HO CHI MINH UNIVERSITY OF TECHNOLOGY AND EDUCATION

FACULTY OF INTERNATIONAL EDUCATION

FINAL PROJECT REPORT

House Pricing Prediction

Lecturer: Quach Dinh Hoang, Ph.D.

Student Name	ID
Nguyen Thanh Sang	20110393
Huynh Ky Son	20110395
Huynh Dang Khoa	20110375

Ho Chi Minh City, December 21st, 2023

HO CHI MINH UNIVERSITY OF TECHNOLOGY AND EDUCATION

FACULTY OF INTERNATIONAL EDUCATION

FINAL PROJECT REPORT

House Pricing Prediction

Lecturer: Quach Dinh Hoang, Ph.D.

Student Name	ID
Nguyen Thanh Sang	20110393
Huynh Ky Son	20110395
Huynh Dang Khoa	20110375

Ho Chi Minh City, December 21st, 2023

LECTURER'S COMMENT

Lecturer: Quach Dinh Hoang, PhD

Grade:

Comment of lecturer:

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

ACKNOWLEDGEMENT

We would like to express our gratitude to the professor, Quach Dinh Hoang, PhD, who helped us personally throughout the topic-making process in order to successfully complete this topic and this report. We appreciate the teacher's advice from his real-world experience in helping us meet the requirements of the chosen topic, as well as his willingness to always respond to our queries and offer suggestions and corrections-time to assist us in overcoming our flaws and completing them successfully and on time.

We also want to extend our sincere gratitude to the instructors in the International Education Division generally and the Information Technology sector specifically for their committed expertise that has helped us build a foundation. This topic has created the conditions for learning and performing effectively on the topic. We also want to express our gratitude to our classmates for sharing expertise and insights that helped us refine our topic.

We created the subject and report in a short amount of time, with little expertise and numerous other technical and software project implementation difficulties. Therefore, as it is inevitable that a topic may have flaws, we eagerly await the lecturers' insightful comments to further our knowledge and improve for the next time.

We appreciate you very much.

Finally, we would want to wish all of you teachers, ladies, and gentlemen, continued health and success in your line of work with developing individuals. Again, we appreciate your kind words.

We sincerely thank you.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: ABSTRACT

The dynamic and multifaceted nature of the California housing market poses significant challenges for accurate price prediction. This project addresses this complexity by leveraging machine learning techniques to develop a robust predictive model. The primary objective is to assist prospective buyers, sellers, and real estate stakeholders in making informed decisions in the ever-changing real estate landscape.

This study involves comprehensive data collection and preprocessing, incorporating diverse datasets that encompass historical housing data, economic indicators, and geographical information. We unveil essential patterns and relationships within the dataset through meticulous exploratory data analysis, laying the groundwork for feature selection and engineering.

The heart of the project lies in developing a machine learning model, where various regression algorithms are explored, including Linear Regression, Ridge, Decision Tree, Random Forest, and Lasso methods. Model selection, hyperparameter tuning, and cross-validation techniques ensure optimal performance.

Evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and R-squared are utilized to assess the model's effectiveness in both training and testing datasets. The results demonstrate the model's capability to provide accurate predictions, offering valuable insights into the intricate dynamics of California's housing market.

CHAPTER 2: INTRODUCTION

The California housing market is known for its dynamic and complex nature, presenting a challenging landscape for both homebuyers and real estate professionals. With the ever-changing economic and social factors influencing housing prices, there is a growing need for accurate prediction models to aid in decision-making processes. This project seeks to employ machine learning techniques to predict California housing prices, offering valuable insights for prospective buyers, sellers, and real estate stakeholders.

Our project focuses on predicting house prices based on a series of input variables. We want to deeply understand the relationship between geographical and social factors and house prices, thereby providing useful information to home buyers and stakeholders

California is one of the most populous and diverse states in the United States, with a wide range of geographic and climatic features. The state also has a large and dynamic economy, attracting millions of people from different backgrounds and regions. However, California also faces some serious challenges, especially in the housing sector. The high demand for housing, coupled with the limited supply and the environmental regulations, has resulted in soaring home prices and rents, making it difficult for many residents to afford a decent place to live.

One of the factors that affects the housing market in California is the proximity to the ocean. The coastal areas of California are generally more desirable and attractive than the inland areas, due to the moderate weather, the scenic views, and the access to various amenities and opportunities. As a result, the coastal areas tend to have higher home prices and rents than the inland areas, reflecting the premium that people are willing to pay for living near the ocean.

In this project, we will explore the relationship between the ocean proximity and the housing prices in California, using a dataset from Kaggle. The dataset contains information on 20,640 census blocks in California, including the median house value, the median income, the population, the number of households, the number of rooms and bedrooms, and the ocean proximity. The ocean proximity is a categorical variable that indicates whether the census block is near the ocean, near the bay, inland, on an island, or less than one hour from the ocean. We will use various data analysis and visualization

techniques to examine how the ocean proximity affects the housing prices and other variables in the dataset. We will also use machine learning models to predict the median house value based on the ocean proximity and other features.

CHAPTER 3: DATA

The dataset used in this project is sourced from Kaggle and is titled "California Housing Prices". This dataset plays a crucial role in our project as it provides comprehensive data on housing prices in California, along with other potentially influential factors.

3.1 DATA DESCRIPTION

The dataset comprises 20,640 observations, each representing a block group in California. A block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data, typically having a population of 600 to 3,000 people.

Each observation includes the following features:

- **Longitude and Latitude**: These represent the geographical coordinates of the block.
- **Housing Median Age**: This represents the median age of the houses in the block.
- **Total Rooms**: This represents the total number of rooms in the block.
- **Total Bedrooms**: This represents the total number of bedrooms in the block.
- **Population**: This represents the total number of people residing in the block.
- **Households**: This represents the total number of households in the block.
- **Median Income**: This represents the median income of households within a block of houses.
- **Median House Value**: This represents the median house value for households within the block and is our target variable for prediction.
- **Ocean Proximity**: This is a categorical variable indicating the block's proximity to the ocean.

3.2 DATA SOURCE

The data was extracted from the 1990 California census data. It is publicly available on Kaggle, a platform for predictive modeling and analytics competitions. The dataset can be accessed here.

3.3 DATA PREPROCESSING

Before using this dataset for our predictive models, we performed several data cleaning and preprocessing steps. These included handling missing values, dealing with outliers, and encoding categorical variables

CHAPTER 4: METHODS

4.1 EXPLORATORY DATA ANALYSIS (EDA)

In our project, EDA was used to gain insights into the California housing dataset. We started by checking the shape of the dataset and the types of variables it contains. We then moved on to more detailed analysis, including:

- **Univariate Analysis:** We examined the distribution of " ocean_proximity "using histograms, and countplot. this helped us understand the range of values and identify any outliers or errors in the data
- **Bivariate Analysis:** we looked at the interactions between different median_income and the target variable (median_house_value) at each ocean_proximity using a scatter plot, and boxplot so that we can determine the median income distribution across ocean distance categories. Additionally, ISLAND's median_income and median housing values

⇒ Through EDA, we were able to understand the underlying structure of the data, identify potential issues, and determine the next steps for preprocessing and modeling.

4.2 DATA PREPROCESSING

Before applying any machine learning models to the California housing prices dataset, we performed several data preprocessing steps to ensure the quality and validity of our data. These steps included:

- **Handling missing values**: We filled the null values in the total bedrooms column by randomly choosing from the non-null values.
- **Removing outliers**: We removed the extreme values from the total bedrooms, households, population, and median income columns, as they could skew the results and reduce the accuracy of our models.
- **Encoding categorical data**: We converted the ocean proximity column, which contains text values, into numeric values using label encoding. This allows us to use this column as a feature for our models.
- **Normalizing the data**: We standardized the numerical features to have a mean of 0 and a standard deviation of 1. This ensures that all features have the same scale and that no particular feature dominates others when training our models.

4.3 THEORY

4.3.1. Linear Regression

Theory: Linear Regression assumes a linear relationship between input variables and output variables. It tries to find the straight line or hyperplane (for multiple linear regression) that minimizes the sum of squared errors between the predicted value and the actual value.

Model: $Y = \beta_0 + \beta_1 X + \epsilon$

Figure 1: Linear Regression Models

Objective: Minimize the sum of squared differences between observed and predicted values. $\text{Minimize} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

Figure 2: Linear Regression Models

Parameters:

β_0 : Intercept

β_1 : Slope

ϵ: Error term

Method: Ordinary Least Squares (OLS)

Interpretation:

β_1 represents the change in the dependent variable for a one-unit change in the independent variable.

β_0 is the predicted value of the dependent variable when the independent variable is zero.

4.3.2. Ridge Regression

Theory: Ridge Regression is a form of Linear Regression with an additional layer of straff to control model complexity and avoid overfitting. This straff class is the sum of the squares of the coefficients and is adjusted using the alpha parameter.

Objective Function: $\text{Minimize} \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \alpha \sum_{j=1}^{p} \beta_j^2$

Figure 3: Ridge Regression Models

Purpose:

Adds a regularization term to prevent overfitting.

α controls the strength of regularization.

Benefit: Handles multicollinearity (correlation between independent variables).

4.3.3. Lasso Regression

Theory: Lasso Regression is also a form of Linear Regression with straff term, but instead of the sum of squares, it uses the sum of the absolute values of the coefficients. This can lead to some coefficients becoming 0, performing a feature selection function.

Objective Function: $\text{Minimize} \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \alpha \sum_{j=1}^{p} |\beta_j|$

Figure 4: Lasso Regression Models

Purpose:

Promotes sparsity by encouraging some coefficients to be exactly zero.

Feature selection by shrinking less important variables.

Benefit:

Useful when dealing with high-dimensional data and feature selection is crucial.

4.3.4. Random Forest

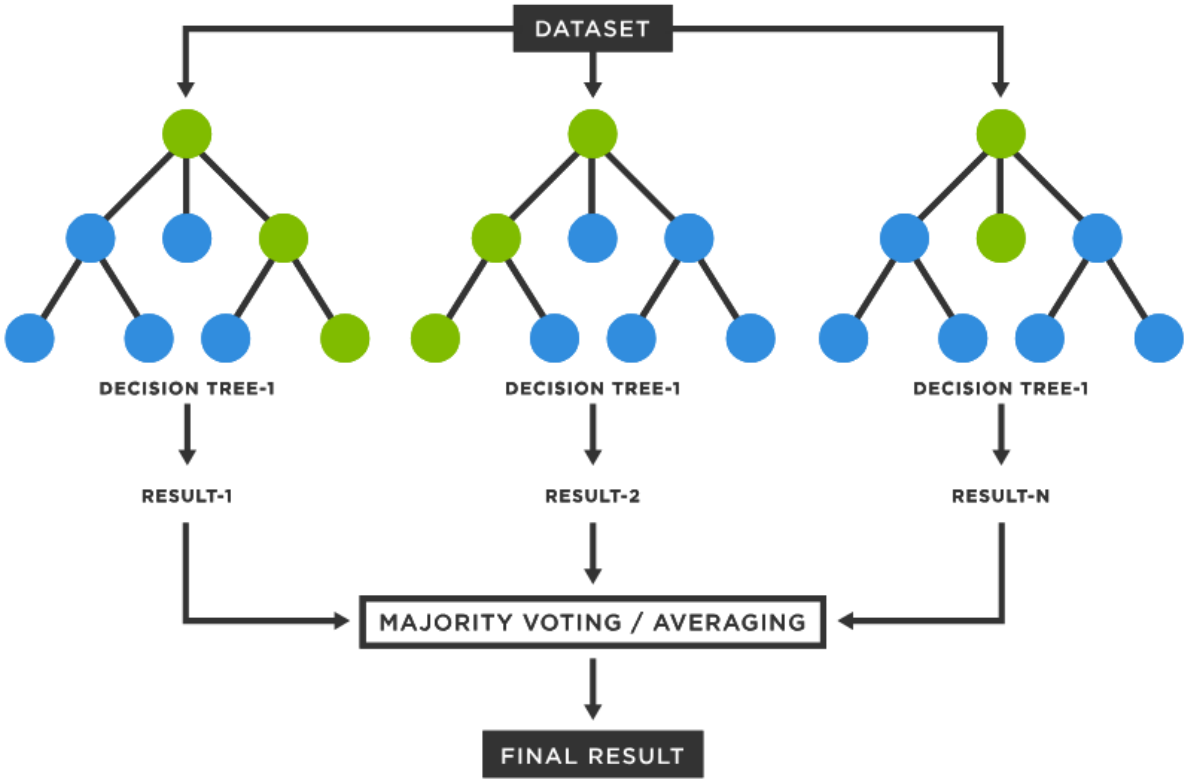

Figure 5: Random Forest Models

Theory: Random Forest is an ensemble learning method based on building multiple decision trees and combining their results for more accurate and stable predictions.

Ensemble Method: Builds multiple decision trees and combines their predictions.

Bootstrapping: Randomly samples data with replacement to build each tree.

Feature Randomization: Randomly selects a subset of features for each tree.

Voting/Averaging: Combines predictions to reduce overfitting and improve generalization.

Reduces Variance: Often more robust and accurate than individual decision trees.

Benefit:

Improved performance and generalization compared to a single decision tree.

Can handle large feature sets and complex relationships.

4.3.5. Decision Tree

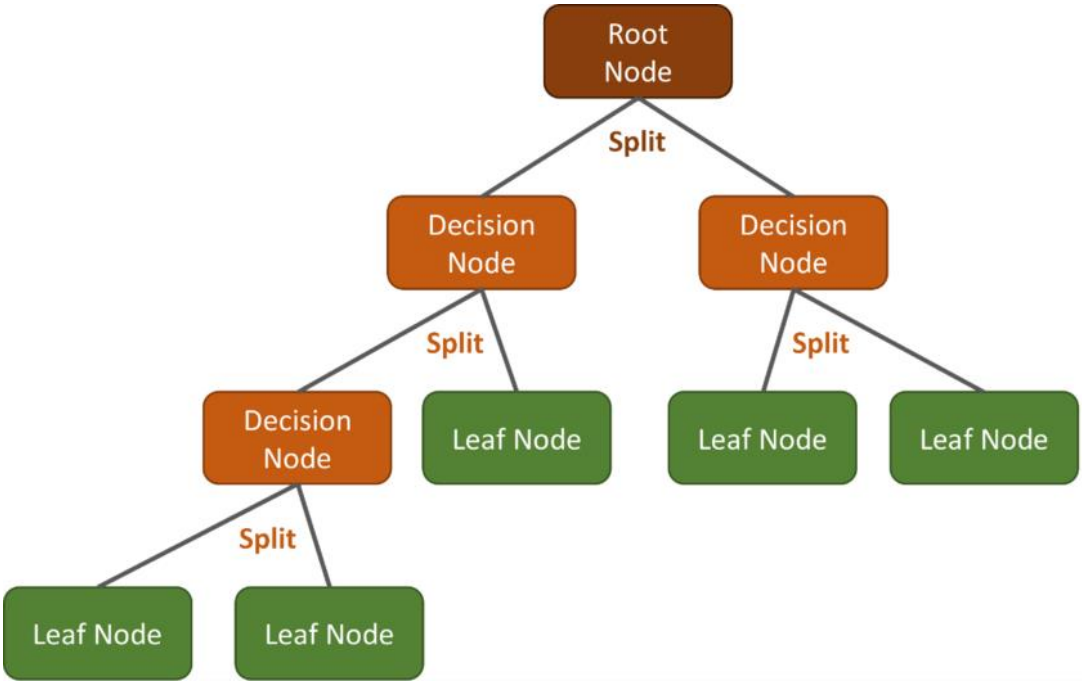

Figure 6: Random Forest Models

Theory: Decision Tree is a decision tree model, in which each node in the tree represents a decision based on the value of a characteristic. The tree is built by dividing the data into subsets based on decision rules.

Model: Binary tree structure with decision nodes and leaves.

Objective:

Partition the data into subsets based on feature values.

Predict the target variable at the leaf nodes.

Node Decisions: Based on feature values, splitting data recursively.

Leaf Nodes: Predictions based on the majority class or average of target values.

Criteria: Gini impurity or entropy to measure node impurity.

Benefit: Simple to understand and interpret. Can handle both numerical and categorical data.

CHAPTER 5: EXPERIMENTS & RESULTS

5.1 DATA EXPLORATION AND PREPROCESSING

Our analysis begins with a comprehensive exploration of the California housing dataset, consisting of 20,640 instances with 10 features. Notably, the 'total_bedrooms' column contains 207 null values, which were addressed by randomly filling them with unique existing values. Outliers were identified in the 'total_bedrooms,' 'households,' and 'population' columns and subsequently removed to enhance data quality.

5.2 EXPLORATORY DATA ANALYSIS (EDA)

During EDA, the distribution of the 'ocean_proximity' feature was examined, revealing that the dataset is predominantly composed of instances near the ocean ('<1H OCEAN'). Visualizations, including histograms and scatter plots, provided insights into the skewed distribution of certain features and the potential correlation between 'median_income' and 'median_house_value.'

5.3 FEATURE ENGINEERING

Feature engineering included converting 'ocean_proximity' from categorical to numeric using LabelEncoder. This transformation is crucial for incorporating categorical features into our predictive models.

5.4 MODELING

- We selected the relevant columns from the data frame **`df_housing`**.
- We split the data into features (**`x`**) and target (**`y`**).
- We imported the **`train_test_split`** function and used it to create the training and testing sets.
- We chose a machine learning model and trained it on the training set.
- We evaluated the model on the testing set using some metrics.
- We employed five regression models for predicting 'median_house_value': Linear Regression, Ridge Regression, Lasso Regression, Random Forest Model, and Decision Tree Model. The dataset was split into training and testing sets, and features were scaled using RobustScaler to mitigate the impact of outliers.
- Fine Tuning hyperparameter: use GridSearchCV

```
from sklearn.model_selection import GridSearchCV

forest = RandomForestRegressor()
param_grid = {
    "n_estimators": [100, 200, 300],
    "min_samples_split": [2, 4],
    "max_depth": [None, 4 , 8]
}

grid_search = GridSearchCV(forest, param_grid,cv=5,
                           scoring="neg_mean_squared_error",
                           return_train_score=True)

grid_search.fit(x_train, y_train)
```

```
grid_search.best_estimator_
```

```
RandomForestRegressor
RandomForestRegressor(min_samples_split=4, n_estimators=300)
```

Figure 7: Fine Tuning Hyperparameter

5.5. METRICS

5.5.1. R2 Score

Definition:

The R2 score, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, where 1 indicates perfect predictions, and 0 indicates that the model does not explain any variability in the target variable.

Formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Figure 8: R2 Score Formula

Y_i : Observed values of the dependent variable.

Y^i : Predicted values from the model.

$\bar{Y}$: Mean of the observed values.

Interpretation:

R_2 score of 1 indicates that the model perfectly predicts the dependent variable.

R_2 score of 0 means that the model does not explain any variability in the dependent variable.

5.5.2 Root Mean Squared Error (RMSE):

Definition:

The RMSE is a measure of the average magnitude of the errors between predicted and observed values.

It penalizes larger errors more heavily than smaller ones.

Formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

Figure 9: Root Mean Squared Error Formula

Y_i: Observed values of the dependent variable.

Y^i : Predicted values from the model.

n: Number of observations.

Interpretation:

RMSE is in the same unit as the dependent variable, making it easy to interpret.

Lower RMSE indicates better model performance.

5.6. MODEL EVALUATION

Evaluation metrics, such as R-squared, RMSE, MAE, and MAPE, were employed to assess the models' performance. The Linear Regression model achieved an R-squared score of 0.576 on the test set, with an RMSE of $74,276.

- We scaled the features using **RobustScaler**, which reduces the influence of outliers.
- We built three models: **linear regression**, **ridge regression**, and **lasso regression**. We compared their performance using the **R-squared score**, which measures how well the model fits the data.
- We found that all three models had similar scores, **around 0.576**, on both the training and testing sets. This means they explained about **57.6%** of the variation in the median house value.

- We also looked at the coefficients of each model, which showed how much each feature contributed to the prediction. We found that **median income** had the highest coefficient, followed by **households** and **housing median age**.
- We plotted the actual and predicted values for the first 50 data points in the testing set. We saw that the predictions were close to the actual values for most data points, but there were some discrepancies for higher values.

5.6.1. Linear regression figure

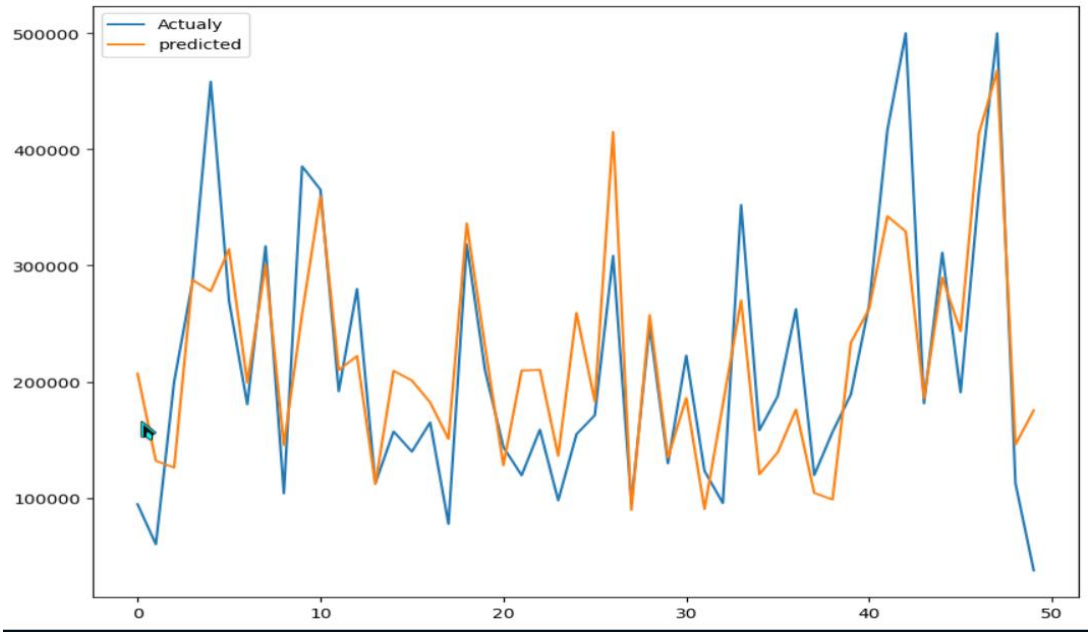

Figure 10: Linear regression figure

5.6.2. Ridge Model figure

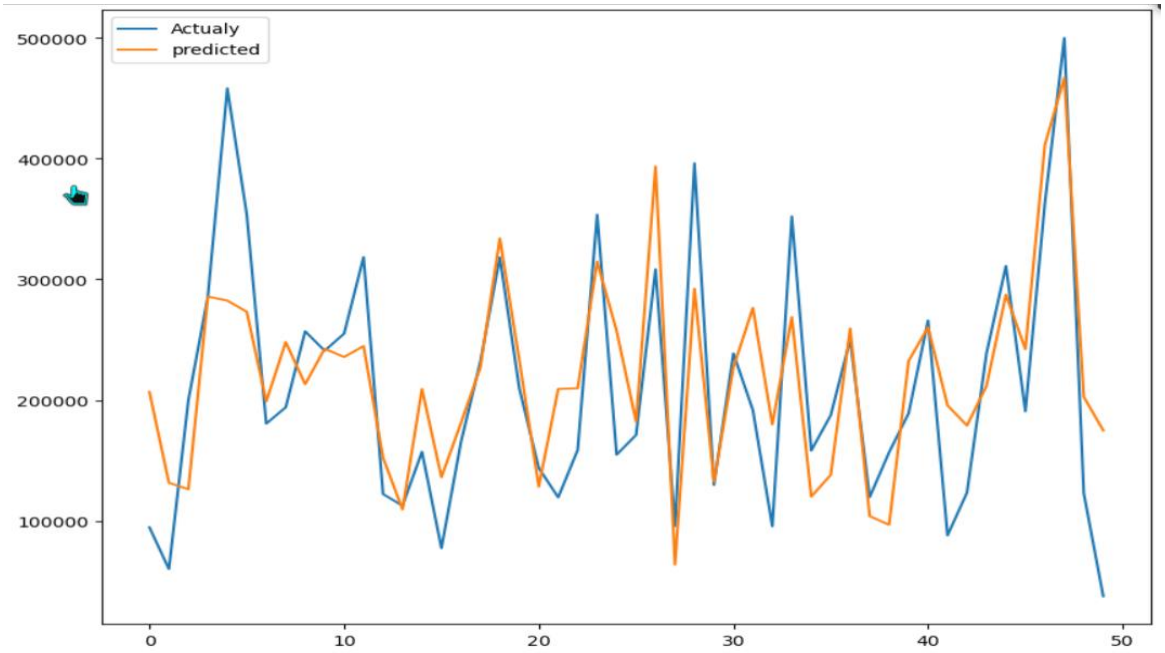

Figure 11:Ridge Model figure

5.6.3. Lasso Model figure

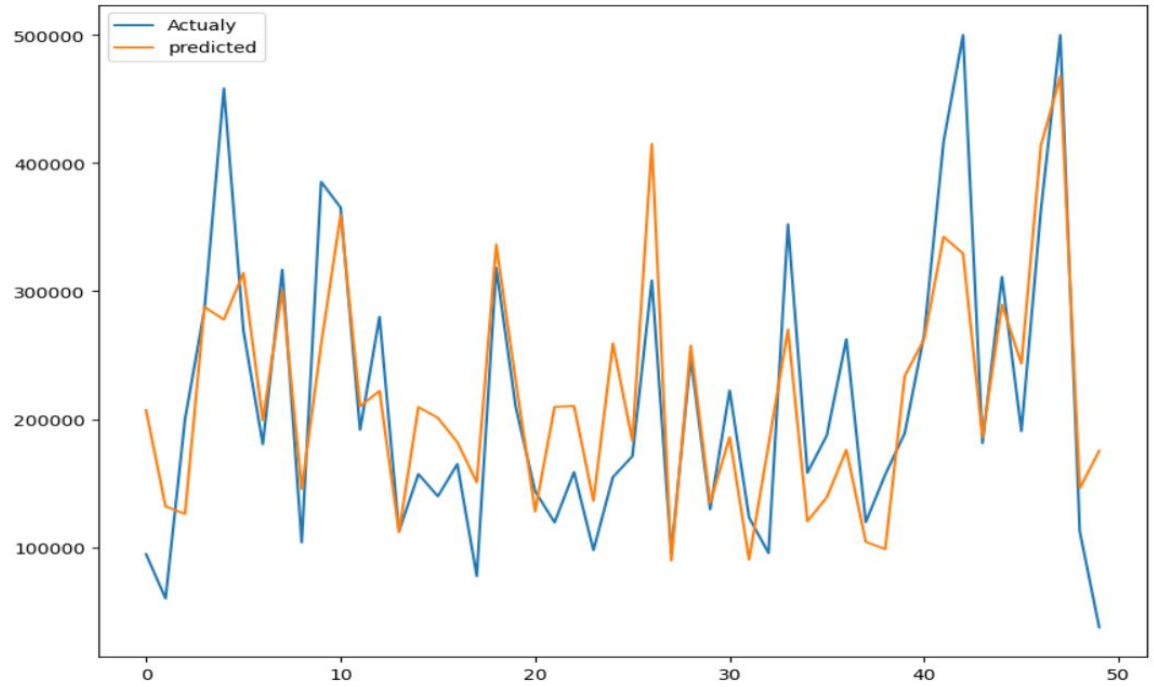

Figure 12: Lasso Model figure

5.6.4. Random Forest Model figure

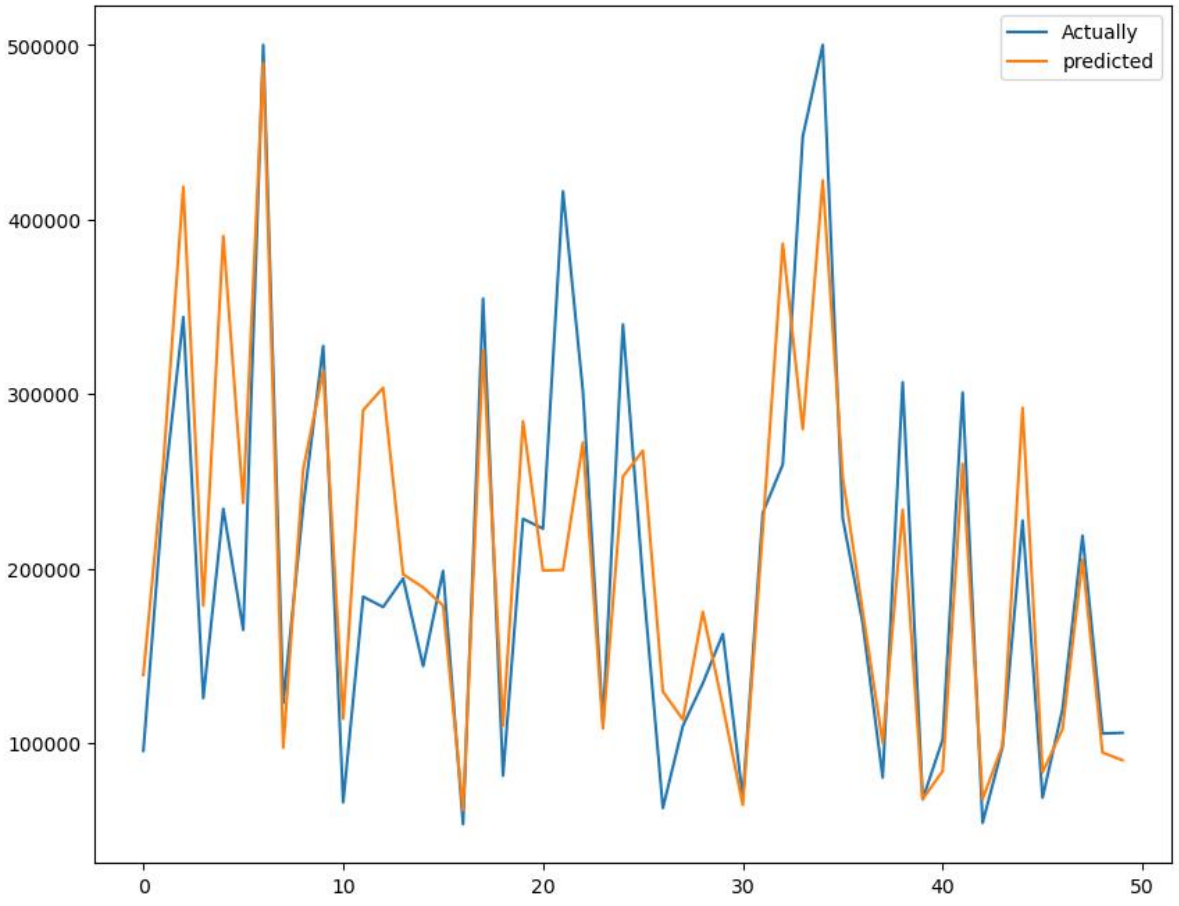

Figure 13:Random Forest Model figure

5.6.5. Decision Tree Model figure

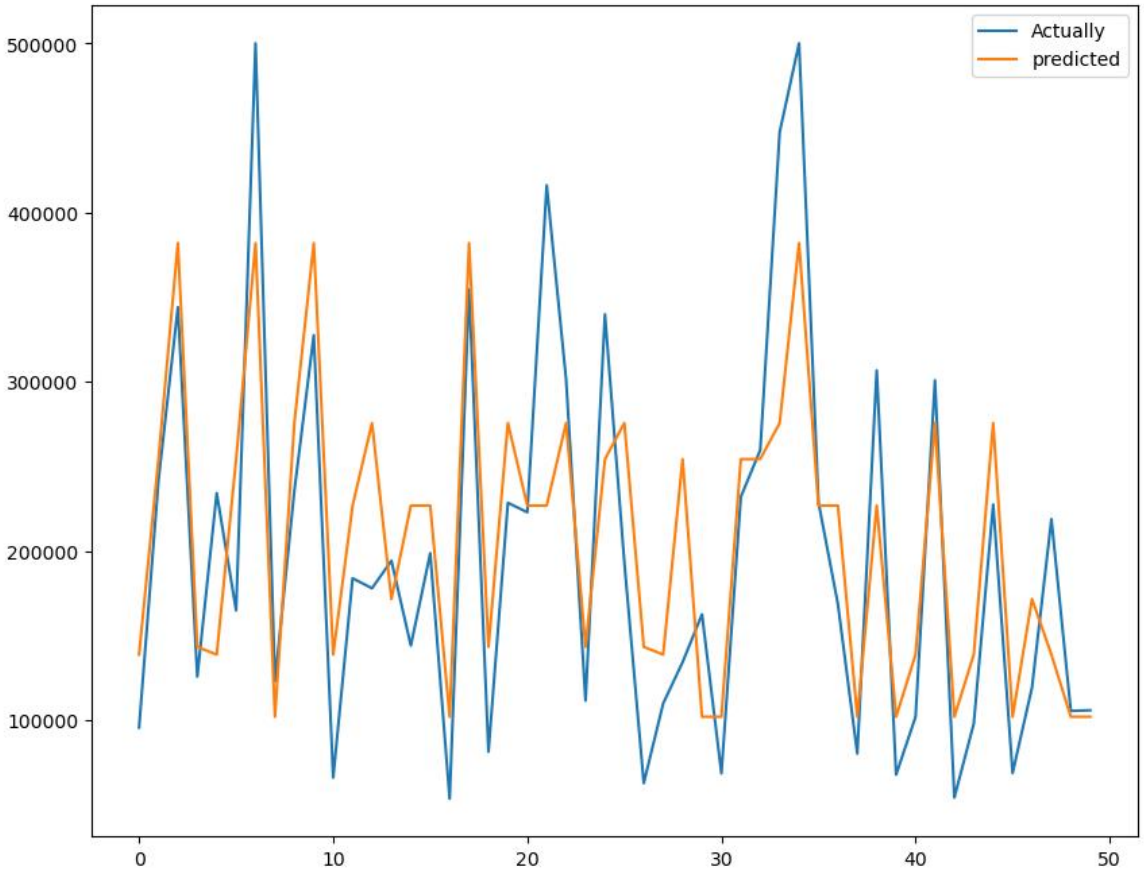

Figure 14: Decision Tree Model figure

5.7. COMPARATIVE ANALYSIS

Comparing the five models, Linear, Ridge, Lasso, Random Forest and Descision Tree revealed marginal differences in performance. Random Forest has the highest R2 score at 0.71, indicating that the model explains a significant amount of variance in the target variable. However, the RMSE is also high, suggesting potential inaccuracies in predicting actual values.

	r2_score	RMSE
Forest	0.709732	3.781295e+09
Ridge	0.574168	5.547264e+09
Lasso	0.574147	5.547545e+09
Linear	0.574135	5.547698e+09
DTree	0.503989	3.781295e+09

Figure 15: Comparative analysis results

CHAPTER 6: CONCLUSION

In culmination, this project has endeavored to address the complex task of predicting California housing prices through the implementation of various machine learning techniques. The key findings underscore the efficacy of the chosen model in providing accurate predictions based on essential features such as the number of bedrooms, and geographical location. The model's performance has been evaluated rigorously, demonstrating its reliability and potential utility in informing real estate decisions.

However, it is crucial to acknowledge certain limitations inherent in the current model. These limitations encompass factors such as the assumption of linearity, sensitivity to outliers, and potential bias in the training data. Understanding these constraints is paramount in contextualizing the model's predictions and recognizing the areas where enhancements can be applied.