ECE 225A Project Report - California Housing Prices Prediction

Kirtan Shah
Department of Electrical and Computer Engineering, University of California, San Diego

Abstract

This project analyzes California Housing Prices and predicts median house values using Linear Regression, K Nearest Neighbors, Random Forest, and XGBoost.

1 Introduction

The California housing market has long been a subject of interest for economists, policymakers, and residents. This paper presents an exploratory data analysis of the California Housing Prices dataset, a collection of housing-related statistics from the 1990 California census [1]. For this project, I focus on uncovering insights into patterns within the data, especially in relation to median house values across different census blocks in California. Through data preprocessing, statistical modeling, and visualization techniques, I explore features and their predictive value in determining median house values. I also compare regression models for median house values, which provide further insight into the complex dynamics of California's housing market.

2 Dataset

The California Housing Prices dataset [1], provided by the US Census and available on Kaggle, comprises housing-related statistics for different blocks in California. Each entry represents statistics for a group of homes in a given block.

2.1 Dataset Description

These variables include the following:

- Latitude and longitude
- Median house age within block
- Total rooms, total bedrooms
- Total population

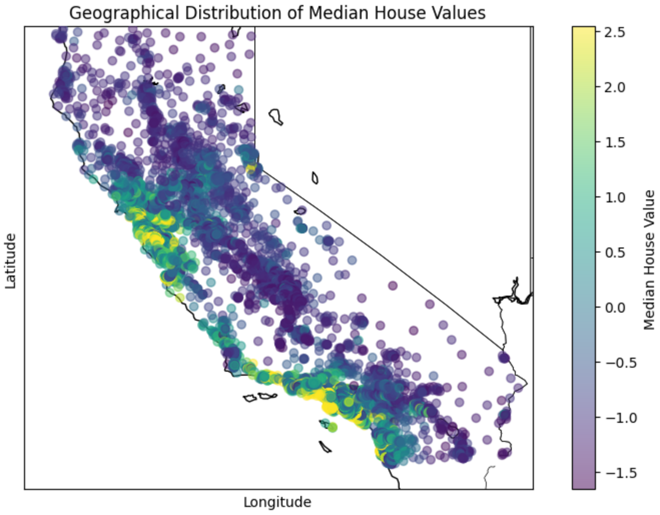

Figure 1: Visualization of California median house values

- Number of households in block
- Median income for households in block
- Median house value for households in block
- Ocean proximity (<1H OCEAN, INLAND, ISLAND, NEAR BAY, NEAR OCEAN)

The dataset contains 20640 entries corresponding to a different census block, each with the features described above.

2.2 Exploratory Data Analysis

The dataset contains a variety of numerical and categorical features, wich need proper preprocessing before they can be used for analysis. First, we recognize the categorical variable of ocean proximity, for which it would make little sense to represent numerically; for example, if ISLAND is mapped to 1 and NEAR OCEAN is mapped to 2, it implies that NEAR OCEAN is somehow twice whatever ISLAND represents,

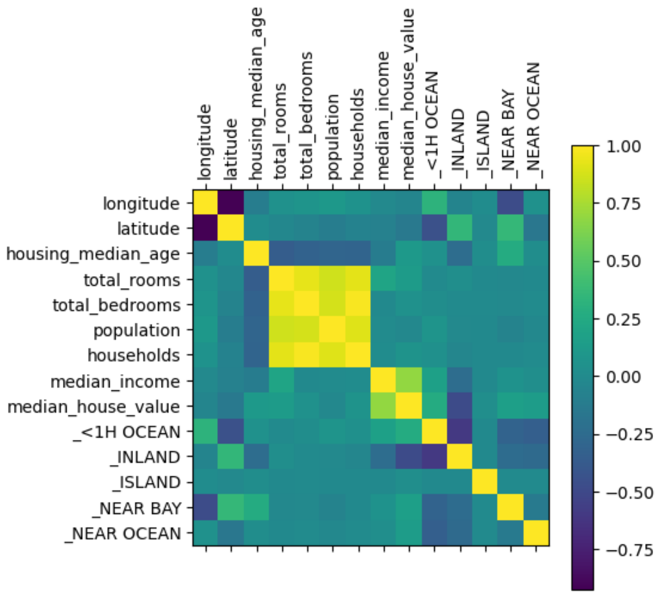

Figure 2: Correlation matrix for housing features

which doesn't necessarily make sense. Thus, we use one-hot encoding to represent the ocean proximity variable as a set of binary variables, each representing a different category. Here is an example one hot vector for an entry with ISLAND:

$$V_{\text{ISLAND}} = \begin{array}{c} \text{<1H OCEAN} \\ \text{INLAND} \\ \text{ISLAND} \\ \text{NEAR BAY} \\ \text{NEAR OCEAN} \end{array} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \tag{1}$$

Next, we normalize the numerical features to ensure that they are on similar scales and have relatively small values. This is important because we don't want to bias metrics like correlation or our predictor towards features with larger values (e.g. latitude/longitude) over those with smaller values (e.g. total rooms). Therefore, we normalize each feature to have a mean of 0 and a standard deviation of 1 as follows:

$$x' = \frac{x - \mu}{\sigma} \tag{2}$$

With the preprocessed features, we can now analyze the data to uncover patterns and relationships between different features. One way to do this is by computing the correlation matrix for the features, which gives us a sense of how each feature is related to the others. This is done by computing a matrix C where C_{ij} is defined in equation 3.

$$C_{ij} = \frac{Cov(X_i, X_j)}{\sigma_i \sigma_j} \tag{3}$$

The correlation matrix is visualized in Figure 2. Notice that similar features like total rooms and total bedrooms have a high correlation, while features relating to ocean proximity have low correlation with population and number of households. Most importantly, the median house value has the highest correlation with median income, which suggests that income could be key factor in determining house prices.

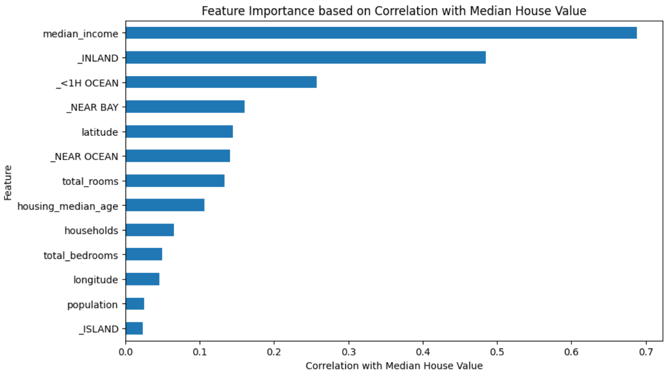

Figure 3: Inferred Feature Importance based on correlation with median house value

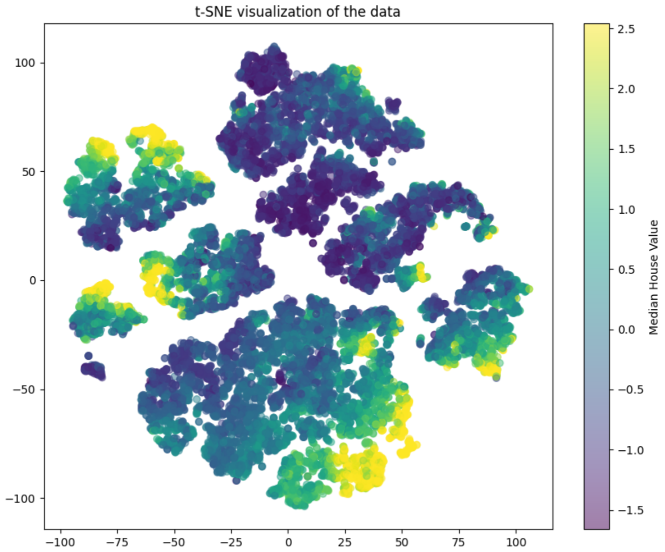

Figure 4: t-SNE visualization of California housing data

Zooming into median house value correlations specifically, we plot the magnitudes of the correlations of each feature and median house value in Figure 3. From correlations alone, it seems that median income and ocean proximity seem to be the most important features in determining median house value.

With 20640 rows of multi dimensional data, it is hard to visualize the complexity (or hidden simplicity) of the data in a condensed plot. However, a technique called t-SNE (t-distributed stochastic neighbor embedding) can be used to reduce the dimensionality of the data to 2D for visualization purposes [4]. Figure 4 shows the t-SNE visualization of the California hous-

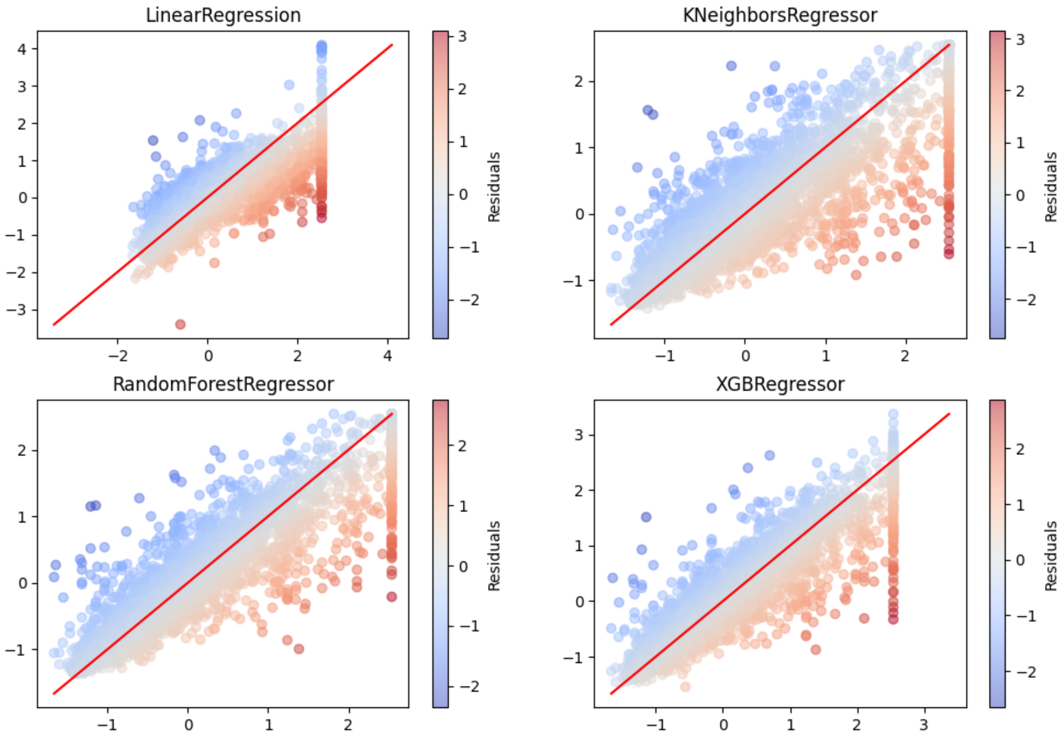

Figure 5: Predicted vs Actual Median House Values

ing data, and reveals the presence of several clusters of data points. This suggests that an algorithm like K-Nearest Neighbors, which relies on values of proximal data points, could be effective in predicting median house values.

3 Predictive Modeling

In this section, we will explore the use of a few different regression models to predict median house values based on the features in the dataset.

3.1 Models

We will start off with a Linear Regression model. Based on the findings from the exploratory data analysis, we will try to improve results with a K Nearest Neighbors regressor. Finally, we utilize a Random Forest regressor and XGBoost regressor to see if we can further improve our predictions with state-of-the-art, popular models.

3.2 Evaluation

In order to accurately evaluate the performance of each model, we split the data into a training set and a test set. For this project, we use a split of 80% training data and 20% test data. The evaluation metric is the mean squared error (MSE), which measures the average squared difference between the predicted and actual median house values. A lower MSE indicates a better model.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

4 Results

The models are evaluated based on their mean squared error (MSE) on the test set. The results are summarized in Table 1.

Model	MSE
Linear Regression	0.3624
K Nearest Neighbors	0.2884
Random Forest	0.1838
XGBoost	0.1741

Table 1: Model Performance

For a more holistic view of the model performance, we plot the predicted vs actual median house values for each model in Figure 5. The hypothesis that median house values tend to be clusters seems to be somewhat correct, with K Nearest Neighbors performing better than Linear Regression. Yet, Random Forest and XGBoost regressors perform the best, with XGBoost having the lowest MSE of 0.1741. Random Forest is an ensemble model that uses multiple decision trees

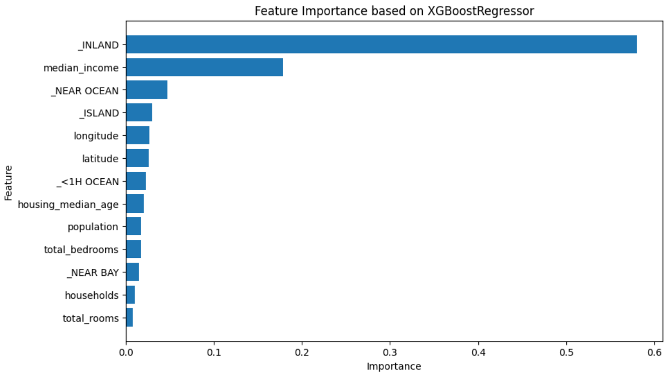

Figure 6: Feature Importance in XGBoost Model

to make predictions [3], while XGBoost is a gradient boosting model combined with decision trees [2]. The superior performance of these models suggests that the relationships between the features and median house values are complex and non-linear. Interestingly, features like ocean proximity and median income continue to be important in the XGBoost, corroborating what we found with the correlation matrix. I conclude that the top most features in Figure 6 are the most important in determining house value.

5 Conclusion

In this project, I conducted an exploratory data analysis of the California Housing Prices dataset, focusing on uncovering patterns and relationships between different features. I found that median income and ocean proximity are the most important features in determining median house values, specifically whether or not a home is inland.

6 Code

The code for this project is uploaded to GitHub at `https://github.com/kirtan-shah/ECE225A-California-Housing-EDA`.

References

[1] California Housing Prices — kaggle.com. `https://www.kaggle.com/datasets/camnugent/california-housing-prices`. [Accessed 15-12-2024].

[2] Chen, T., and Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2016), KDD '16, ACM, pp. 785–794.

[3] Ho, T. K. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (1995), vol. 1, IEEE, pp. 278–282.

[4] van der Maaten, L., and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research 9*, 86 (2008), 2579–2605.