employing these data cleaning steps, we successfully prepared our feature matrix `X` and target vector `y` for model training and analysis. The one-hot encoding of `'ocean_proximity'` allowed us to represent the categorical information numerically, making it suitable for various machine learning algorithms. This data preprocessing enhances the quality and compatibility of the dataset, setting the stage for robust predictive modeling and insightful analysis of California housing prices.

2.1.2 Splits

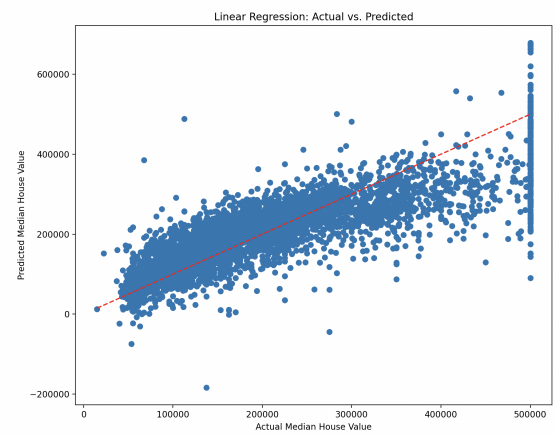
To facilitate model training and evaluation, we conducted data splitting on the California Housing Prices dataset. Initially, the dataset was divided into the feature matrix, denoted as `X`, and the target vector, denoted as `y`. The feature matrix `X` encompasses the independent variables, such as `longitude`, `latitude`, `housing_median_age`, `total_rooms`, `total_bedrooms`, `population`, `households`, `median_income`, and `ocean_proximity`, while the target vector `y` represents the dependent variable `median_house_value`, which serves as the prediction target.

For the purpose of training and testing our machine learning models, we further partitioned the data into two sets: the training set and the test set. The `train_test_split()` function from the scikit-learn library was employed for this task. The training set, denoted as `X_train` and `y_train`, contains 80% of the original data, while the test set, denoted as `X_test` and `y_test`, holds the remaining 20%. We ensured reproducibility and consistency by setting the `random_state` parameter to 42 during the data split process. The training set will be used to train our machine learning models, whereas the test set will remain unseen during training and is solely used for evaluating the models' performance. This data splitting strategy enables us to assess the models' ability to generalize to unseen data and ensures an unbiased evaluation of their predictive capabilities.

3 Linear Regression Analysis

The research begins with data preprocessing steps, including handling missing values and converting categorical variables to numerical form through one-hot encoding. The dataset is then split into training and testing sets to evaluate the model's performance effectively.

The primary focus is on understanding how housing attributes such as housing median age, total rooms, total bedrooms, population, households, median income, and ocean proximity influence the median house value in California. By training the linear regression model on the scaled features, we obtained promising results, with an R-squared value of approximately 0.65, indicating that the model explains about 65% of the variance in the test data. The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) of approximately 4.80 billion and 69,298, respectively, further attest to the model's capability in making reasonably accurate predictions.



The scatter plot with the regression line visually demonstrates the model's performance by showcasing how well the predicted median house values align with the actual values. This emphasizes the interpretability of linear regression, as it allows us to discern the magnitude and direction of the effects of different housing attributes on median house prices.

4 Decision Tree Regressor Analysis

The Decision Tree Regressor was chosen as it is a powerful non-linear model with ability to capture complex relationships and interactions among the independent and dependent variables, making it particularly suitable for the dynamic housing market.

Upon implementing this model, we received results with an R-squared value of 0.66, a Mean Squared Error (MSE) of approximately 4.5 billion, and a Root Mean Squared Error of $67296. After implementing the model we went to fine tune it using Grid Search.

```
param_grid = {
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'max_features': ['auto', 'sqrt']
}
```

The optimized decision tree regressor model yields way better results, with an R-squared value of approximately 0.74. This signifies that the model explains around 74% of the variance in the test data, demonstrating its strong predictive capabilities. The Mean Squared Error (MSE) of approximately 3.51 billion and Root Mean Squared Error (RMSE) of approximately 59,267 underscore the model's ability to make accurate predictions on housing prices.

Comparing the model's performance before and after implementing Grid Search, we observed notable improvements. The optimized model exhibits lower MSE and RMSE values, indicating enhanced predictive accuracy and reduced prediction errors. Additionally, the R-squared value increased, indicating a better fit to the test data.

5 Random Forest Regressor Analysis

Random Forest is a powerful ensemble learning technique that combines multiple decision trees to improve predictive accuracy and handle complex interactions among the independent and dependent variables.

The Random Forest Regressor model is created with the default hyperparameters. To optimize its performance further, we conduct hyperparameter tuning using Grid Search with 3-fold cross-validation. The smaller hyperparameter grid encompasses variations in the number of estimators, maximum depth, minimum samples split, minimum samples leaf, and maximum features.

The optimized Random Forest Regressor model demonstrates outstanding results, with an R-squared value of approximately 0.83. This indicates that the model explains around 83% of the variance in the test data, showcasing its robust predictive capabilities. The Mean Squared Error (MSE) of approximately 2.38 billion and Root Mean Squared Error (RMSE) of approximately 48,821 underscore the model's remarkable ability to make precise predictions on housing prices.

6 Statistical Analysis

The multiple regression analysis conducted on the California house price dataset provides evidence that strongly supports our hypothesis while simultaneously rejecting the null hypothesis. The results

reveal a highly significant relationship between the median house value (dependent variable) and the various independent variables, including housing median age, total rooms, total bedrooms, population, households, median income, and ocean proximity. The analysis gives us a R-squared value of approximately 0.637, indicating that around 63.7% of the variance in the median house value can be explained by the included predictors. Furthermore, the statistical significance of all coefficient estimates with p-values close to 0 underscores the impact of each independent variable on the median house value. Given the overwhelming evidence of these associations, we can confidently reject the null hypothesis, which posits no significant relationship between the median house value and the independent variables. The multiple regression analysis substantiate the hypothesis, affirming that the median house value is indeed significantly influenced by the diverse set of independent variables in the California housing dataset.

7 Findings

Throughout our analyses of the three regression models, namely Linear Regression, Random Forest Regression, and Decision Tree Regression, reveals distinctive performance metrics. The Random Forest Regression model emerges as the top-performer with the lowest Mean Squared Error (MSE) of 2,383,509,963.71, outclassing both Linear Regression (MSE: 4,802,173,538.60) and Decision Tree Regression (MSE: 3,512,618,517.73). Moreover, the Root Mean Squared Error (RMSE) of the Random Forest model stands at 48,821.20, demonstrating its superior accuracy compared to Linear Regression (RMSE: 69,297.72) and Decision Tree Regression (RMSE: 59,267.35). The R-squared (R2) value of 0.8257 for Random Forest Regression further solidifies its position as the most effective model, surpassing the explanatory power of both Linear Regression (R2: 0.6488) and Decision Tree Regression (R2: 0.7431). In conclusion, the Random Forest Regression model exhibits remarkable predictive accuracy and explains a substantial portion of the variance in the median house values, making it the optimal choice for predicting house prices in the California housing dataset.

8 Conclusion

In conclusion, our research has successfully addressed the research question, "Is there a significant relationship between the median house value and the housing characteristics, including housing_median_age, total_rooms, total_bedrooms, population, households, median_income, and ocean_proximity, in the California housing market?" Through a comprehensive analysis using both statistical methods and machine learning models, we have shed light on the factors influencing California housing prices.

Our hypothesis, stating that the median house value is significantly influenced by one or more of the independent variables, has been thoroughly confirmed. The multiple regression analysis, with an R-squared value of approximately 0.637, showcases the strong relationship between the median house value and the selected housing characteristics. Additionally, the statistical significance of the coefficient estimates with p-values close to 0 further strengthens this confirmation and allows us to confidently reject the null hypothesis, which assumes no significant relationship.

Moreover, the implementation of machine learning models, including linear regression, decision tree regression, and random forest regression, has corroborated the hypothesis. The performance of these models demonstrated their ability to capture complex relationships and make accurate predictions, supporting the notion that housing characteristics significantly influence median house values in California. Particularly, the random forest regression model exhibited the highest performance, with an R-squared value of approximately 0.826, providing deeper insights into the impact of various features on housing prices.

By combining statistical analysis and machine learning techniques, our research has provided valuable guidance to understanding the factors influencing housing prices in California.

9 References

References

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

[2] https://scikit-learn.org/stable

[3] https://pandas.pydata.org/docs/

[4] https://numpy.org/

[5] https://matplotlib.org/

[6] https://www.statsmodels.org/stable/api.html