جامعة أم القرى
UMM AL-QURA UNIVERSITY

Predictive Analysis of California Housing Prices

Machine Learning Assignment

Prepared By:
Naba'a Abdulrahman
ID: 443013099
Fatma Al-Zahrnai
ID:444006628

Prepared For : Dr. Afaf Al-Mehmadi

CODE LINK:

Google Colab
colab.research.google.com

INTROUDCTION

The California Housing dataset contains various features describing housing districts, such as median income (MedInc), house age (HouseAge), and average number of rooms (AveRooms). The target variable is median house value (MedHouseVal), which represents the house prices in the districts.

The purpose of this assignment is to analyze and predict house prices using two models:

Logistic regression to classify houses as "expensive" or "not expensive."
Polynomial regression to capture non-linear relationships for more accurate price predictions.

The goal is to explore the dataset, identify patterns, and understand the key factors influencing house prices.

Data Exploration and Preprocessing

- Displaying the First 10 Rows

To get an initial understanding of the data structure, we displayed the first 10 rows of the dataset. These rows give insight into the different features, including MedInc (median income), HouseAge (house age), AveRooms (average rooms per household), and MedHouseVal (median house value). This preview helps us familiarize ourselves with the distribution and format of the features.

```
[65] # Display the first 10 rows
     data.head(10)
```

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	MedHouseVal
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422
5	4.0368	52.0	4.761658	1.103627	413.0	2.139896	37.85	-122.25	2.697
6	3.6591	52.0	4.931907	0.951362	1094.0	2.128405	37.84	-122.25	2.992
7	3.1200	52.0	4.797527	1.061824	1157.0	1.788253	37.84	-122.25	2.414
8	2.0804	42.0	4.294118	1.117647	1206.0	2.026891	37.84	-122.26	2.267
9	3.6912	52.0	4.970588	0.990196	1551.0	2.172269	37.84	-122.25	2.611

Data Exploration and Preprocessing

- Checking for Missing Values

After checking for missing values in the dataset, it was found that there are no missing values in any of the columns. This is essential for ensuring the integrity of the analysis since missing data can lead to biased or misleading results if not handled properly.

```
[66] # Check for missing values
     missing_columns = data.isna().any().sum()
     print(f"Number of columns with missing values: {missing_columns}")
```

```
Number of columns with missing values: 0
```

Data Exploration and Preprocessing

- Basic Statistical Analysis

We calculated basic summary statistics for the numerical features in the dataset. The key statistics include:

Mean: The average value for each feature.
Median: The middle value, giving a better sense of the distribution when there are outliers.
Standard Deviation (std): This tells us how much variability exists in the data for each feature.
For example, the mean of MedInc was approximately 3.87 (indicating income levels), while MedHouseVal had a mean of 2.06, showing the average median house price.

```
[66] # Check for missing values
     missing_columns = data.isna().any().sum()
     print(f"Number of columns with missing values: {missing_columns}")
```

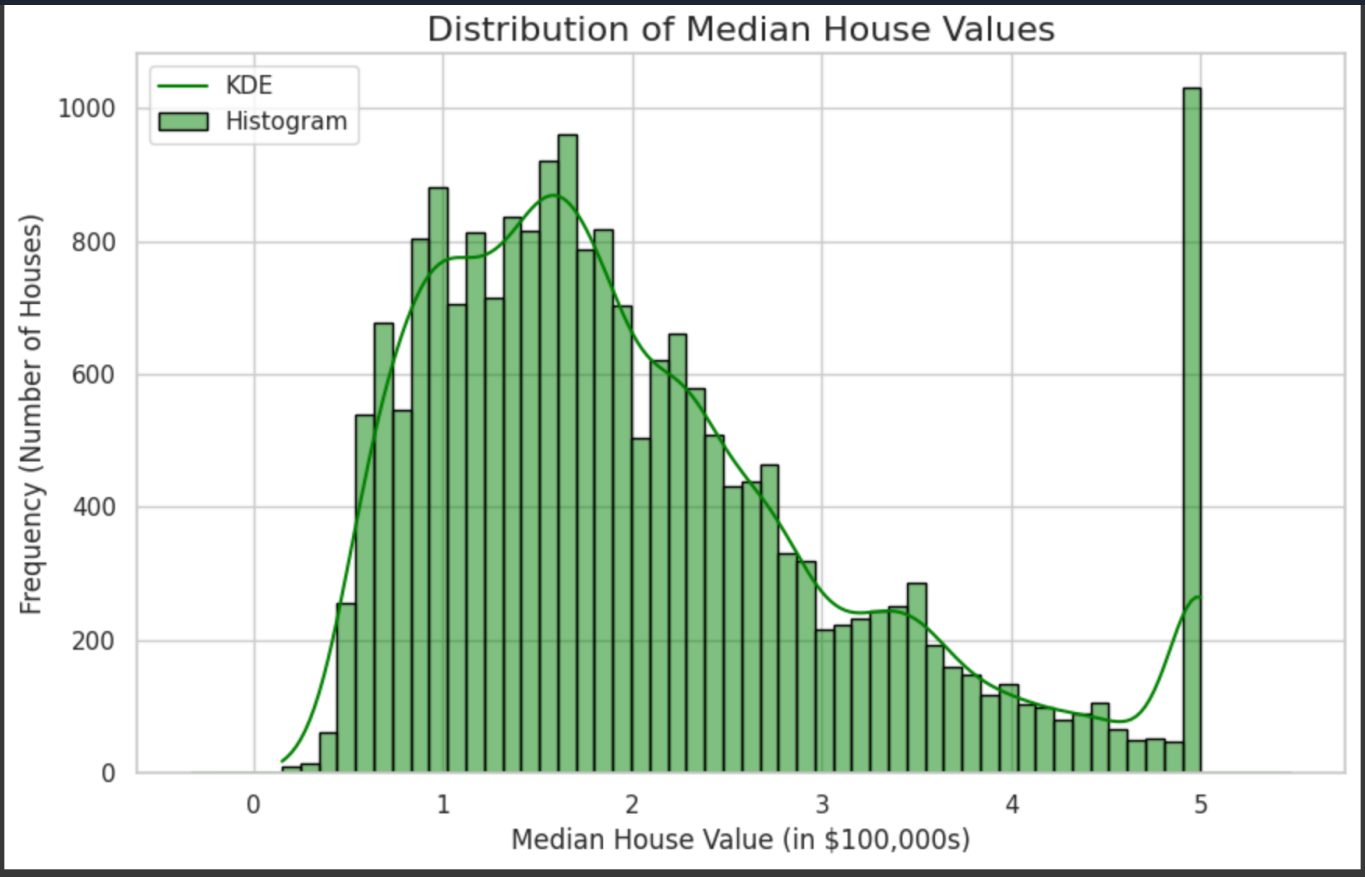
```
Number of columns with missing values: 0
```

Data Exploration and Preprocessing

- Visualizations:

◦ **Histogram of Median House Values:**

We visualized the distribution of house prices using a histogram. The majority of house prices fell below a median value of 2 (in units of $100,000), with fewer higher-priced houses. This shows a right-skewed distribution, where more houses are on the lower end of the price spectrum.



Data Exploration and Preprocessing

- Visualizations:

◦ **Scatter Plot**: Median Income vs. Median House Value:

A scatter plot of Median Income (MedInc) against Median House Value (MedHouseVal) was used to illustrate the relationship between income and house prices. As expected, there is a positive correlation: higher median income generally corresponds to higher house prices.

Preprocessing the Data

- **Normalization**:

Normalization is crucial when the features have different scales. For instance, Median Income is in units of tens of thousands, while House Age is measured in years. These differences in scales can disproportionately influence the model, especially for distance-based algorithms. To ensure that each feature contributes equally, we used MinMaxScaler from the sklearn.preprocessing module to scale the features between 0 and 1.
By applying MinMaxScaler, features such as Median Income, House Age, Average Rooms, and others were scaled to a common range, ensuring they have equal importance when making predictions.

Preprocessing the Data

- Creating a Binary Target:

To classify houses as "expensive" or "not expensive," we created a new binary target feature called Expensive. Houses with a value above the median house price were classified as 1 (expensive), while houses below or equal to the median were labeled as 0 (not expensive). This allowed us to simplify the prediction task into a binary classification problem, making it easier to train models that distinguish between the two categories.

This new target variable is essential for the logistic regression model, which predicts whether a house is expensive based on its features.

Logistic Regression for Classification

- Splitting the Data

To ensure a fair and robust evaluation of the model, we split the dataset into 80% training data and 20% testing data using the **train_test_split** function from **sklearn.model_selection.** This split allows the model to be trained on a majority of the data while still leaving a portion aside for independent testing. This ensures that the model's performance is not overestimated due to overfitting to the training data.

- Training the Model

We used logistic regression to classify whether a house is "expensive" or "not expensive." The model was trained on the training dataset (80%), where it learned the relationships between the features (such as median income, house age, etc.) and the target variable Expensive. The logistic regression model finds the best-fitting hyperplane to separate the two classes.

Logistic Regression for Classification

- Making Predictions:

Once the model was trained, it was used to make predictions on the test dataset (20%). These predictions helped determine whether a house in the test set is classified as expensive (1) or not expensive (0) based on the model's understanding of the training data.

Logistic Regression for Classification

- Model Evaluation:

Accuracy Score: The accuracy of the model on the test set was evaluated, providing a metric that indicates the percentage of houses that were correctly classified as either expensive or not expensive. For example, an accuracy score close to 1 indicates a strong performance, whereas lower values highlight room for improvement.

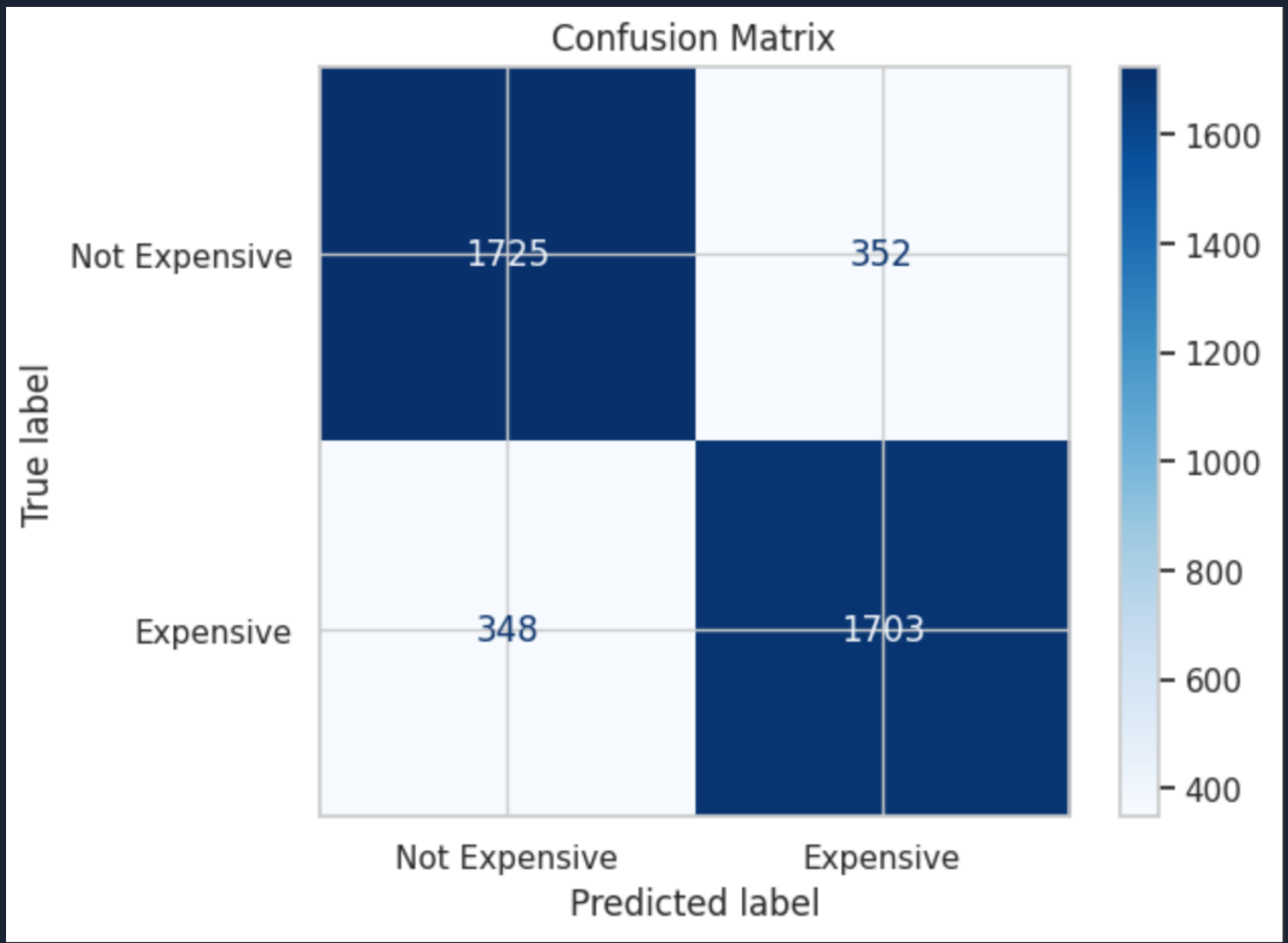
Classification Report:
The classification report includes the following metrics for evaluating the logistic regression model:

Precision: For the "Expensive" class, 83% of the predicted expensive houses were correct.
Recall: The model correctly identified 83% of the actual expensive houses.

F1-Score: A balanced score of 0.83, combining precision and recall.
Support: There were 1703 expensive and 1725 non-expensive houses in the test set.

Logistic Regression for Classification

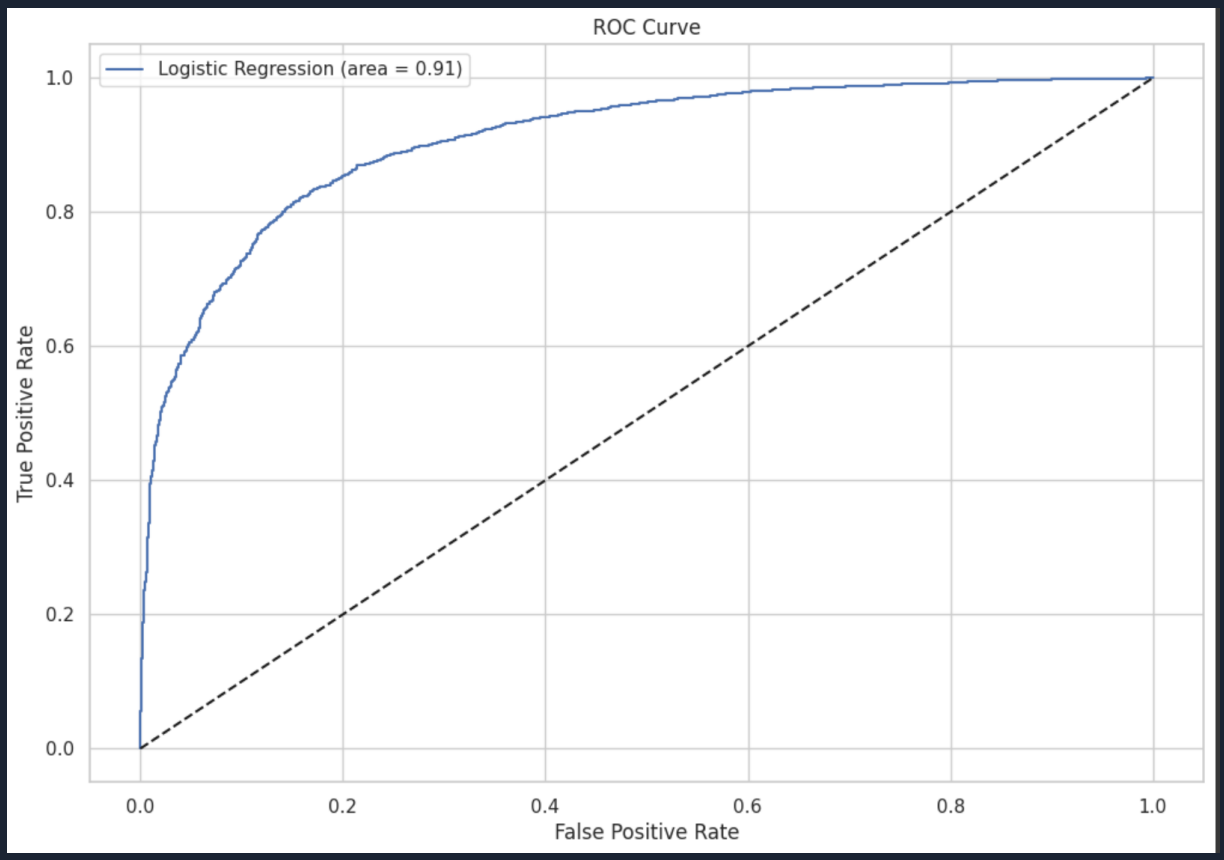
Confusion Matrix: The confusion matrix revealed the distribution of:
True Positives (TP): Correctly predicted expensive houses.
False Positives (FP): Houses incorrectly predicted as expensive (but actually not expensive).
True Negatives (TN): Correctly predicted not expensive houses.
False Negatives (FN): Houses incorrectly predicted as not expensive (but actually expensive).



Logistic Regression for Classification

ROC-AUC:

The ROC curve illustrates the model's performance in distinguishing between expensive and not expensive houses by plotting the true positive rate (sensitivity) against the false positive rate. The curve demonstrates how well the model performs at various classification thresholds. In this case, the area under the curve (AUC) is 0.91, indicating a strong ability of the logistic regression model to differentiate between the two classes. A higher AUC score closer to 1 represents better overall performance.



Non-Linear Regression

- Polynomial Regression Model:

To improve predictive performance, a polynomial regression model with degree 2 was implemented. This method allows the model to capture non-linear relationships between the features and house prices by generating polynomial features. The transformation takes the original features (e.g., median income, house age) and creates additional features such as the squares and interactions of those features. This helps capture more complex patterns in the data that linear regression cannot.

- Splitting the Data:

The dataset was split into training and testing sets using **train_test_split** with an 80-20 split. The training data (80%) was used to train the polynomial regression model, while the remaining 20% was set aside for testing. This ensures that the model's performance is evaluated on unseen data, providing a fair assessment of its predictive ability.

Non-Linear Regression

- Training the Model and Making Predictions:

Once the data was split, the polynomial regression model was trained on the training data using the transformed polynomial features. The model learned from the training set, attempting to capture the underlying patterns in the data. After training, the model was used to predict house prices on the test set, outputting predicted values for the house prices based on the learned polynomial relationships.

Non-Linear Regression

- **Model Evaluation:**

Mean Squared Error (MSE):
The Mean Squared Error (MSE) for the polynomial regression model is 0.0146. This value is quite low, indicating that the model's predicted house prices are very close to the actual prices. MSE measures the average squared differences between the actual and predicted values, and a lower value suggests better predictive performance.

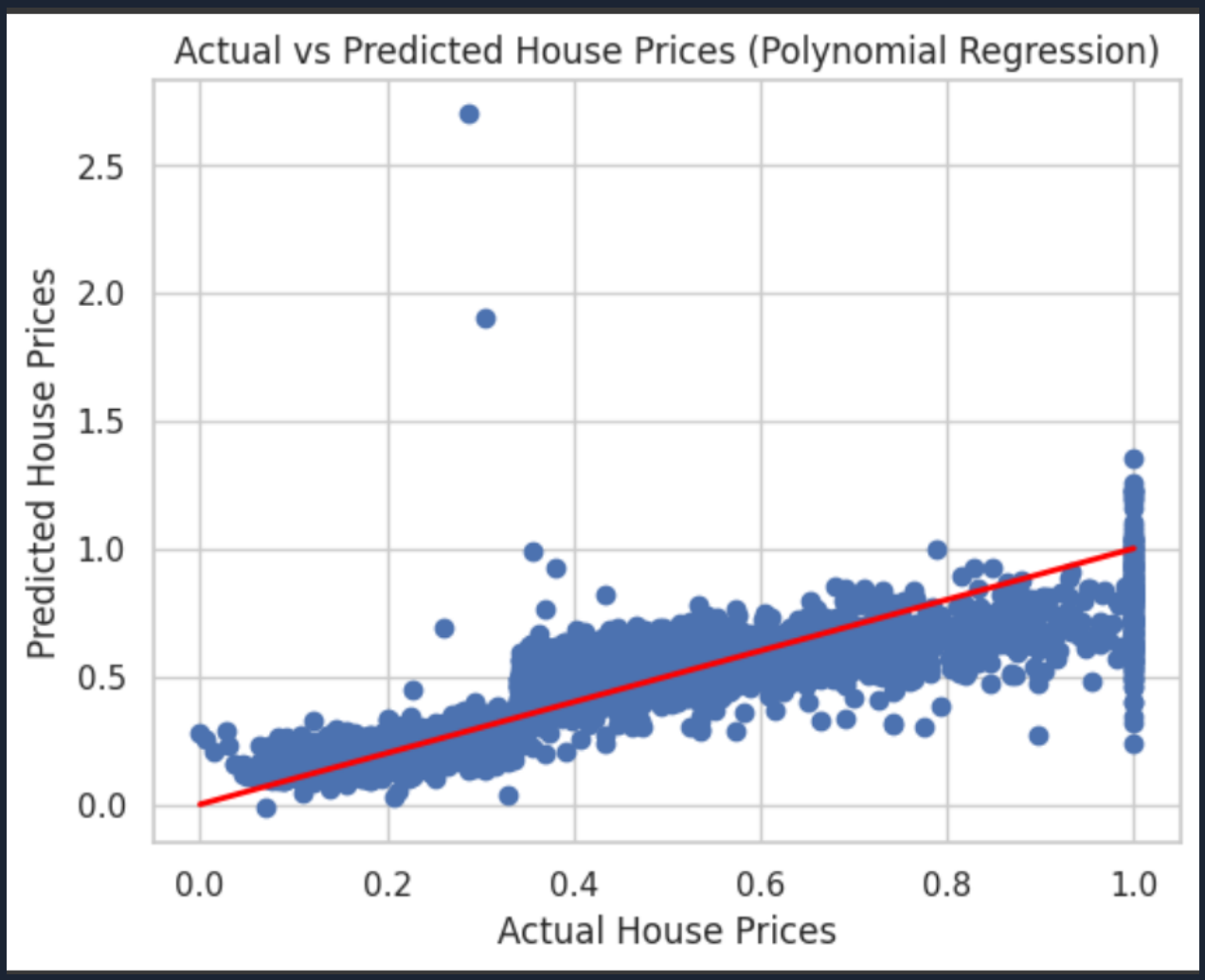
R-squared Score:

The R^2 score for the polynomial regression model is 0.737. This means that the model explains approximately 73.7% of the variance in house prices based on the features provided. A higher R^2 score indicates a better fit to the data, and while this score is reasonably high, there is still some room for improvement in capturing all the variance within the dataset.
These metrics provide an effective understanding of how well the polynomial regression model performs in predicting house prices based on the given data.

Non-Linear Regression

- **Actual vs. Predicted Plot:**

The plot compares the true house prices with the predictions from the polynomial regression model. The red diagonal line represents perfect predictions. While the model performs well for lower-priced houses, many points deviate from the line as house prices increase, indicating that the model struggles with higher prices. This suggests the need for further refinement or a more complex approach to improve accuracy, particularly for extreme values.



Analysis and Reflection

a. **Once you loaded the dataset, provides a description of the dataset, including the dependent and independent variables.**

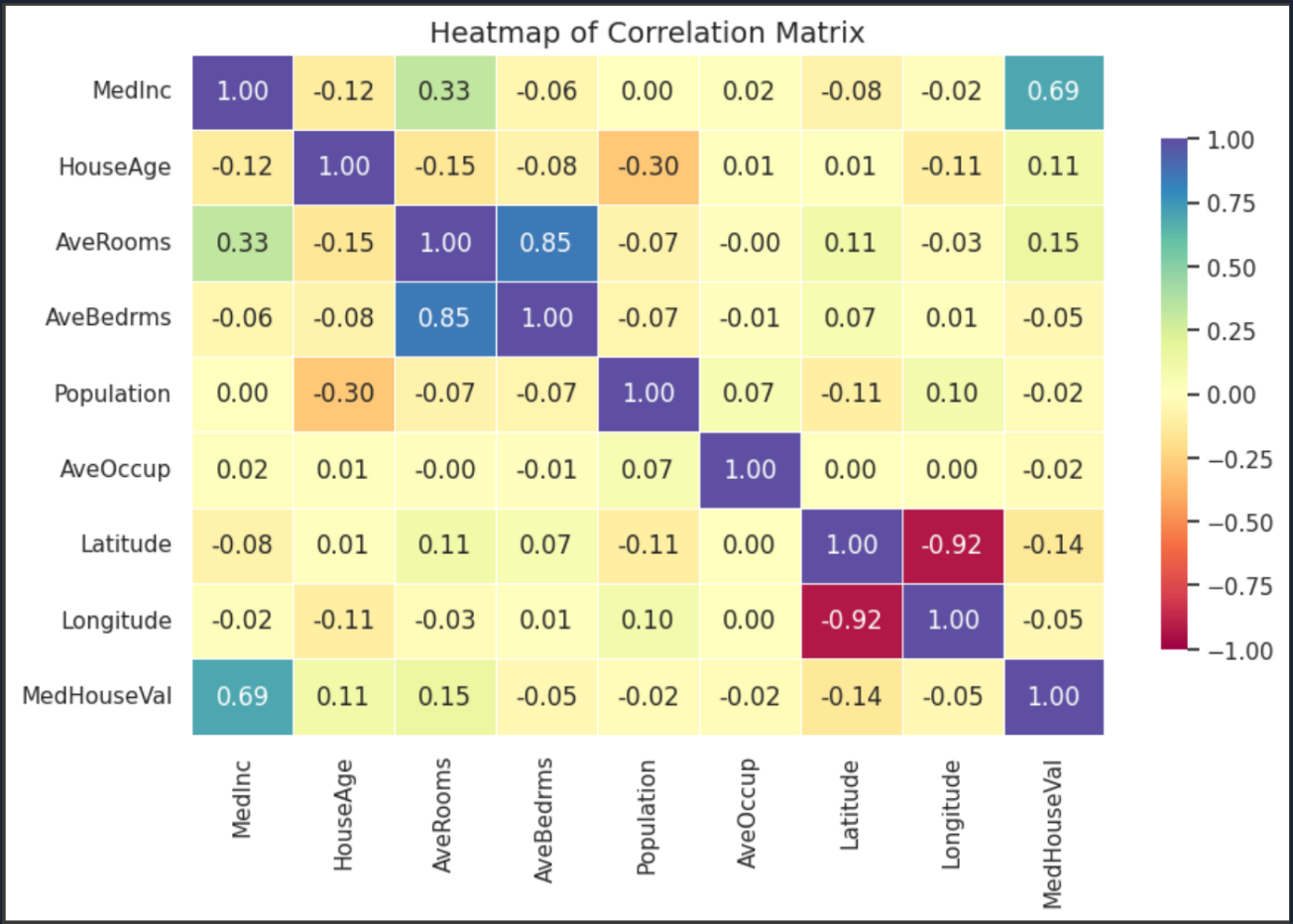
Summary of Dataset Variables

Variable Name	Description	Type	Dependent/Independent
MedHouseVal	Median house value for each area	Continuous	Dependent
MedInc	Median income of households	Continuous	Independent
HouseAge	Median age of houses in each area	Continuous	Independent
AveRooms	Average number of rooms per household	Continuous	Independent
AveBedrms	Average number of bedrooms per household	Continuous	Independent
Population	Population of the area	Continuous	Independent
AveOccup	Average number of occupants per household	Continuous	Independent
Latitude	Latitude of the area	Continuous	Independent
Longitude	Longitude of the area	Continuous	Independent

Analysis and Reflection

b. Which features seemed to be the most important in predicting house prices?

Answer: Median Income (MedInc) shows the strongest positive correlation with MedHouseVal **(0.69)**, indicating that it plays a crucial role in predicting house prices. As median income increases, house prices tend to rise. Average Number of Rooms (AveRooms) also shows a moderate positive correlation (**0.15)** with house prices. These two features appear to be the most influential in determining house prices according to the correlation analysis.



Analysis and Reflection

c. How well did the logistic regression model perform in classifying expensive houses? Discuss potential reasons for its performance.

Answer: The logistic regression model performed with an accuracy of **83.09%**, meaning approximately **83**% of predictions for expensive houses were correct. Based on the confusion matrix, the model correctly identified **1,703** expensive houses (True Positives) and **1,725** non-expensive houses (True Negatives). However, it also made **352** False Positive errors (predicting a house was expensive when it wasn't) and **348** False Negatives (failing to identify some expensive houses).

The model's relatively high accuracy suggests that it performs well overall, but it struggles with certain cases, possibly due to overlapping features between the expensive and non-expensive house categories. This is reflected in the balance of false positives and false negatives.

Analysis and Reflection

d.Compare the performance of the linear regression (which you can assume was poor) with the polynomial regression. Why might the polynomial regression perform better? .

- **Linear Regression Performance:** Linear regression assumes a straight-line relationship between the features and the target variable (house prices). If the relationship between the independent variables and the dependent variable is not linear, the model will fail to capture important patterns, leading to poor performance.

- **Polynomial Regression Performance:** Polynomial regression can handle non-linear relationships by introducing polynomial terms (e.g., squares, interactions) into the model. By using these additional features, the model can capture more complex patterns in the data. In your case, a second-degree polynomial regression (degree = 2) is used, which means it can capture quadratic relationships, which linear regression cannot.

- **Why Polynomial Might Perform Better:** Since house prices are likely influenced by non-linear relationships between features like number of rooms, and other factors, polynomial regression can fit the data better by accounting for those non-linearities. The improved R^2 score for the polynomial model compared to the linear model confirms that it is fitting the data more effectively.

Analysis and Reflection

e)What are some limitations of the models you used? How might you improve them?

Limitations of the Models:

Polynomial Regression:

Overfitting risk: It may fit the training data too well, reducing its ability to generalize.

Increased complexity: More polynomial terms make the model harder to interpret and more computationally expensive.

Logistic Regression:

Binary classification only: It's effective for binary tasks but struggles with unclear class boundaries.

Analysis and Reflection

e)What are some limitations of the models you used? How might you improve them?

Improvements:

Feature Engineering: Create new features to capture complex relationships.

Regularization: Use Lasso or Ridge to prevent overfitting in regression models.

Advanced Models: Try decision trees or random forests for better handling of non-linearity.

Cross-Validation: Ensure the model generalizes well by testing on different data subsets.

Conclusion (Findings)

Feature Importance: Median Income (MedInc) and Average Number of Rooms (AveRooms) were the most significant predictors of house prices.

Logistic Regression Performance: Achieved 83% accuracy with an ROC-AUC score of 0.91 but struggled with higher-priced houses.

Polynomial Regression: Showed better performance than linear regression with lower MSE and higher R-squared scores, capturing more complex relationships in the data.

Limitations: Both models faced challenges in predicting extreme house prices. Future improvements could include more advanced models and feature engineering.

References