Capstone Project Module 3: California Housing Price

Ivan Taufiqurrahman

Link Video:
https://youtu.be/BvHkUiws-sc

Presentation overview

Background

User/Stakeholders

People who have interest in Real Estate Industry, Machine learning enthusiast.

Analytical Approach

Follows a structured machine learning pipeline such as Data preparation, Transformation, Model Development, Tuning, Final Modeling and Interpretation.

Metric Evaluations

- RMSE
- MAE
- MAPE

Context

Dataset is based on the California Housing dataset from 1990 census. Data Collection method use block group and not individual.

Problem

With block group in this house census dataset, how well machine learning predicting median house value with features and condition back in 1990.

Goals

- Identify features that play the biggest role in predicting house price
- Creating Model that can be a reflection for people to improve their prediction model
- Minimize error in pricing house in California

Data Dictionary

Column Name	Data Type	Description
Latitude	Float	Latitude coordinates of the block group
longitude	Float	Longitude coordinates of the block group
housing_median_age	Float	Median age of the houses within the block group
total_rooms	Float	Total number of rooms across all houses in the block group
total_bedrooms	Float	Total number of bedrooms across all houses in the block group
population	Float	Total population living in the block group
households	Float	Total number of households in the block group
median_income	Float	Median income of households in the block group (US$)
ocean_proximity	Object	Categorical variable describing the block group's proximity to the ocean
median_house_value	Float	Median house value in the block group

Data Cleaning

1. Data Collection

Data_california_house
(14448 rows × 10 columns)

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

2. Data Preprocessing

Load Dataset

Load data_california_house With 14448 rows and 10 columns

Handling Duplicate

No Duplicate

Make a copy

First Dataset (df) is raw dataset.
2nd Dataset (df2) will be cleaned from outliers and Missing Value.

Handling Outliers & Missing Value

df2 -> 12592 rows and 10 columns.
137 Missing Value has been filled.

Multiply by 2

Create df3 as copy of df And df4 as copy of df2.

From this X1 and y1 created from df.
X2 and y2 created from df2, X3 and y3 created from df3, and X4 and y4 created from df2

Modeling

Splitting X1 and y1 into train and test

Transformer use:
OnehotEncoder and RobustScaler

Use Raw Data and RobustScaler

Splitting X2 and y2 into train and test

Transformer use:
OnehotEncoder and StandardScaler

Use Clean Data and StandardScaler

Splitting X3 and y3 into train and test

Transformer use:
OnehotEncoder and StandardScaler

Use Raw Data and StandardScaler

Splitting X4 and y4 into train and test

Transformer use:
OnehotEncoder and RobustScaler

Use Clean Data and RobustScaler

Total Model Use: 12

6 Normal Model:
LinearRegression, KNeighborsRegressor, DecisionTreeRegressor, RandomForestRegressor, XGBRegressor, XGBRFRegressor
6 Transformed Model:
Same but use log for target (median_house_value) and inverse for interpretation

Metric Evaluation: RMSE, MAE, MAPE

Metric Evaluation: Prediction with 3 Best Models before tuning and baseline

From 4 Different Strategy in using Raw Data and Cleaned Data, the best evaluation comes from X1, and y1 (using Raw Data and RobustScaler)

Use X1, y1	RMSE	MAE	MAPE
Baseline (No ML)	114151.30	90142.45	0.631
Normal RandomForest	50386.679414	33337.277391	0.187669
Normal XGBRegressor	48865.848493	32753.710570	0.187000
Transformed XGBRegressor	48203.011286	31652.709776	0.167941

Because of Normal XGBRegressor and Transformed XGBRegressor has little difference in RMSE, MAE, and MAPE, the best approach is to tuning both model

How XGBRegressor Works

XGBRegressor is an advanced implementation of gradient boosting that use decision tree as a base model that always Improves mistakes from previous trees and Corrects errors sequentially. That is why type of algorithm of xgboost is boosting. This is the formula:

$$Fm(x)=Fm-1(x)+\gamma mhm(x)$$

- $Fm-1(x)$ = previous model
- $hm(x)$ = new tree trained on errors that use features to preduce
- γm= weight (learning rate), parameter = learning_rate (default = 0.5)

Example:

House	Median_income	Actual Price
A	2	120K
B	3	180K
C	5	250K

$$F0(x)=mean(y)=183,333$$

House	Median_income	y	F0(x)	Residual	Pattern
A	2	120K	183K	-63K	If Median_income < 3, prediction -60K
B	3	180K	183K	+3K	If Median_income < 4, prediction +0K
C	5	250K	183K	+67K	Else, prediction +60K

$$F1(x)=F0(x)+\gamma1h1(x)$$

- If learning rate γ1=0.1

House	F0(x)	h1(x)	γ1h1(x)	F1(x)
A	183K	-60K	-6K	176K
B	183K	+0K	+0K	183K
C	183K	+60K	+6K	189K

After this, the process repeat like find the new residual and pattern, and then implement that with learning rate 0.1 to how many we want with parameter n_estimator

Business Evaluation: Comparison of After Tuning, Before Tuning, and Baseline

Normal XGBRegressor

	RMSE	MAE	MAPE
Before Tuning	48865.848493	32753.710570	0.187000
After Tuning	45964.023161	30781.209424	0.175253

Transformed XGBRegressor

	RMSE	MAE	MAPE
Before Tuning	48203.011286	31652.709776	0.167941
After Tuning	46564.677258	29966.121437	0.159083

	RMSE	MAE	MAPE
Baseline (No ML)	114151.30	90142.45	0.631

When comparing Normal XGB after tuning with baseline, RMSE has been reduced for about $68K, MAE lower for about $59361, and MAPE also goes down from 63% to 18% or you can say that accuracy is increased from only 37% to 82%.

Business Impact when using 2 models

Minimize error in use case

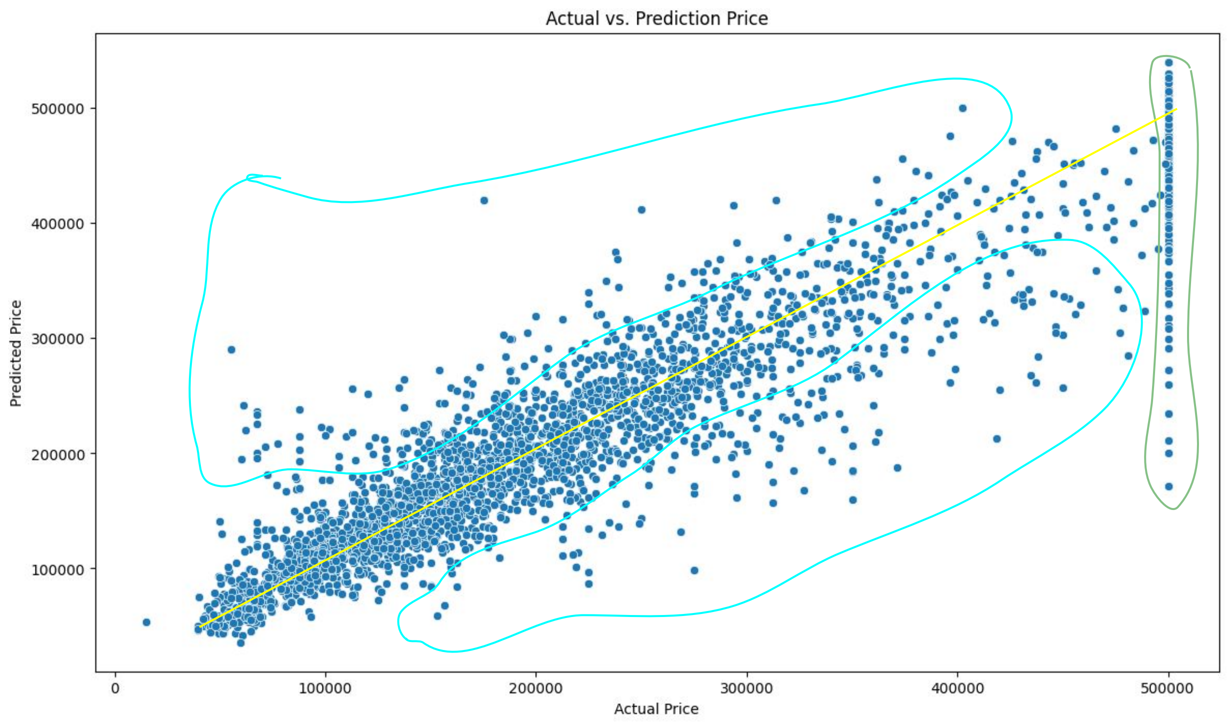
Cheap House has Lower value than $287.5K

Use Transformed XGBRegressor to predict because MAPE is no more than 16% of total value

Luxury House has higher value than $287.5K

Use Normal XGBRegressor to predict because RMSE only $46K

Actual vs Prediction Price with Transformed XGBRegressor



1. **Diagonal Trend**

The strong diagonal pattern shows that the actual prices closely up to around $500,000, which indicates solid performance in this range.

2. **Spread of Predictions**

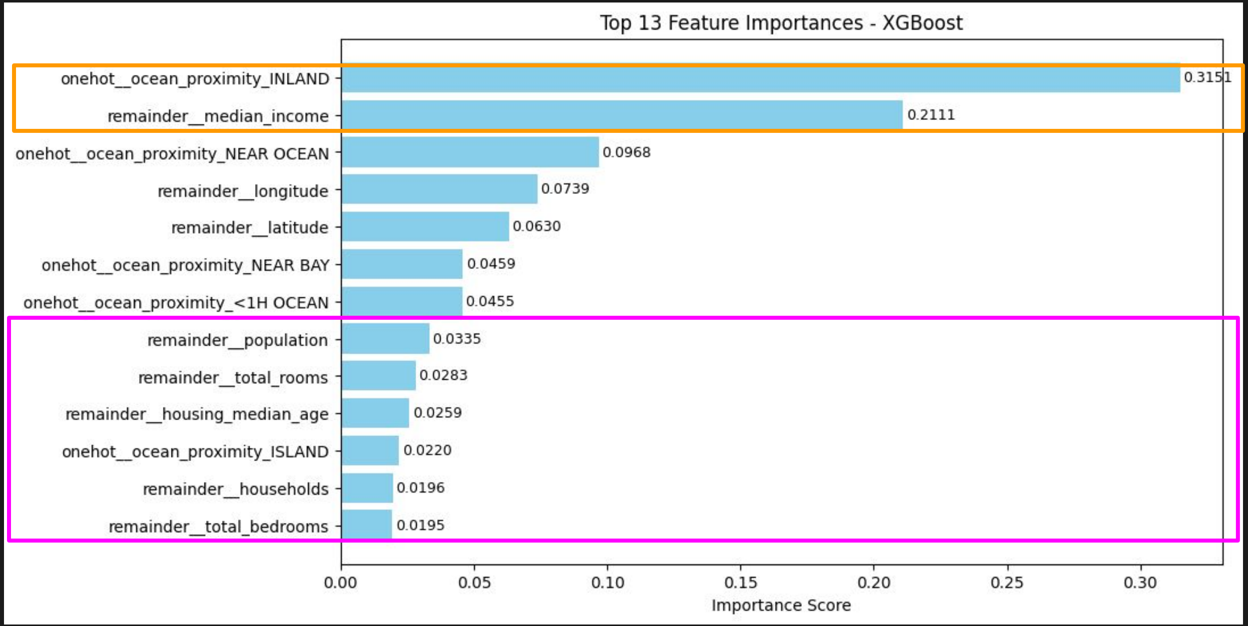
some scatter around the diagonal line reflects noise in the data and highlights that while the model predicts well on average, individual houses may still have noticeable errors.

3. **Capping Effect at $500,000**

Due to a cap in the dataset, the model struggles to predict and producing varied predictions where the target values are artificially fixed.

The model is reliable for estimating homes below the $500,000 threshold but not recommended for higher than $500,000.

Feature Importances

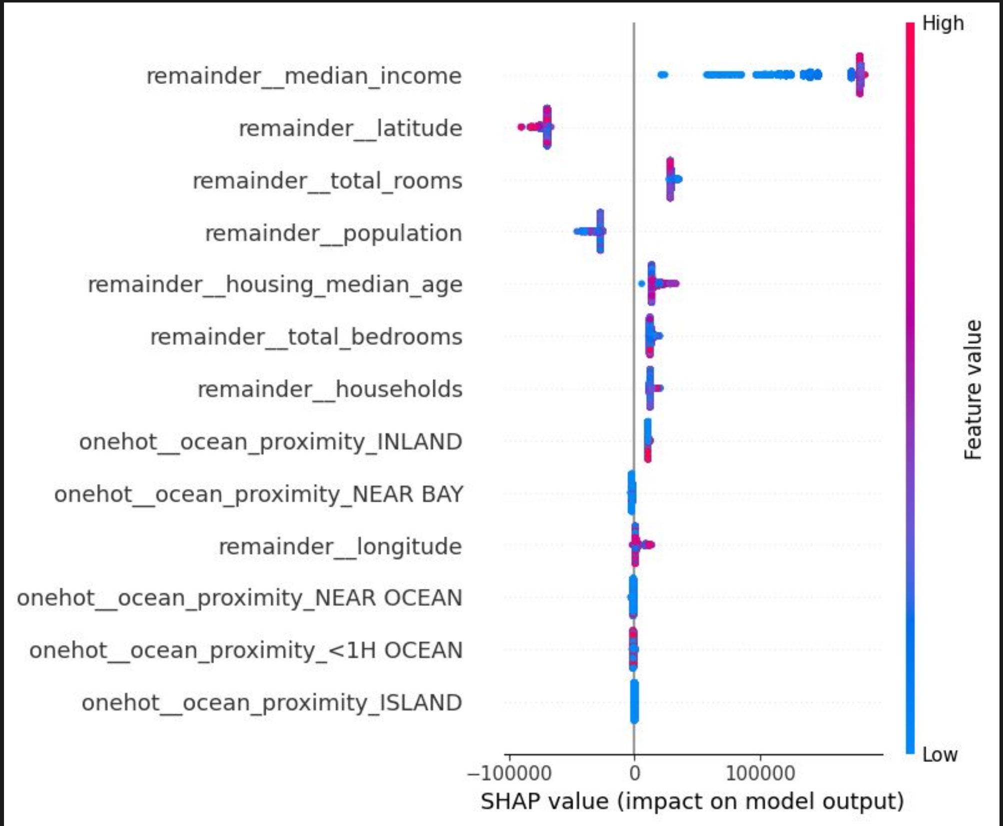


Dominant Drivers
These features capture both geographic desirability and the economic profile of households.

Supporting Factors
Other location-based variables like longitude and latitude also carry weight and further reinforces the importance of geography.

Lesser Contributions
Features such as total rooms, population, and housing median age still contribute but to a much smaller degree.

SHAP Analysis



Key Insights on Median Income
The SHAP plot confirms that median income has a strong positive impact. Higher income values are consistently linked to higher predicted housing prices, and this effect is linear and clear across the data.

Geographic Features
Specially latitude play noticeable roles, though in different directions. SHAP values suggest that being in certain coastal areas increases predicted price, while inland positioning (especially the INLAND category) often lowers it.

Other than 2 dominant features, other features have minimum influence, that means median income in that group block and geographic features like latitude really determine the value of median house value in California back in 1990.

Conclusion

Improved Model vs BASELINE

Both tuned models delivered strong performance that can reduce more than 60% of Baseline RMSE, MAE, MAPE.

3

Models application

models predict within ~15–17% (off) of actual prices on average, but not suitable to predict home values more than $500K.

2

Strong features

Median income and geographic features are the strongest predictors, while other features has small impact to home values.

1

Recommendation

Modeling Improvement

Explore newer models such as LightGBM or deep learning models. Implement ensemble models with base learn XGB, XGBRF, and RF.

01

Feature Improvements

Add more reliable features like economic features, neighbourhood features, and even environmental risk features.

02

Data Collection Improvement

Changing from census group block level to individual property level. Use recent and continuous data like 2010-2025 housing transactions.

03