Nick Tran

Writeup

In this project, we aimed to predict the median house values in California, given various metrics like population, median income, etc., for each district. Our dataset, sourced from the US Census Bureau, comprised 20640 records and 10 metrics, with 'median_house_value' as the target variable.

To begin with, we loaded our data from the 'housing.csv' file and examined its structure. We identified the input (X) and output (Y) variables and found that our data had both numerical and categorical variables.

Next, we performed data cleaning, a crucial step in preparing our dataset for predictive modeling. We filled the missing values in our dataset with the mean of the respective column, which provided us a complete dataset for further steps. It's important to note here that while filling missing values with mean values works well for numerical columns, we might want to consider other strategies like mode or specific value imputation for categorical columns.

For our categorical column 'ocean_proximity', we used one-hot encoding to convert the categories into a binary (0 and 1) format, making it suitable for our machine learning models.

With our data cleaned and formatted correctly, we split the dataset into a training set and a test set, with an 80:20 split. This partitioning allowed us to train our models on a portion of the data and then test it on unseen data to evaluate its performance.

Before feeding our data into machine learning algorithms, we standardized it to bring all features to the same scale, making the training process more efficient and the resulting model more accurate.

We then used three different regression algorithms - Linear Regression, Decision Tree Regression, and Random Forest Regression - to train models and predict the median house value. For each model, we calculated the Root Mean Squared Error (RMSE) to measure the differences between the values predicted by the model and the actual values.

As a bonus exercise, we also performed Linear Regression with 'median_income' as the sole independent variable and plotted the results, offering an intuitive visual understanding of the model's performance.

Overall, this project provided valuable insights into the process of predictive modeling, including data loading, cleaning, preparation, model training, prediction, and evaluation. With more fine-tuning and use of advanced techniques, the models can be further improved for more accurate predictions.