PART A: Property Price Prediction Model Report

Introduction

This project focuses on predicting the median house value in California districts using linear regression models (Simple and Multiple Linear Regression). Initially, the model was built using basic preprocessing and feature engineering techniques.

Data Preprocessing

Steps taken in the preprocessing phase:

1. Handled missing values in the 'total_bedrooms' column by imputing the median value.
2. Created composite features to address multicollinearity:
   - 'rooms_per_household' = total_rooms / households
   - 'bedrooms_per_room' = total_bedrooms / total_rooms
   - 'population_per_household' = population / households
3. Dropped highly correlated features: 'total_rooms', 'total_bedrooms', 'population', 'households'.
4. Encoded the categorical column 'ocean_proximity' using one-hot encoding.
5. Normalized numerical features using MinMaxScaler to scale all numerical data to a range between 0 and 1.
6. Added polynomial features (e.g., median_income squared and cubed) to capture non-linear relationships.
7. Introduced interaction terms to account for combined effects between variables (e.g., median_income * housing_median_age).
8. Applied log transformations to stabilize variance in skewed features (e.g., median_income, median_house_value).

Modeling Steps: Simple Linear Regression

1. Selected 'median_income' as the sole predictor variable.
2. Split the dataset into training (80%) and testing (20%) sets.
3. Trained a Simple Linear Regression model using the training set.
4. Evaluated the model using Root Mean Squared Error (RMSE) and R^2 score on the test set.

Modeling Steps: Multiple Linear Regression

1. Used all features (except the target variable) after preprocessing.
2. Split the dataset into training (80%) and testing (20%) sets.
3. Trained a Multiple Linear Regression model using the training set.
4. Evaluated the model using Root Mean Squared Error (RMSE) and R^2 score on the test set.

Initial Model Performance

Initial performance metrics were as follows:

- Simple Linear Regression:
  - RMSE = 0.1736
  - R^2 = 0.4589
- Multiple Linear Regression:
  - RMSE = 0.1578
  - R^2 = 0.5530

These metrics highlighted the limitations of the initial approach, particularly for Multiple Linear Regression, which struggled to explain the variance in the target variable.

Decision to Refine

Given the relatively low R^2 scores and the potential for non-linear relationships, it was decided to enhance feature engineering and introduce polynomial and interaction terms. This decision aimed to better capture the complexities in the data while staying within the linear regression framework.

Improved Model Performance

After applying refinements, the models showed the following improvements:

- Simple Linear Regression (unchanged):
  - RMSE = 0.1736
  - R^2 = 0.4589
- Multiple Linear Regression:
  - RMSE = 0.01697
  - R^2 = 0.9948

The dramatic improvement in Multiple Linear Regression performance demonstrates the value of the added features and transformations.

Conclusion

The refinements applied, such as polynomial features, interaction terms, and log transformations, effectively captured non-linear relationships in the data, leading to a much better fit for the Multiple Linear Regression model. Future work can focus on validating the model on unseen data, further feature selection, and exploring more advanced algorithms if needed.