California Housing Price Prediction

Presented By: Muhammad Osama

Introduction

- **Objective:**

Predict median house values using geographic, economic, and demographic features.

- **Dataset:**
  - Source: 1990 U.S. Census.
  - Size: 20,640 entries, 10 features.
- **Key Features**: median_income, ocean_proximity, total_rooms, geographic coordinates.
- **Importance:**

Real-world application for real estate pricing and policy-making.

Data Preprocessing

- **Importing Libraries:** Importing all necessary libraries
- **Missing Values:** Dropped rows with missing total_bedrooms.
- **Duplicates:** Removed duplicates.
- **Encoding:** One-hot encoded ocean_proximity into 5 binary columns.
- **Train Test Split:** 80% training, 20% testing.

```
# Removing any rows with missing data
df.dropna(inplace=True)

# Remove all duplicate rows
df.drop_duplicates(inplace=True)

# Hot encoding
data_encoded = pd.get_dummies(df, columns=['ocean_proximity'], drop_first=True)
print(data_encoded.head())

# Spilitting the data
X = data_encoded.drop('median_house_value', axis=1)
y = data_encoded['median_house_value']

# 80% Training, 20% Testing Data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
print("Training set shape:", X_train.shape)
print("Test set shape:", X_test.shape)
print("Dataset Shape after Data Cleaning:", df.shape)
```

Correlation Matrix

- Correlation measures the statistical relationship between two variables, indicating the strength their linear relationship, ranging from -1 to +1 with 0 indicating no linear relationship
- **Strong Positive Correlation:** median_income: +0.69
  - Higher income strongly predicts higher house prices.
- **Negative Correlation:** ocean_proximity_INLAND: -0.48
  - Inland locations correlate with lower prices.

Feature	Correlation with Target:
median_house_value	1
median_income	0.688075
ocean_proximity_<1H OCEAN	0.256617
ocean_proximity_NEAR BAY	0.160284
ocean_proximity_NEAR OCEAN	0.141862
total_rooms	0.134153
housing_median_age	0.105623
households	0.065843
total_bedrooms	0.049686
ocean_proximity_ISLAND	0.023416
population	-0.02465
longitude	-0.045967
latitude	-0.14416
ocean_proximity_INLAND	-0.484859

MAE and MSE

- **Mean Absolute Error :** Average of absolute differences between predicted and actual values
  - **Easy to Interpret**: Directly represents average error in the target's units
  - **Robust to Outliers**: Less sensitive to extreme errors
  - **Ignores Error Severity**: Treats large and small errors equally
- **Mean Squared Error:** Average of squared differences between predicted and actual values.
  - **Penalizes Large Errors:** Highlights severe mistakes
  - **Hard to Interpret:** Squared units (e.g., dollars^2) lack intuitive meaning.
  - **Outlier Sensitivity**: Overemphasizes rare, extreme errors (e.g., skewed by a few poor predictions).

Algorithms & Hyperparameters

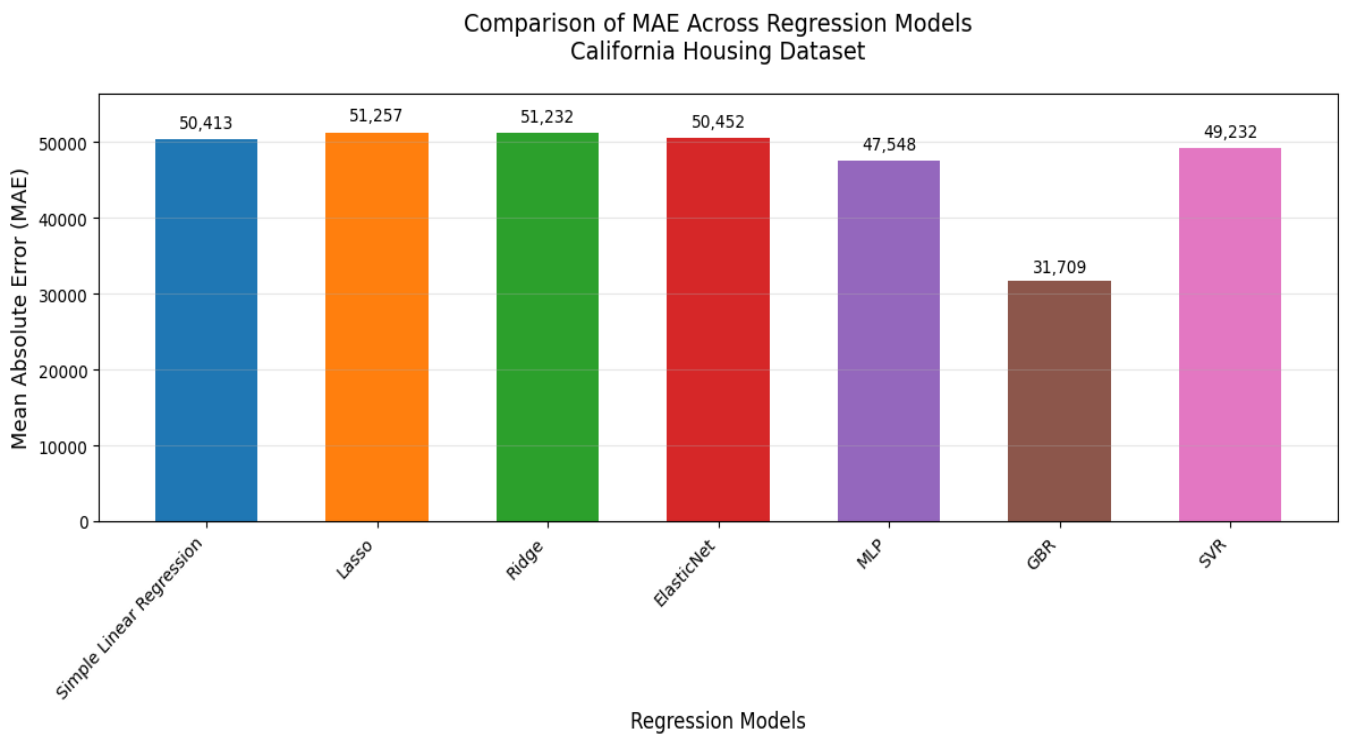
- **Linear Models:**
  - **Simple Linear Regression** is a statistical method used to model the relationship between two variables: **$Y=\beta 0+\beta 1X+\varepsilon$**
    - One independent variable (X) – the predictor or input.
    - One dependent variable (Y) – the outcome or target
  - **Lasso Regression** adds an **L1 regularization** penalty which helps reduce model complexity and performs feature selection by shrinking some coefficients to zero
    - **Best alpha=100 (Lasso)**
  - **Ridge Regression** adds an **L2 regularization** penalty which reduces coefficient size but does not set any of them exactly to zero.
    - **Best alpha= 10 (Ridge)**
- **ElasticNet Regression** combines both **L1 and L2 regularization** penalties, balancing feature selection (L1) and coefficient shrinkage (L2)
  - **alpha=0.1, l1_ratio=0.5**

Algorithms & Hyperparameters

- **Gradient Boosting (GBR):**
  - GridSearchCV: n_estimators=300, learning_rate=0.1, max_depth=5.
  - Grid Search to 27 combination
  - K Fold cv = 5
- **MLPRegressor:**
  - Architecture: 2 hidden layers (10, 5 neurons). Max iter = 5000
- **SVR:**
  - kernel=linear, C=100, epsilon=0.1

Model Performance Comparison

- **Gradient Boosting (GBR):** GBR outperforms others due to non-linear pattern capture and tuning.

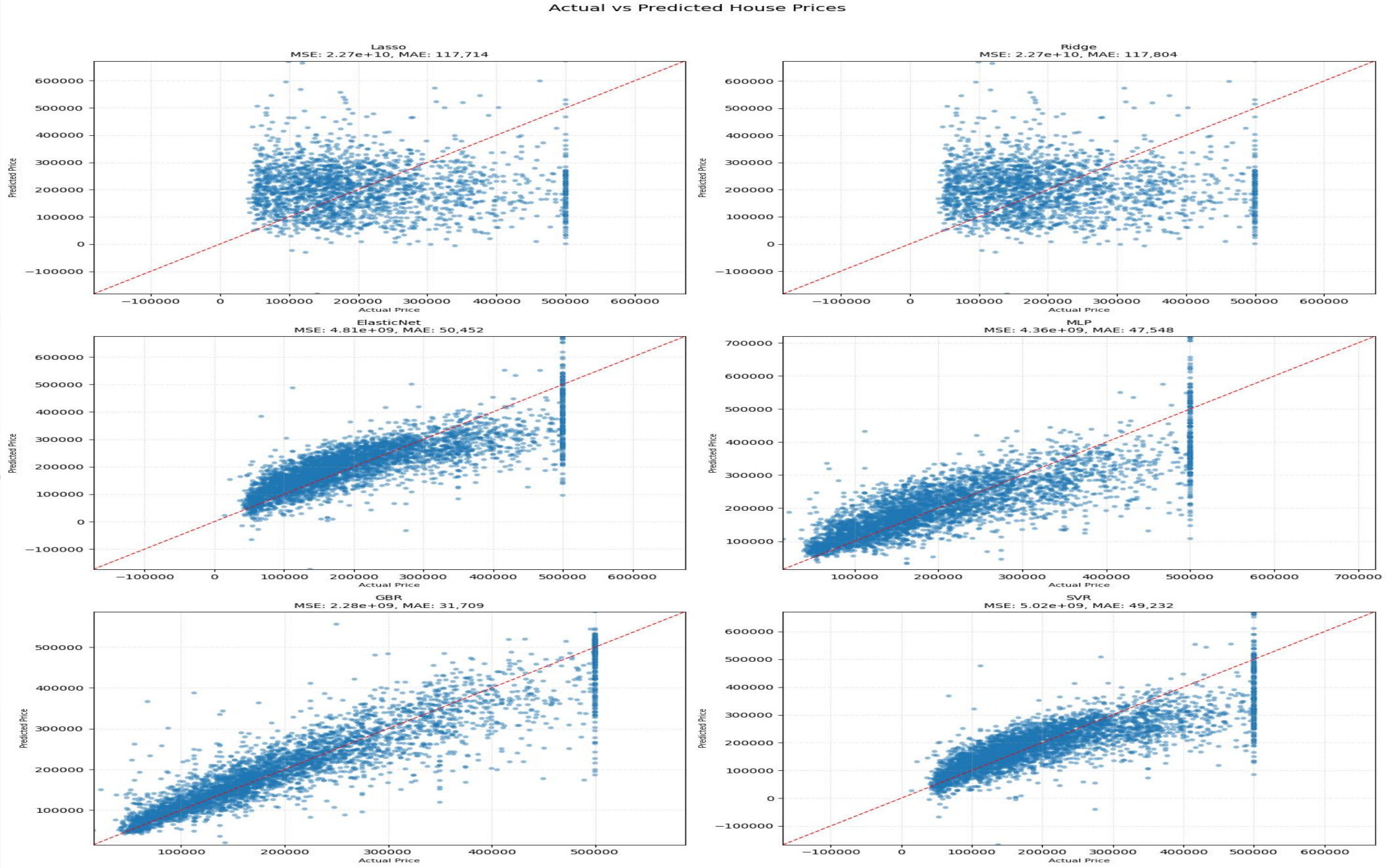


Model	MAE	MSE
Gradient Boosting (GBR)	31,709	2.28×10^9
MLP	48,249	4.47×10^9
Linear Regression	50,413	4.80×10^9
ElasticNet	50,452	4.81×10^9
Ridge	51,232	5.01×10^9
Lasso	51,257	5.02×10^9
Support Vector Regression	49,232	5.02×10^9

Actual vs. Predicted Values

- **Visual**: Scatter plots for GBR, Linear Regression, and SVR.
  - **GBR**: Points cluster tightly around the red ideal line.
  - **Linear Models**: Wider spread, higher deviation.
- **Takeaway**:
  - GBR's predictions align closely with actual prices.

Actual vs. Predicted Values



Thank you