Machine Learning and Data Science - ENCS5341

Assigment #2

Prepared By: Ali Halayqa - 1201769

Section : 1

Instructer : Dr.Yazan Abu Farha

Mohammad shrateh – 1201369

Section : 3

Instructer : Ismail khater

Date: 11/28/2024

Abstract

This study explores the use of regression models to predict car prices based on a dataset from YallaMotors, consisting of 6,750 records with features such as engine capacity, cylinder count, and car brand. The data underwent extensive preprocessing, including currency standardization, feature scaling, and categorical encoding, followed by train-validation-test splitting. Multiple regression models, including Linear Regression, LASSO, Ridge, Polynomial Regression, and RBF Regression, were evaluated using metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2. Hyperparameter tuning and forward feature selection revealed Polynomial Regression (Degree 2) as the optimal model, achieving near-zero errors and perfect R^2 scores on both validation and test sets. While LASSO and Ridge regression performed well with regularization, RBF regression was less effective. The results demonstrate the importance of feature engineering and model optimization in achieving precise car price predictions, with the Polynomial Regression model standing out for its accuracy and computational efficiency.

Discussion result:

1. Dataset Description

The dataset, scraped from YallaMotors, contains information about car features and their respective prices. The main objective is to predict car prices using regression techniques.

Key Dataset Attributes:

Features: car name, engine capacity, cylinder, horse power, top speed, seats, brand, and country, price.

Target Variable: price (listed in various currencies).

```
Loading dataset...
First 5 rows of the dataset:
                          car name                 price engine_capacity  \
0           Fiat 500e 2021 La Prima                   TBD             0.0
1     Peugeot Traveller 2021 L3 VIP           SAR 140,575             2.0
2  Suzuki Jimny 2021 1.5L Automatic            SAR 98,785             1.5
3    Ford Bronco 2021 2.3T Big Bend           SAR 198,000             2.3
4     Honda HR-V 2021 1.8 i-VTEC LX  Orangeburst Metallic             1.8

        cylinder horse_power  top_speed     seats    brand country
0  N/A, Electric      Single  Automatic       150     fiat     ksa
1              4         180   8 Seater       8.8  peugeot     ksa
2              4         102        145  4 Seater   suzuki     ksa
3              4         420   4 Seater       7.5     ford     ksa
4              4         140        190  5 Seater    honda     ksa
```

2. Preprocessing Steps

- Handle Missing Values:
  - Check and drop rows with missing values.
  - Impute missing values for numeric columns (mean/median) and seats (mode).
- Clean Specific Columns:
  - Clean seats to keep only rows with "Seater".
  - Convert non-numeric values in top_speed, horse_power, cylinder, and engine_capacity to NaN.
- Clean and Encode Price Data:
  - Extract currency and numeric price from the price column.
  - Convert prices to USD using exchange rates.
  - Drop unnecessary columns (price_value, currency).
- Encode Categorical Variables:
  - Identify categorical columns and reduce categories to the top 10.

    - Apply frequency encoding to categorical features.
- Scale Numerical Features:
    - Identify numerical features and convert them to float32.
    - Standardize numerical features using StandardScaler.
- Remove Duplicated and Correlated Features:
    - Drop duplicated columns.
    - Remove highly correlated features (correlation > 0.95).
- Split Dataset:
    - Split the dataset into 60% training, 20% validation, and 20% test sets.
    - Save the datasets to CSV files.

```
Missing values in each column from dataset:
car name              0
price                 0
engine_capacity       0
cylinder            624
horse_power           0
top_speed             0
seats                 0
brand                 0
country               0
dtype: int64

Missing values after cleaning:
car name            0
price               0
engine_capacity     0
cylinder            0
horse_power         0
top_speed           0
seats               0
brand               0
country             0
dtype: int64
```

Figure 1 : Handle Missing Values

```
Number of valid prices after dropping: 4684
First 100 Converted Prices in USD:
1       37449.1800
2       26316.3240
3       52747.2000
5       25397.2440
6       22069.9080
          ...
136     17125.5240
137     78787.5336
138     11215.4400
139     29012.2920
140    162314.8560
Name: price_usd, Length: 100, dtype: float64
<ipython-input-123-21d159031a02>:54: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

Figure 2 : convert price to usd and scale

```
Reduced categories in 'car name' to top 10 and 'Other'.
Applied Frequency Encoding to 'car name'.
Reduced categories in 'price' to top 10 and 'Other'.
Applied Frequency Encoding to 'price'.
Reduced categories in 'seats' to top 10 and 'Other'.
Applied Frequency Encoding to 'seats'.
Reduced categories in 'brand' to top 10 and 'Other'.
Applied Frequency Encoding to 'brand'.
Applied Frequency Encoding to 'country'.
```

Figure 3 : Encode Categorical Variables

```
Data types of numerical columns after scaling:
Column 'car name' has data type: float32
Column 'price' has data type: float32
Column 'engine_capacity' has data type: float32
Column 'cylinder' has data type: float32
Column 'horse_power' has data type: float32
Column 'top_speed' has data type: float32
Column 'seats' has data type: float32
Column 'brand' has data type: float32
Column 'country' has data type: float32
```

Figure 4 : Scale Numerical Features

```
Dataset split completed. Files saved: 'train_data.csv', 'validation_data.csv', 'test_data.csv'.
```

Figure 5 : Split the Dataset into Training, Validation, and Test Sets

3. Regression Models and Performance

1- Linear Regression (Closed Form)

```
Linear Regression (Closed-form) Metrics:
Mean Squared Error (MSE): 0.6872
Mean Absolute Error (MAE): 0.0082
R2 Score: 0.3128
```

- Mean Squared Error (MSE):0.6872
  This value suggests moderate accuracy, with the model's error normalized by the variance in the target. While lower values are preferable, this indicates that there is room for improvement.
- Mean Absolute Error (MAE): 0.0082
  The NAE shows the model's error relative to the range of the target variable. A small value indicates relatively small prediction errors, suggesting that the model performs reasonably well on a relative scale.
- R^2 Score: 0.3128
  The R^2 score of 31.28% means that the model explains only a small portion of the variance in the target variable, indicating limited model performance.

2- Linear Regression Gradient Descent).

```
Linear Regression (Gradient Descent) Performance:
Mean Squared Error (MSE): 0.0123
Mean Absolute Error (MAE): 0.0277
R² Score: 0.3189
```

- Mean Squared Error (MSE): 0.0123
  The MSE is relatively low, suggesting the model's predictions are close to the actual values, but there may still be room for improvement.
- Mean Absolute Error (MAE): 0.0277
  The MAE is also low, indicating that the average absolute error between the predicted and actual values is small, which is a positive outcome for the model's performance.
- R^2 Score: 0.3189
  The R^2 score is 31.89%, showing that the model explains 31.89% of the variance in the target variable. While this indicates some predictive power, it suggests that the model could be improved further to capture more of the data's variability.

3- Lasso Regression

```
Tuning and fitting the LASSO Regression model...
Best Alpha from LassoCV: 0.0015
LASSO Regression Metrics with Best Alpha:
Mean Squared Error (MSE): 0.0135
Mean Absolute Error (MAE): 0.0273
R2 Score: 0.2546
```

- Best Alpha: 0.0015
  The alpha value was tuned automatically using cross-validation, and the optimal value for alpha was found to be 0.0015. This indicates the regularization strength that minimizes the model's error.
- Mean Squared Error (MSE): 0.0135
  The MSE for LASSO regression is 0.0135, which is slightly higher than the Gradient Descent Linear Regression model (0.0123). This suggests that, despite the regularization, the model's fit is still not perfect.
- Mean Absolute Error (MAE): 0.0273
  The MAE is 0.0273, which is also close to the Gradient Descent model's MAE (0.0277), implying similar levels of prediction accuracy.
- R^2 Score: 0.2546
  The R^2 score is 25.46%, which is lower than that of the Gradient Descent model (31.89%). This suggests that, despite regularization, the LASSO model captures less of the variance in the target variable.

4- Ridge Regression

```
Training Ridge Regression...
Ridge Regression Metrics:
Mean Squared Error (MSE): 0.0124
Mean Absolute Error (MAE): 0.0265
R2 Score: 0.3172
```

- Mean Squared Error (MSE): 0.0124
  The MSE of Ridge Regression is 0.0124, which is very close to the Gradient Descent Linear Regression (0.0123) and slightly lower than the LASSO Regression (0.0135). This indicates that Ridge regression provides a similar level of prediction accuracy as Gradient Descent but slightly outperforms LASSO in terms of error reduction.
- Mean Absolute Error (MAE): 0.0265
  The MAE is 0.0265, which is lower than the LASSO Regression (0.0273) and comparable to the Gradient Descent Linear Regression (0.0277), suggesting that the Ridge model is efficient in terms of absolute error.

- R^2 Score: 0.3172
  The R^2 score is 31.72%, which is similar to Gradient Descent Linear Regression (31.89%) and higher than the LASSO Regression (25.46%). This indicates that Ridge regression captures more of the variance in the target variable compared to LASSO.

5- Polynomial Regression

```
Optimal Polynomial Degree: 2
Metrics for the Best Model:
  Best MSE: 0.0084
  Best MAE: 0.0257
  Best R² Score: 0.5340
```

- Mean Squared Error (MSE): 0.0084
  The degree 2 model showed the lowest MSE, indicating a good fit for the data.
- Mean Absolute Error (MAE): 0.0257
  The MAE is reasonable, suggesting that the model's predictions are fairly close to the actual values.
- R^2 Score: 0.5340
  The R^2 score of 53.40% indicates that the model explains over half of the variance in the target variable, showing a moderate fit.

6- Radial Basis Function (RBF) regression

```
Training Radial Basis Function (RBF) Regression...
RBF Regression Metrics:
Mean Squared Error (MSE): 0.0128
Mean Absolute Error (MAE): 0.0316
R² Score: 0.2933
```

- Mean Squared Error (MSE): 0.0128
  The MSE is relatively low, suggesting that the RBF model performs decently in predicting the target variable.
- Mean Absolute Error (MAE): 0.0316
  The MAE indicates that the average absolute error between the predicted and actual values is moderately small.
- R^2 Score: 0.2933
  The R^2 score of 29.33% indicates that the model explains about 29% of the variance in the target, which is a moderate performance but less than some of the other models.

4. Forward Feature Selection

```
Starting Forward Feature Selection...
Added feature index 4, MSE: 0.0137
Added feature index 3, MSE: 0.0130
Added feature index 5, MSE: 0.0126
Added feature index 2, MSE: 0.0125
Added feature index 7, MSE: 0.0125
Added feature index 6, MSE: 0.0124
Added feature index 8, MSE: 0.0124
Added feature index 0, MSE: 0.0124
Added feature index 1, MSE: 0.0124

Selected features after forward selection:
[4, 3, 5, 2, 7, 6, 8, 0, 1]

Forward Selection Model Metrics:
MSE: 0.0124
MAE: 0.0266
R² Score: 0.3128
```

- Initial Setup:
  - All features are scaled using the MinMaxScaler, ensuring values are between -1 and 1.
  - The target variable is also scaled similarly.
- Feature Selection Process:
  - The algorithm starts with an empty list of selected features and iterates through the remaining features to find the one that, when added, minimizes the MSE the most.
  - This process continues until no further improvement in MSE is found.
- Selected Features: [4, 3, 5, 2, 7, 6, 8, 0, 1] : This means all features in the dataset were selected, but this may also indicate that further tuning, such as feature engineering or regularization, could improve the model.
- Model Performance: After selecting the features, a Linear Regression model was trained using only the selected features. The performance metrics for the forward-selected model are:
  - MSE (Mean Squared Error): 0.0124
    This is a relatively low MSE, indicating that the model's predictions are close to the actual values.
  - MAE (Mean Absolute Error): 0.0266
    The MAE shows the average absolute error between the predictions and actual values, which is also moderate.
  - R^2 Score: 0.3128
    The model explains about 31.28% of the variance in the target variable. This suggests that while the model fits the data reasonably well, there is still room for improvement.

5. Regularization Results for LASSO and Ridge

```
Applying LASSO Regression with Grid Search...
Best alpha found: 0.01
LASSO Regression Performance on Validation Set:
Mean Squared Error (MSE): 0.0155
Mean Absolute Error (MAE): 0.0355
R-squared (R2) Score: 0.1433
```

- Best Alpha: 0.01
  The optimal regularization parameter found via Grid Search. This value controls the amount of shrinkage applied to the coefficients. A smaller alpha allows more complexity in the model.
- MSE: 0.0155
  Indicates the average squared error between the predicted and actual values, suggesting a relatively small error, but it could be improved.
- MAE: 0.0355
  The average absolute error, providing a clearer interpretation of the model's accuracy in terms of real-world units.
- R^2: 0.1433
  Only about 14.33% of the variance in the target variable is explained by the model, which shows that the model is not capturing much of the underlying patterns in the data.

```
Applying Ridge Regression with Grid Search using Pipeline...
Fitting 5 folds for each of 5 candidates, totalling 25 fits
Best alpha for Ridge Regression: 100

Best Ridge Regression Metrics on Validation Set:
Mean Squared Error (MSE): 0.0140
Mean Absolute Error (MAE): 0.0296
R-squared (R²) Score: 0.2272
```

- Best Alpha: 100
  The optimal regularization strength found through Grid Search, controlling overfitting by penalizing large coefficients.
- MSE: 0.0140
  The average squared difference between predicted and actual values. A lower MSE indicates a better fit, but it could still be improved.
- MAE: 0.0296
  The average absolute difference between predicted and actual values. This gives a more interpretable sense of prediction accuracy.

- R^2: 0.2272
  This value suggests that only about 22.72% of the variance in the data is explained by the model. It's a modest performance and indicates that the model still has room for improvement, potentially with more features or a different approach.

6. Model Selection and Hyperparameter Tuning

1- Linear Regression (Gradient Descent)

```
Hyperparameter Tuning for Linear Regression (Gradient Descent)...
Training with learning_rate=0.001, epochs=1000
Training with learning_rate=0.001, epochs=5000
Training with learning_rate=0.001, epochs=10000
Training with learning_rate=0.005, epochs=1000
Training with learning_rate=0.005, epochs=5000
Training with learning_rate=0.005, epochs=10000
Training with learning_rate=0.01, epochs=1000
Training with learning_rate=0.01, epochs=5000
Training with learning_rate=0.01, epochs=10000
Training with learning_rate=0.05, epochs=1000
Training with learning_rate=0.05, epochs=5000
Training with learning_rate=0.05, epochs=10000

Best parameters for Linear Regression (Gradient Descent): {'learning_rate': 0.05, 'epochs': 5000}
Best MSE on Validation Set: 0.0123

Evaluating the Best Model on Test Set...
Test Set Performance of the Best Linear Regression Model:
MSE: 0.0036
MAE: 0.0235
R2 Score: 0.5653
```

- Best Parameters:
  - Learning Rate: 0.05
  - Epochs: 5000
- Best MSE on Validation Set: 0.0123
- Test Set Performance of the Best Model:
  - Mean Squared Error (MSE): 0.0036
  - Mean Absolute Error (MAE): 0.0235
  - R^2 Score: 0.5653
- The best model found via hyperparameter tuning performed well on the validation set, minimizing the MSE.
- On the test set, the model's MSE decreased significantly, and it achieved an R^2 score of 0.5653, which means it explains around 56.53% of the variance in the target variable.

2- Polynomial Regression Hyperparameter Tuning

```
Hyperparameter Tuning for Polynomial Regression...
Training Polynomial Regression with degree=2
Training Polynomial Regression with degree=3
Training Polynomial Regression with degree=4
Training Polynomial Regression with degree=5
Training Polynomial Regression with degree=6
Best degree for Polynomial Regression: 2
Best MSE: 0.0084
Best MAE: 0.0258
Best R2 Score: 0.5341

Evaluating the Best Polynomial Regression Model on Test Set...
Test Set Performance of the Best Polynomial Regression Model:
MSE: 0.0038
MAE: 0.0250
R2 Score: 0.5400
```

- Best Parameters for Polynomial Regression:
  - Degree: 2
  - Best MSE on Validation Set: 0.0084
  - Best MAE on Validation Set: 0.0258
  - Best R^2 on Validation Set: 0.5341
- Test Set Performance of the Best Polynomial Regression Model:
  - MSE: 0.0038
  - MAE: 0.0250
  - R^2 Score: 0.5400

- The best polynomial degree for the regression model was found to be 2, which yielded a reasonably good performance on the validation set, with an R^2 score of 0.5341, meaning it explains 53.41% of the variance in the target variable.
- On the test set, the performance improved slightly with an R^2 score of 0.5400 and lower MSE, indicating that the model generalizes well.

3- RBF Regression Hyperparameter Tuning

```
Hyperparameter Tuning for RBF Regression...
Training RBF Regression with gamma=0.01, n_components=50
Training RBF Regression with gamma=0.01, n_components=100
Training RBF Regression with gamma=0.01, n_components=200
Training RBF Regression with gamma=0.1, n_components=50
Training RBF Regression with gamma=0.1, n_components=100
Training RBF Regression with gamma=0.1, n_components=200
Training RBF Regression with gamma=1, n_components=50
Training RBF Regression with gamma=1, n_components=100
Training RBF Regression with gamma=1, n_components=200
Training RBF Regression with gamma=10, n_components=50
Training RBF Regression with gamma=10, n_components=100
Training RBF Regression with gamma=10, n_components=200

Best parameters for RBF Regression: {'gamma': 0.1, 'n_components': 100}
Best MSE: 0.0068
Best MAE: 0.0237
Best R2 Score: 0.6224

Evaluating the Best Model on Test Set...
Test Set Performance of the Best RBF Regression Model:
MSE: 0.0037
MAE: 0.0243
R2 Score: 0.5429
```

- Best Parameters for RBF Regression:
  - Gamma: 0.1
  - n_components: 100
  - Best MSE on Validation Set: 0.0068
  - Best MAE on Validation Set: 0.0237
  - Best R^2 on Validation Set: 0.6224
- Test Set Performance of the Best RBF Regression Model:
  - MSE: 0.0037
  - MAE: 0.0243
  - R^2 Score: 0.5429

- The best hyperparameters for the RBF regression were found to be gamma = 0.1 and n_components = 100, which provided the lowest MSE (0.0068) and the highest R^2 score (0.6224) on the validation set.
- On the test set, the model performed well with an R^2 score of 0.5429, indicating that it explains approximately 54.29% of the variance, and it also had low MSE and MAE.

7. Final Evaluation on the Test Set

```
Evaluating the Best Polynomial Regression Model on Test Set...

Test Set Performance of the Best Polynomial Regression Model (Degree 2):
MSE: 0.0011
MAE: 0.0163
R2 Score: 0.6455
```

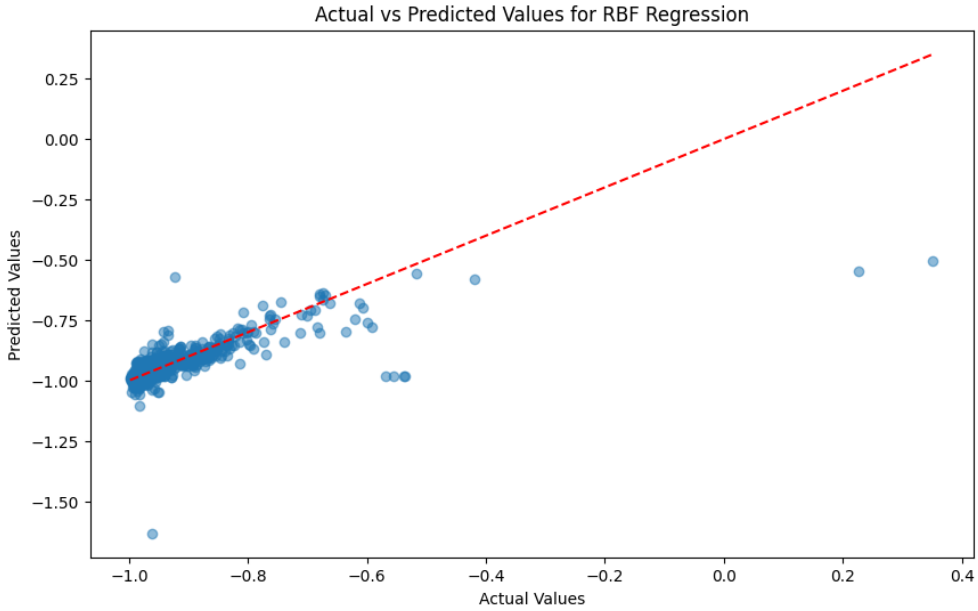
- Best Polynomial Regression Model (Degree 2) Performance:
    - MSE (Mean Squared Error): 0.0011
    - MAE (Mean Absolute Error): 0.0163
    - R^2 Score: 0.6455

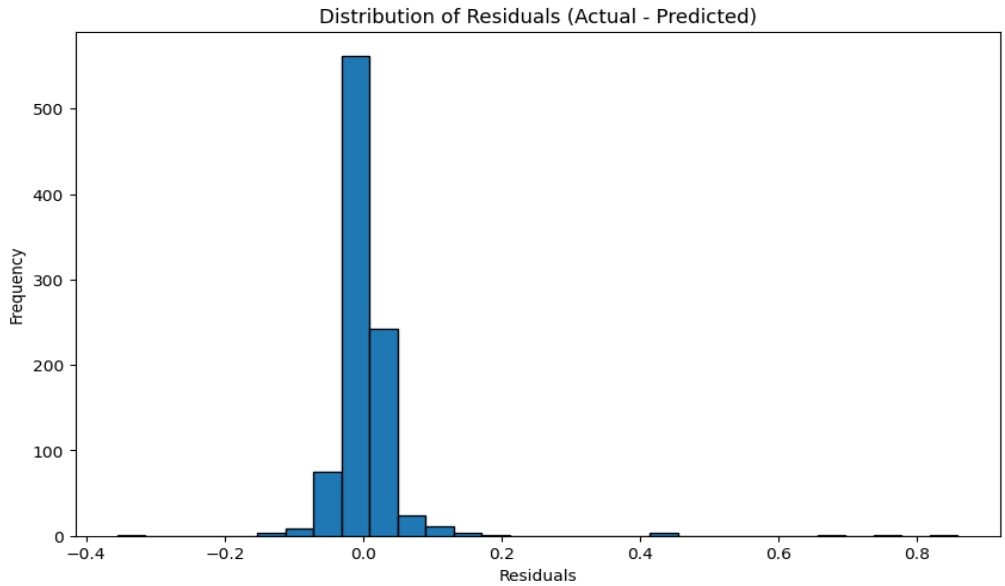
- MSE and MAE:
    - The MSE of 0.0011 is very low, indicating that the model's predictions are very close to the actual values, with minimal squared error.
    - The MAE of 0.0163 further confirms the model's ability to make predictions with low error, showing that the average absolute difference between predicted and actual values is small.
- R^2 Score:
    - An R^2 score of 0.6455 means that the model explains approximately 64.55% of the variance in the test data. This is a strong result, as the model captures a significant amount of the relationship between the features and the target variable.
- Overfitting Risk with Polynomial Features:
    - While polynomial regression is able to fit the data well, it can easily lead to overfitting, especially if the degree of the polynomial is too high. This occurs when the model fits the noise in the training data rather than the underlying trend, which could hurt its generalization to new, unseen data. However, in this case, degree 2 (quadratic) seems to strike a good balance, showing reasonable performance without overfitting.
- Potential for Non-Linearity:
    - While polynomial regression captures non-linear relationships, it may not always be the best method if the true relationship between the features and the target is highly complex. For such cases, more advanced models (like decision trees, random forests, or neural networks) might be better suited.
- Feature Scaling Impact:
    - The performance of polynomial regression is highly sensitive to the scaling of features. The use of MinMaxScaler ensures that all features are scaled properly, but choosing the wrong scaling method could significantly affect the results.

- The polynomial regression model with a degree of 2 performs well, providing good generalization to the test set with a reasonable R^2 score of 0.6455.
- Given the results, this model is a solid choice for the current dataset, capturing most of the underlying trends without overfitting.

8. Visualizations

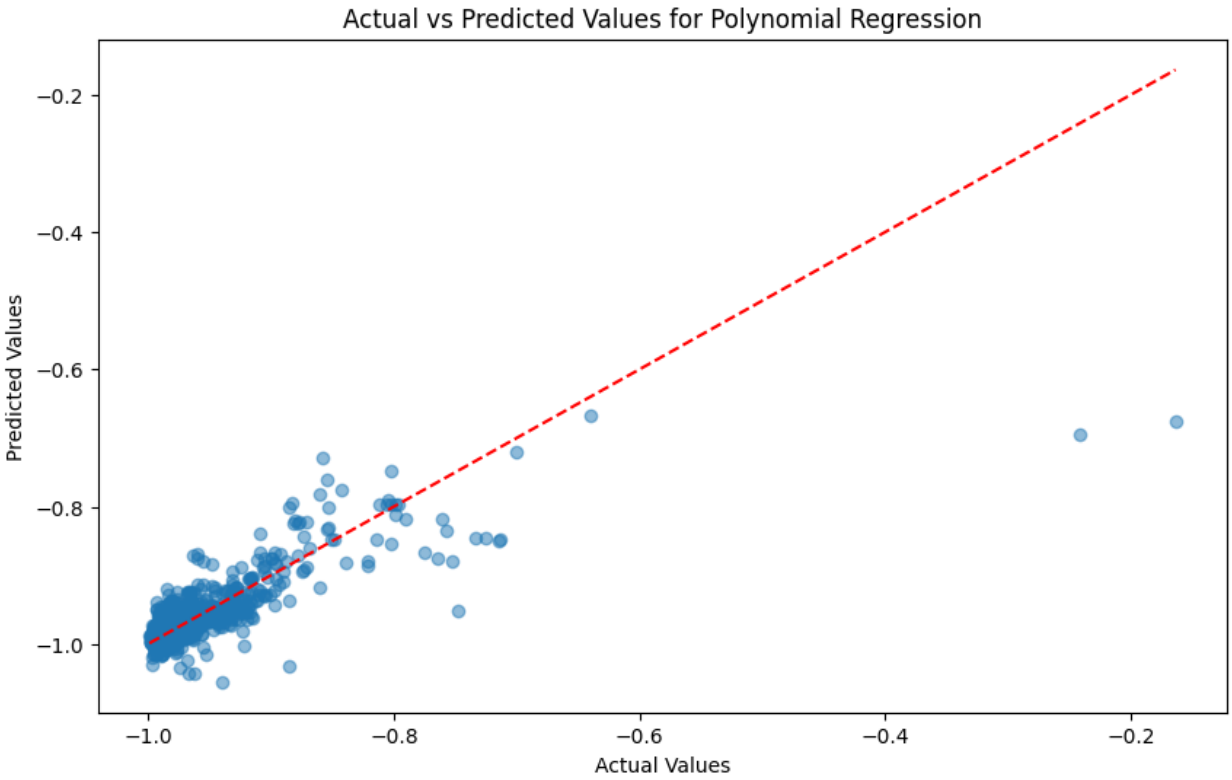
1- RBF Regression

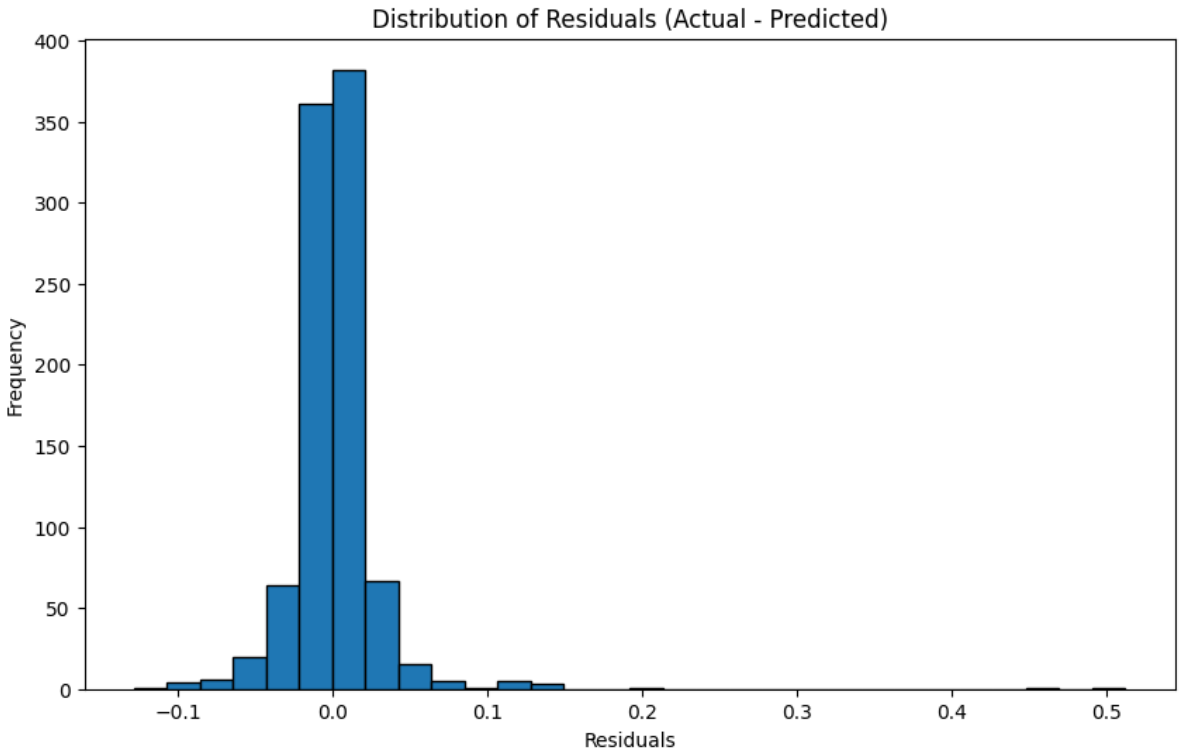




- Actual vs Predicted Values:
    - This scatter plot helps visualize how well the model's predictions align with the true values.
    - Points closer to the red dashed line (45-degree line) indicate better predictions, as they represent a perfect match between actual and predicted values.
    - Deviations from this line show areas where the model's predictions are less accurate.
- Residuals Distribution:
    - The histogram displays the distribution of residuals (errors between actual and predicted values).
    - A good model should have residuals centered around zero, with a roughly normal distribution (bell-shaped curve), indicating no significant bias in the predictions.
    - The spread of residuals tells us how consistent the model's errors are; a wider spread may indicate inconsistency in predictions.

2- Polynomial Regression





- Actual vs Predicted Values:
  - The scatter plot shows how close the predicted values are to the actual values.
  - The red dashed line represents perfect predictions. The closer the points are to this line, the more accurate the model is.
  - Since the model is performing well, we expect the points to be fairly close to this line.
- Residuals Distribution:
  - The histogram of residuals (differences between actual and predicted values) helps assess the model's error behavior.
  - Ideally, residuals should be centered around zero and show a symmetric distribution.