

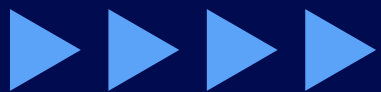


Capstone 3



CALIFORNIA **HOUSING PRICE**

By : Gian Sirait



Business Problems

- A real estate company in California faces a major challenge in determining accurate prices for properties.
- Improper pricing can lead to:
 - Financial losses
 - Lost business opportunities
 - Difficulty selling properties due to high prices
 - Inventory build-up and increased operating costs
 - Selling properties below true market value, leading to long-term losses





PROBLEM STATEMENT

01.

Real estate companies in California face challenges in accurately determining property prices.

02.

Improper pricing can result

03.

Need for a solution to optimize pricing using historical data and predictive analytics

ANALYTIC APPROACH

Perform exploratory data analysis to identify patterns and trends.

Create new features to enhance model performance.

Train models and evaluate using appropriate metrics.

Understanding the Dataset

Data Exploration

Data Cleaning

Feature Engineering

Model Selection

Model Training and Evaluation

Study available dataset and variables affecting housing prices

Rectify missing values, outliers, and invalid data.

Choose models like linear regression, KNN, ridge, decision tree, random forest, and XGBoost.

GOAL



Decision Support:

- Provide actionable insights
- Make the right decisions regarding



Identifying Key Factors:

- Analyze the importance of various features
- Determine the factors that influence changes in house prices in California



Predicting House Prices:

- Develop a regression model to predict median house value based on features.



WHY ML?



Accurate Predictions

Learn complex patterns and trends



Efficiency

Process large amounts of data quickly and efficiently.



Adaptability

Adapt to various types of data and situations.



Complex Factor Identification

Uncover hidden and complex pattern

Data Understanding

- longitude
- latitude
- Housing Median Age
- Total Rooms
- Total Bedrooms
- Population
- Households
- Median Income
- Median House Value
- ocean Proximity

DATA PREPROCESSING

Distribution Analysis

Performing a skewness analysis found no normally distributed numerical features.

Missing Value

Found missing values in the total_bedrooms column, as many as 137 data points.

Outlier

It was found that there were outliers in the feature column. It was decided not to delete the data because it was considered important.

Anomaly

In the median_house_value column there are 678 data with \$500,001 and median_age there are 896 data aged 52 Years.




DATA PREPROCESSING

Duplicate

After performing a duplicate check, it was found that there were no duplicate rows.



Feature Engineering



Add New Feature

rooms_per_household : availability of space in housing

population_per_household : population to number of households

Encoding

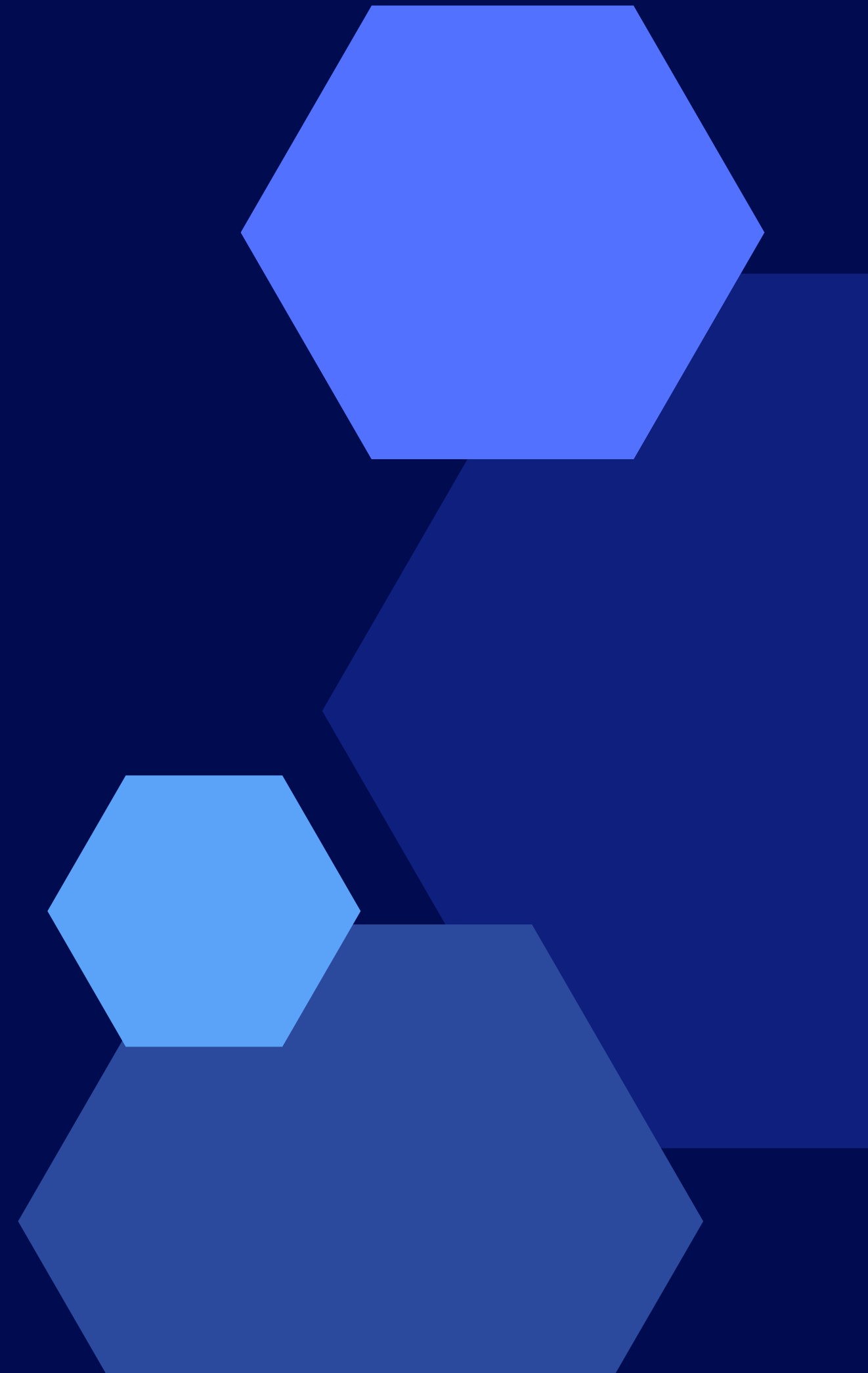
ocean_proximity

Scaling

Robust Scaler



Modelling



Metric Evaluation

MAE

Mean Absolute Error (MAE) is used to measure the average absolute value of the difference between the prediction and the actual value.

MAPE

MAPE measures the average percentage error of the prediction against the actual value. This metric provides an understanding of the extent to which the model can correctly predict housing prices on a percentage scale.

R²

R-squared indicates how well the variability in the data can be explained by the model. The closer it is to 1, the better the model is able to explain the variation in the data. By using a combination of these evaluation metrics, we can measure and understand the model's performance holistically, ensuring that the chosen model can provide accurate and reliable housing price predictions.

BENCHMARK MODEL

Benchmark Model

AdaBoost
Decision Tree
KNN
Gradient Boosting
Random Forest
XGBoost

Best Model

XGBRegressor, Random
Forest, Gradient Boosting

Hyper Parameter Tunning

XGBRegressor =>
MAE : 26662.975565
MAPE : 0.146036
R-2 : 0.812424



XGBREGRESSOR

DEFINITION

XGBRegressor is a machine learning algorithm used for regression tasks, which is part of XGBoost (Extreme Gradient Boosting).

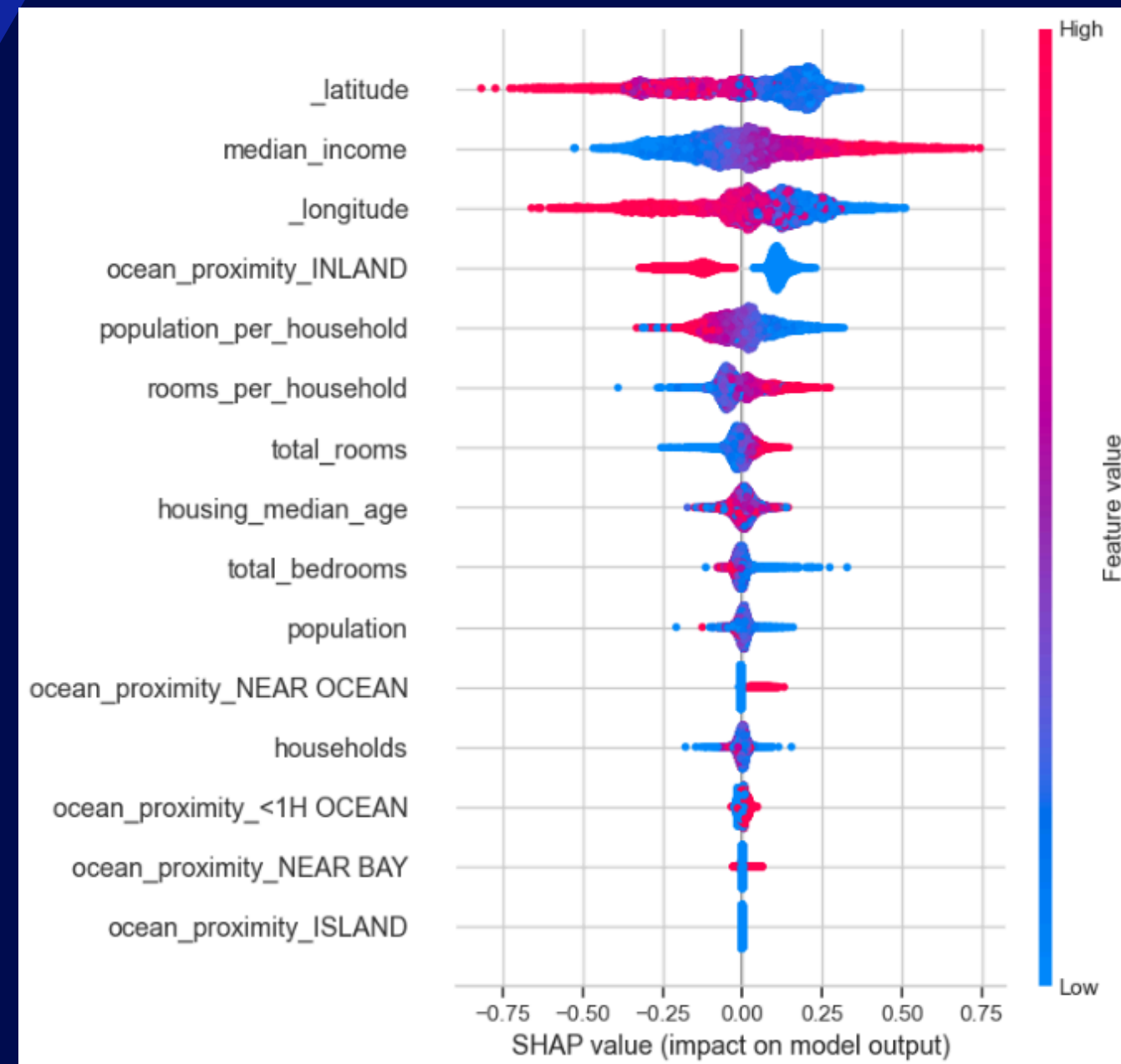
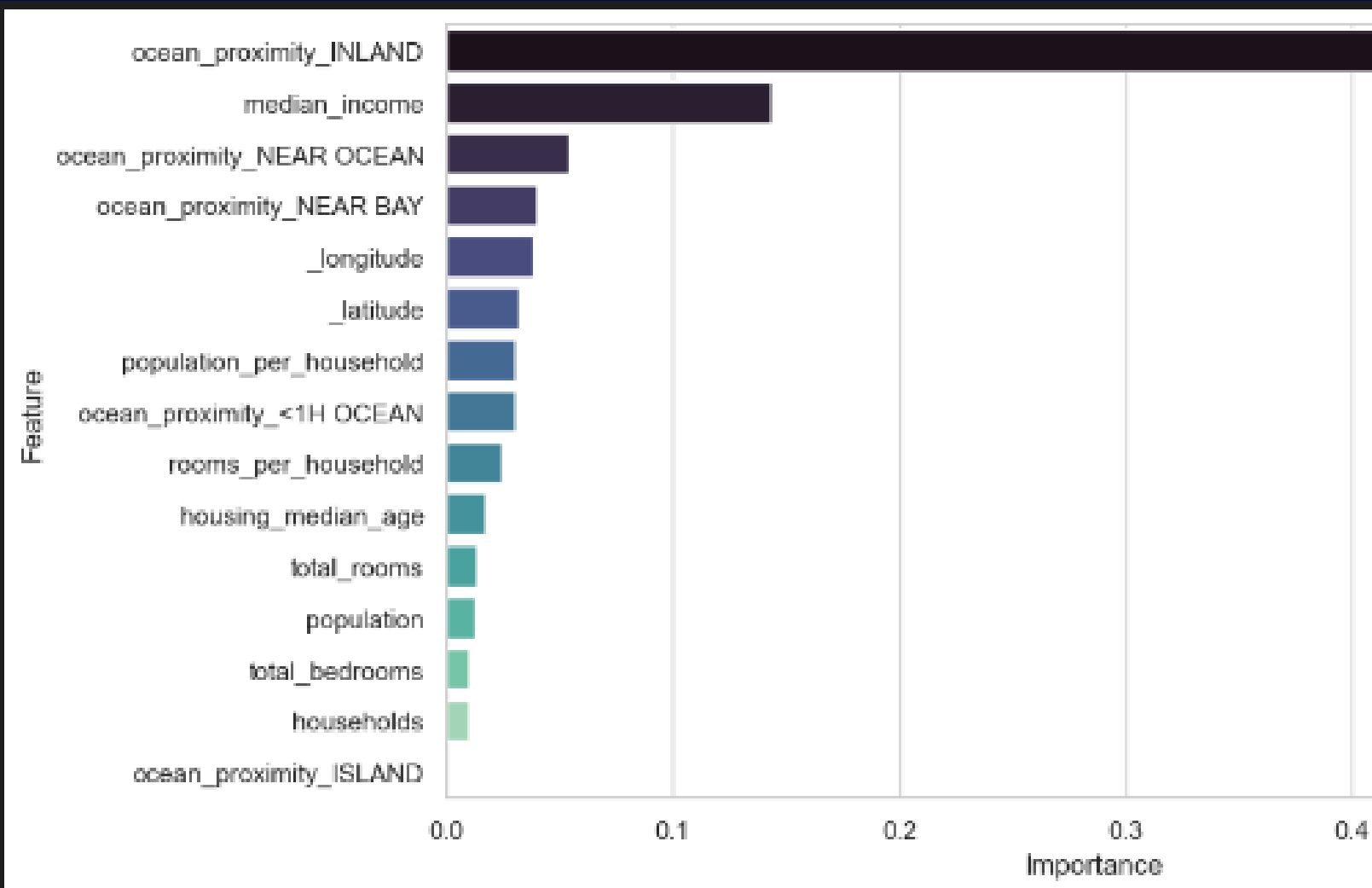
HOW IT WORK

XGBRegressor is a boosting model that uses multiple decision trees to produce accurate predictions. Using a gradient boosting technique, it iteratively builds a new model to reduce the error of the previous model, focusing on the mistakes made by the previous models.

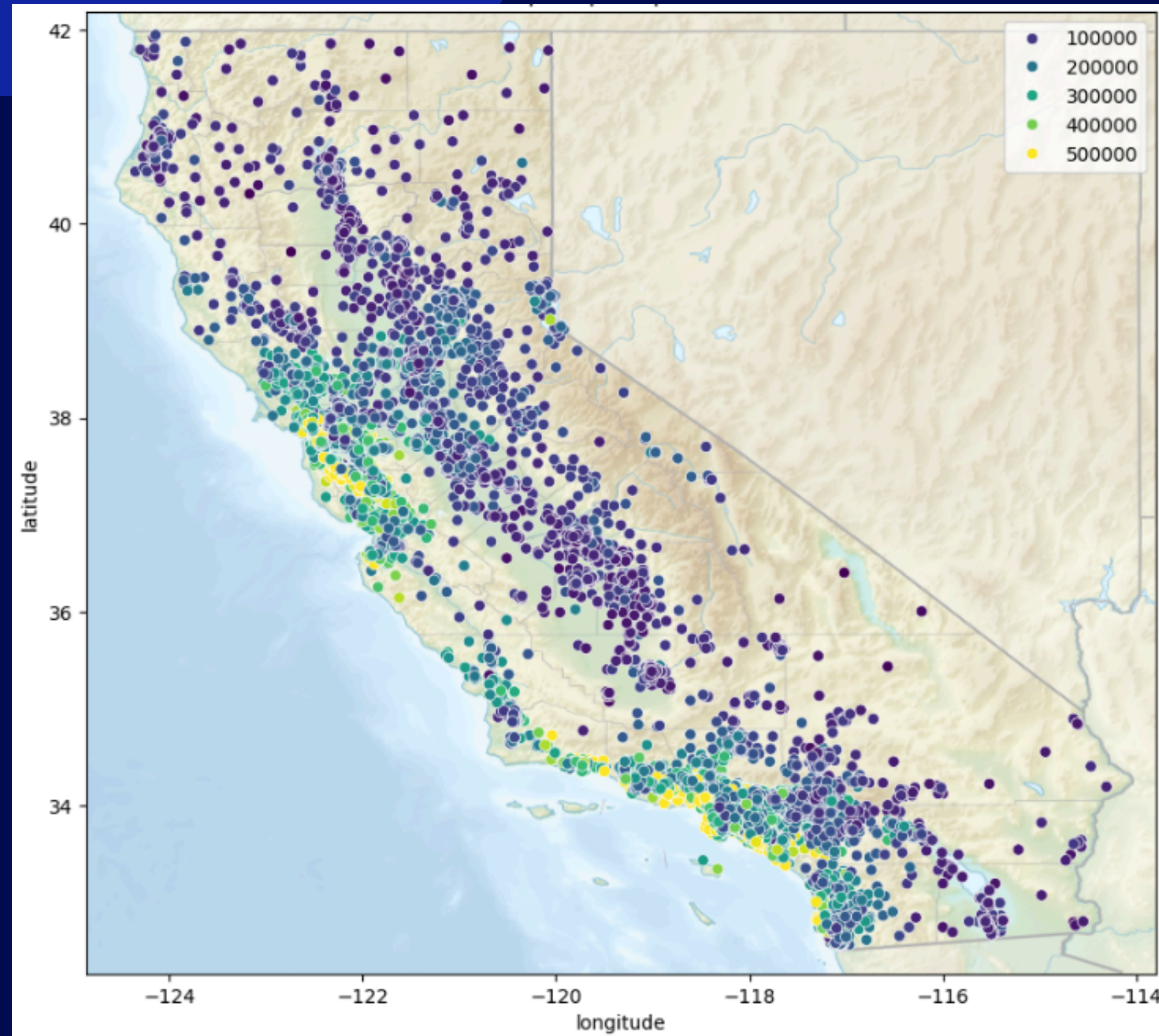
ADVANTAGES

This method is very efficient and can handle large datasets well.

Feature Importance



Feature Importance



Conclusion

MAE	MAPE	R2
26662.975565	14.60%	81.24%

Pricing Strategy

- Price Adjustment
 - Negotiation
-

Development of Real Estate Advisory Services

- Market Analysis
 - Investment Consulting
-

THANK
YOU

