BIRZEIT UNIVERSITY

Faculty of Engineering and Technology

Electrical and Computer Engineering Department

MACHINE LEARNING AND DATA SCIENCE

Assignment #2

Regression Analysis and Model Selection

Prepared by:

Dana Assad	1211452
Mohammad Manasrah	1211407

Instructor:

Dr. Ismail Khater

Section: 3

Date: NOV/28/2024

Abstract

The goal of this research is to create machine learning models that forecast cars prices using a dataset of 6,750 car listings. The data is trained using a variety of regression approaches, such as polynomial regression, ridge regression, LASSO, linear regression, and radial basis function (RBF) kernel methods. To choose the best strategy based on performance measures, the models are verified on a test set. Furthermore, grid search is used for hyper-parameter tweaking in order to maximize model performance, and forward feature selection is utilized to determine the strongest predictors. In order to create a precise and trustworthy automobile price prediction system, the project intends to integrate feature selection and regularization approaches.

Contents

Table of figures

1. Dataset

The Cars Dataset from YallaMotors, sourced from an online car-selling platform (accessible at Kaggle), represents automotive market data from the Middle East. It features car listings from countries such as the UAE, Saudi Arabia, and others. This dataset includes 6,750 samples, covering attributes like price, horsepower, engine capacity, top speed, number of seats, cylinder count, brand, and country of listing.

```
Cars Dataset Overview:
The First 5 Cars:
                        car name                price engine_capacity      cylinder horse_power  top_speed     seat
0          Fiat 500e 2021 La Prima                TBD             0.0  N/A, Electric      Single  Automatic       15
1     Peugeot Traveller 2021 L3 VIP         SAR 140,575             2.0              4         180   8 Seater       8.
2   Suzuki Jimny 2021 1.5L Automatic         SAR 98,785             1.5              4         102        145  4 Seate
3     Ford Bronco 2021 2.3T Big Bend        SAR 198,000             2.3              4         420   4 Seater       7.
4     Honda HR-V 2021 1.8 i-VTEC LX  Orangeburst Metallic           1.8              4         140        190  5 Seate
Dataset Statistics:
The total number of cars: 6,308
Number of unique brands: 82
Countries represented: ksa, egypt, bahrain, qatar, oman, kuwait, uae
```

Figure 1: overview of the Cars Dataset

YallaMotors' Cars Dataset comprises 6,308 records showcasing 82 car brands in Middle Eastern markets like Saudi Arabia, Egypt, UAE, Kuwait, Oman, Qatar, and Bahrain. It includes car details like name, price, engine specs, seating.., etc., as shown in Fig1 above.

The bar chart 'in figure 2 below' displays car brand distribution in the YallaMotors dataset. Mercedes-Benz leads with 500+ listings, followed by Audi, BMW, Toyota, and Ford. These brands dominate the dataset, showing their popularity in the Middle East market. Less common brands like Bugatti, Tata, and Brilliance have minimal representation. This highlights the dominance of some brands in the regional market and the variety of choices for consumers.

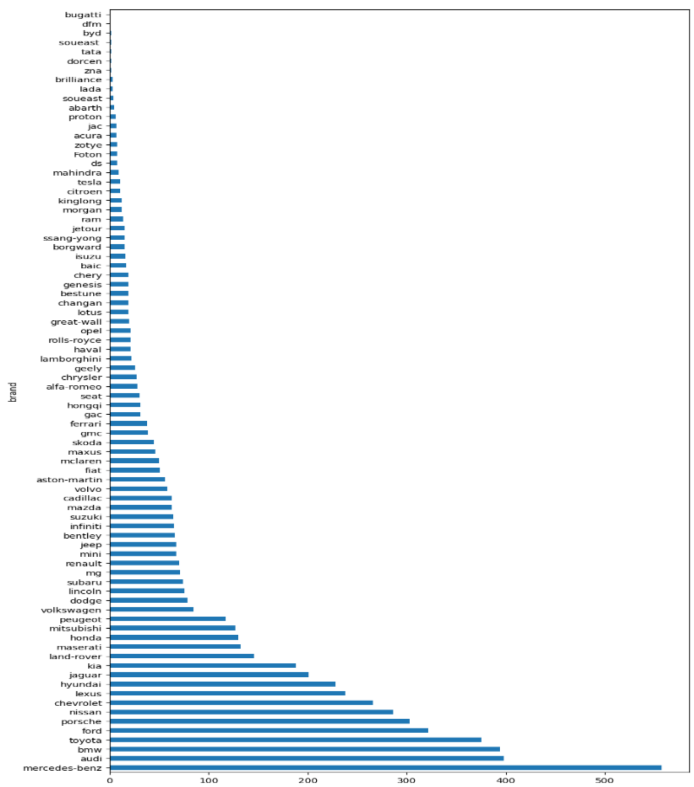


Figure 2: The Distribution of car brands

This Bar chart compares metrics in Egypt, Bahrain, Oman, Qatar, Kuwait, KSA, and UAE. It highlights the UAE's dominance in the dataset and shows clear variations between the countries.

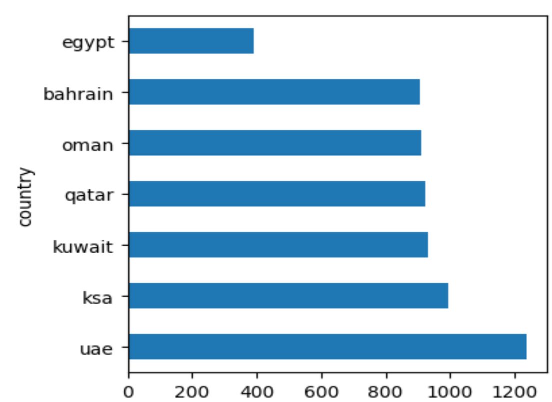


Figure 3: The distribution of cars by country

2. Data Preprocessing

Data preprocessing "refers to the essential step of cleaning and organizing data before it is used in a data-driven neural network algorithm. It involves removing any incorrect or irrelevant data and ensuring that the correct data is inputted into the models"

It is essential for preparing the YallaMotors dataset for analysis. It includes cleaning and organizing data for accuracy and consistency, addressing missing values like "TBD," encoding categorical variables, and scaling numerical features. Irrelevant columns are removed, and outliers in variables like price and horsepower are addressed to maintain data integrity. These steps ensure a clean, structured dataset for meaningful insights.

➔ Here is the first 5 row of data after **preprocessing:**

```
Cleaned Data Has Been Exported To: cleaned_cars_dataset.csv
First few rows of cleaned data:
                           car name         price  engine_capacity cylinder  horse_power   top_speed     seats
0           Fiat 500e 2021 La Prima  18127.809756           2500.0        4        255.0  183.727273  5.096144
1     Peugeot Traveller 2021 L3 VIP  37955.250000           2000.0        4        180.0  200.743119  9.000000
2  Suzuki Jimny 2021 1.5L Automatic  26671.950000           1500.0        4        102.0  145.000000  4.000000
3    Ford Bronco 2021 2.3T Big Bend  53460.000000           2300.0        4        420.0  185.128814  8.000000
4     Honda HR-V 2021 1.8 i-VTEC LX  27616.689111           1800.0        4        140.0  190.000000  5.000000
```

Figure 4: Dataset after preprocessing

➔ We implemented the following **preprocessing steps:**

- Removed duplicate rows in the dataset.
- Replaced missing values in 'cylinder' with None and Filled with **mode**
- Cleaned 'horse_power', replacing missing values with **median**
- Cleaned 'price', converting currencies and filling missing values
- Standardized 'engine_capacity', filling missing values with **median**
- Filled missing 'top_speed' values **with mean per brand**
- Cleaned 'seats' column, replacing outliers and filling missing values
- Encoding 'brand' and 'country' Using Weighted Frequency Encoding

Fully prepared dataset for modeling exported as encoded_standardized_cars_dataset.csv with 6,162 rows and 8 columns, as shown in figure below. And it is ready for analysis and machine learning applications.

```
Processed Data Has Been Exported To: encoded_standardized_cars_dataset.csv
Final Dataset Shape: (6162, 8)
Columns In Final DataSet: ['price', 'engine_capacity', 'cylinder', 'horse_power', 'top_speed', 'seats', 'brand', 'country']
```

3. Regression Models

3.1 Definition of Regression Models

A regression model provides a function that describes the relationship between one or more independent variables and a response, dependent, or target variable.

Regression models **aim** to predict a continuous target variable using input features. The procedure includes the utilization of both linear and nonlinear models, each designed to address distinct data patterns. Linear models are appropriate for less complicated relationships, whereas nonlinear models are able to depict more intricate patterns.

3.2 Type of Regression Models

- **Linear Models:**

1. **Linear Regression:**
   "Is a model that estimates the linear relationship between a scalar response (dependent variable) and one or more explanatory variables (regressor or independent variable)."
   - ➔ This forms the basis of regression analysis, where a straight line is determined by reducing the mean squared error (MSE) between predicted and actual values.

2. **LASSO Regression**
   "Is a regularization technique that applies a penalty to prevent overfitting and enhance the accuracy of statistical models".
   - ➔ Includes L1 regularization, which imposes a penalty on the absolute coefficients. This method both decreases overfitting and selects features by setting unimportant coefficients to zero.

3. **Ridge Regression**
   "Is a statistical regularization technique. It corrects for overfitting on training data in machine learning models."
   - ➔ Implements L2 regularization, penalizing the square of coefficients. This approach shrinks coefficients towards zero without completely eliminating them, making it useful for datasets with multicollinearity or numerous features.

Additionally, a **closed-form solution** for linear regression, derived from the normal equation, allows direct computation of coefficients and is efficient for small datasets. In contrast, gradient descent, an iterative optimization algorithm, scales better for larger datasets.

- **Nonlinear Models**
  1. **Polynomial Regression**
     "Is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an nth degree polynomial in x."
  ➔ This means: Adding polynomial terms of the input features. Exploring different polynomial degrees (e.g., ranging from 2 to 10) helps examine the balance between under-fitting and over-fitting.

  2. **Radial Basis Function (RBF)**
     A kernel-based method that maps inputs into a higher-dimensional space using Gaussian functions. RBF regression excels in modeling complex relationships, especially when the data isn't linearly separable.

Regularization prevents overfitting by penalizing intricate models, enhancing generalization to new data. LASSO (L1) regularization eliminates some coefficients by shrinking them to zero in order to select important features. In contrast, Ridge (L2) regularization decreases large coefficients without getting rid of any features, making it suitable for when all features are important. Grid Search can be used to optimize λ (lambda) in order to balance underfitting and overfitting, ensuring robust, flexible, and high-performing models

3.3 Results of Model Implementation

➔ In our assignment, Regression models were built using a combination of linear and nonlinear approaches. Linear models included Simple Linear Regression, LASSO, and Ridge Regression. Model parameters were derived using closed-form solution and gradient descent. Nonlinear models included Polynomial Regression and RBF regression to capture complex relationships in the data. After Split the dataset into 60% training, 20% validation, and 20% test sets.

➔ After building and training all the models, we analyzed their performance and compared the results using MSE and R^2 metrics (on Validation dataset) to identify the best-fit model for the data.

As shown in figure below, the **LASSO regression model** explains 70.6% of variance (R^2 = 0.706) with MSE of 948,971,368.21, RMSE of 30,805.38, and MAE of 20,450.07. Further improvements in accuracy possible with alpha tuning and data exploration.

```
LASSO Regression Results:
================================================================================

Alpha = 1 Metrics:
----------------------------------------
MSE:  948,971,368.21
RMSE: 30,805.38
MAE:  20,450.07
R2:   0.7060
{'Alpha': 1,
 'MSE': 948971368.2084042,
 'RMSE': 30805.37888435077,
 'MAE': 20450.068924520656,
 'R2': 0.7059760836095299}
```

Figure 5: LASSO Regression Result

Figures below shows R^2, MSE, MAE values For Ridge Regression and all other Models:

```
Ridge Regression Results:
================================================================================

Alpha = 1 Metrics:
----------------------------------------
MSE:  948,360,046.37
RMSE: 30,795.45
MAE:  20,440.93
R2:   0.7062
{'Alpha': 1,
 'MSE': 948360046.367692,
 'RMSE': 30795.45496283002,
 'MAE': 20440.928275187198,
 'R2': 0.7061654921078291}
```

Figure 6: Ridge Regression Result

```
Summary of All Polynomial Degrees:
================================================================================
 Degree          MSE        RMSE        MAE       R2
      1 9.490218e+08  30806.1982 20450.6379   0.7060
      2 7.270993e+08  26964.7783 17176.1678   0.7747
      3 6.144384e+08  24787.8673 15918.9541   0.8096
      4 1.201550e+09  34663.3873 15690.2062   0.6277
      5 1.269421e+11 356289.3772 36443.9453 -38.3310

Best Performing Degree:
Degree 3.0 with R2: 0.8096
```

Figure 7: Polynomial Degrees Result

```
Closed-Form Linear Regression Results:
================================================================================

Training Set Metrics:
----------------------------------------
MSE:  781,963,194.68
RMSE: 27,963.60
MAE:  18,975.16
R2:   0.7555

Validation Set Metrics:
----------------------------------------
MSE:  949,021,848.85
RMSE: 30,806.20
MAE:  20,450.64
R2:   0.7060
```

Figure 8: Closed-Form Linear Regression Result

```
Summary Results:
================================================================================
Metric     Training   Validation
   MSE 7.819632e+08 9.490218e+08
  RMSE 2.796360e+04 3.080620e+04
   MAE 1.897516e+04 2.045064e+04
    R2 7.555000e-01 7.060000e-01
{'Model': 'Closed-Form Linear Regression',
 'Train_MSE': 781963194.676245,
 'Train_RMSE': 27963.60482263052,
 'Train_MAE': 18975.15624215004,
 'Train_R2': 0.7554567098177134,
 'Val_MSE': 949021848.8533545,
 'Val_RMSE': 30806.198221354003,
 'Val_MAE': 20450.637882475676,
 'Val_R2': 0.7059604429722806,
 'Coefficients': array([ 1.16687436e+05, -3.75718228e+04,  4.91908601e+04,  8.20413094e+04,
        3.74348724e+04, -8.48446046e+03,  8.24388263e+01,  4.99065038e+03])}
```

Figure 9: Summary Result of Closed-Form Linear Regression

```
RBF Kernel Ridge Regression Results:
================================================================================

Alpha = 0.01, Gamma = 0.1 Metrics:
----------------------------------------
MSE:  626,609,176.42
RMSE: 25,032.16
MAE:  15,690.54
R2:   0.8059

Summary of All Parameter Combinations:
================================================================================
 Alpha  Gamma          MSE       RMSE        MAE     R2
  0.01    0.1 6.266092e+08 25032.1628 15690.5396 0.8059
```

Figure 10: RBF Kernel Ridge Regression Results

3.4 Comparison Results of All Models on the Validation Set

As shown in Figure 11, this summary evaluates machine learning models by looking at their validation R^2, MSE, RMSE, and MAE. The RBF Kernel model surpassed Ridge, LASSO, and Closed Form models with a validation R^2 of 0.8059, whereas the other models achieved 0.7060, which accounts for approximately 87.6% of the RBF Kernel's performance. Important measurements for the RBF Kernel consist of MSE at 626,610,595.99 and MAE at 15,690.47. Optimization was carried out for the model with hyper-parameters Alpha (0.01) and Gamma (0.1). In general, the RBF Kernel shows better performance, but there could be scope for enhancement.

```
Summary of All Parameter Combinations:
================================================================================
 Alpha  Gamma          MSE       RMSE        MAE     R2
  0.01    0.1 6.266106e+08 25032.1912 15690.4651 0.8059

Model Comparison Results:
================================================================================
      Model  Validation R2
 RBF Kernel         0.8059
      Ridge         0.7062
      LASSO         0.7060
Closed Form         0.7060

Best Performing Model:
Model: RBF Kernel
Validation R2: 0.8059

Relative Performance to Best Model:
================================================================================
      Model  Validation R2  Relative Performance (%)
 RBF Kernel         0.8059                  100.0000
      Ridge         0.7062                   87.6297
      LASSO         0.7060                   87.6062
Closed Form         0.7060                   87.6043

Best Model Metrics:
----------------------------------------
Alpha: 0.01
Gamma: 0.1
MSE:   626,610,595.99
RMSE:  25,032.19
MAE:   15,690.47
```

Figure 11: Model Comparison Results

➔ Including polynomial degrees Models the Comparison result will be as shown in Figure 12 below:

The results indicate that Polynomial Regression (degree=3) performs the best with a validation R^2 of 0.8096, slightly better than the RBF Kernel (R^2 = 0.8059). Ridge, LASSO, and Closed Form models achieved a notably lower score (R^2 = 0.7060). Compared to the top model, the RBF Kernel obtained a performance of 99.54%, whereas the rest achieved 87.2%. Important measures for Polynomial Regression comprise a Mean Squared Error of 614,438,364.18, Root Mean Squared Error of 24,787.87, and Mean Absolute Error of 15,918.95.

```
Model Comparison Results:
================================================================================
                Model  Validation R2
Polynomial (degree=3)         0.8096
           RBF Kernel         0.8059
                Ridge         0.7062
                LASSO         0.7060
          Closed Form         0.7060

Relative Performance to Best Model:
================================================================================
                Model  Validation R2  Relative Performance (%)
Polynomial (degree=3)         0.8096                  100.0000
           RBF Kernel         0.8059                   99.5430
                Ridge         0.7062                   87.2283
                LASSO         0.7060                   87.2036
          Closed Form         0.7060                   87.2036

Best Model Details:
----------------------------------------
Model: Polynomial Regression (degree=3)
MSE:  614,438,364.18
RMSE: 24,787.87
MAE:  15,918.95
R2:   0.8096
```

Figure 12: The Best Model

4. Feature Selection with Forward Selection

Forward selection is a feature selection method where a predictive model is built by starting with no features and adding one at a time. The process evaluates which feature results in the most significant improvement in model performance, reducing overfitting and improving interpretability. It stops when either no more improvement is seen with additional features or a maximum number of features is reached. This method is effective for high-dimensional datasets, focusing on features that impact the model's predictive power while ignoring those with little to no effect.

The process of forward feature selection involves adding features one by one based on how they enhance the model's performance. The initial characteristic, *horse_power*, results in the most notable enhancement, reaching an R^2 score of 0.6305 with a 100% relative improvement. Additional attributes such as *maximum velocity*, *number of cylinders*, and *engine size* continue to improve the model, but their impact lessens with each iteration. Upon reaching the 5th step, a minimal increase of 1.12% is observed when *seats* are included. Nevertheless, incorporating *brand* in the 6th stage does not enhance the model, as the R^2 stays the same and indicates a minor negative relative improvement (-0.0066%). This shows that the model's effectiveness levels off after the initial five features, indicating decreased benefits from more features as shown in figure 13

```
Starting Forward Feature Selection...

Feature Selection Progress:
------------------------------------------------------------

Step 1:
Added feature: horse_power
MSE: 1192590986.27
R² Score: 0.6305
Relative Improvement: 100.0000%

Step 2:
Added feature: top_speed
MSE: 1026875484.27
R² Score: 0.6818
Relative Improvement: 13.8954%

Step 3:
Added feature: cylinder
MSE: 993215460.87
R² Score: 0.6923
Relative Improvement: 3.2779%

Step 4:
Added feature: engine_capacity
MSE: 959646936.92
R² Score: 0.7027
Relative Improvement: 3.3798%

Step 5:
Added feature: seats
MSE: 948911424.06
R² Score: 0.7060
Relative Improvement: 1.1187%

Step 6:
Added feature: brand
MSE: 948974505.99
R² Score: 0.7060
Relative Improvement: -0.0066%
```

Figure 13: forward Selection Result

4.1 Feature Selection Result

Feature selection had an impact on the model's performance, especially for more complex models such as Polynomial degree 3 and RBF Kernel, resulting in slight decreases in their R^2 scores. Less complex models like Ridge and LASSO were less impacted, with their R^2 scores remaining at approximately 0.7060. This indicates that by selecting certain features, important characteristics necessary for accurate predictions in complex models may have been eliminated, resulting in a decrease in their ability to forecast outcomes.

```
Model Performance with Selected Features:
================================================================================
                  Model      R2           MSE        RMSE         MAE
4  Polynomial (degree=3)  0.7833  6.994106e+08  26446.3730  16984.1010
6                    RBF  0.7717  7.367228e+08  27142.6376  17020.1411
3  Polynomial (degree=2)  0.7599  7.748658e+08  27836.4112  17585.9593
5  Polynomial (degree=4)  0.7541  7.937552e+08  28173.6608  16024.3944
1                  Ridge  0.7060  9.489054e+08  30804.3089  20566.6606
0                  LASSO  0.7060  9.489110e+08  30804.3993  20566.7380
7            Closed Form  0.7060  9.489114e+08  30804.4059  20566.7444
2  Polynomial (degree=1)  0.7060  9.489114e+08  30804.4059  20566.7444
```

Figure 14: Model performance with selected Features

5. Regularization Techniques

Regularization methods like LASSO and Ridge regression help prevent overfitting by imposing a penalty on the coefficients of the model. LASSO (L1 regularization) promotes sparsity by possibly setting some coefficients to zero, effectively conducting feature selection. Ridge regression, also known as L2 regularization, decreases coefficients without eliminating them, decreasing model complexity while keeping all features. Our goal is to find the right balance between model accuracy and generalization on the validation set by trying out different regularization parameters (λ) and using Grid Search to optimize them.

→ The Grid Search Result:

```
Grid Search Results:
============================================================
LASSO Best λ: 166.810054
LASSO Validation MSE: 945211122.60
LASSO Validation R2: 0.7071

Ridge Best λ: 14.849683
Ridge Validation MSE: 945145219.77
Ridge Validation R2: 0.7072
```

Figure 15: Grid Search Result

As shown in figure above, the optimal regularization parameters for LASSO and Ridge regression were identified according to the results of the Grid Search. The optimal value of **(λ) for LASSO** was found to be **166.810054**, leading to a validation MSE of 945,211,122.60 and a (R^2) score of 0.7071. The optimal **(λ) value for Ridge** regression was found to be **14.849683**, resulting in a validation MSE of 945,145,219.77 and a (R^2) score of 0.7072. These findings underscore how both regularization techniques effectively balance model complexity and performance.

6. Hyper-parameter Tuning with Grid Search

Utilizing Grid Search for hyperparameter tuning is an essential process in enhancing model performance by methodically exploring for the optimal set of hyperparameters. Grid Search was employed to determine the best regularization parameter value **(λ)** for models such as LASSO and Ridge regression. By testing the model's performance on a validation set for every potential **(λ)** **,** we verified that the selected parameter reduces error and improves generalization. This method ensures that the models are adjusted to strike a balance between accuracy and complexity, ultimately enhancing their ability to predict on new data.

➔The Hyperparameter Tuning Result:

```
Tuning LASSO...

Tuning Ridge...

Tuning Polynomial...

Tuning RBF...

Final Results:
================================================================================

LASSO:
----------------------------------------
Validation MSE: 948,835,764.33
Validation R²: 0.7060
Best Parameters: {'lasso__alpha': 10.0, 'lasso__max_iter': 3000, 'lasso__tol': 0.0001}

Ridge:
----------------------------------------
Validation MSE: 948,073,166.49
Validation R²: 0.7063
Best Parameters: {'ridge__alpha': 10.0, 'ridge__tol': 0.0001}

Polynomial:
----------------------------------------
Validation MSE: 614,438,364.18
Validation R²: 0.8096
Best Parameters: {'poly__degree': 3, 'poly__interaction_only': False}

RBF:
----------------------------------------
Validation MSE: 464,224,606.74
Validation R²: 0.8562
Best Parameters: {'rbf__alpha': 0.01, 'rbf__gamma': 0.1}

Best Model: RBF
Validation MSE: 464,224,606.74
Validation R²: 0.8562
```

Figure 16: Hyper-parameter Tuning Result

The performance of various models is underscored by the Grid Search results. Out of all the models, the RBF model outperformed the LASSO, Ridge, and Polynomial models with a validation MSE of 464,224,606.74 and an (R^2) score of 0.8562, achieving the best results. The best hyperparameters for each model were found, showing how fine-tuning can enhance both accuracy and generalization.

7. Model Evaluation on Test Set

Evaluating the model is an essential part of determining the performance of a trained model on unseen data. It entails evaluating the model's capacity to generalize by assessing it on an independent test set that was not utilized during training or validation.

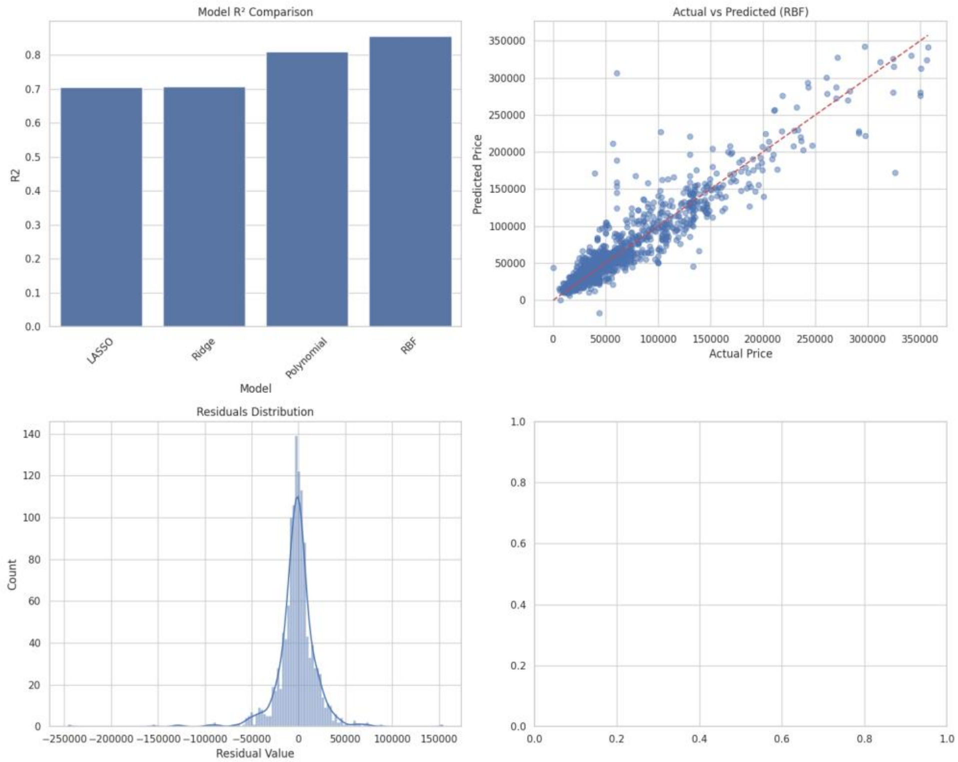
Once the top model has been chosen using performance metrics from the validation set, the next phase involves assessing the selected model on the test set. This assessment offers a final determination of how effectively the model generalizes to data that has not been previously seen. By examining how well the model performs on the test set in comparison to its performance on the training and validation sets, we can assess whether the model is suffering from overfitting or underfitting. A strong model will exhibit consistent performance on these datasets, suggesting it has grasped the fundamental patterns in the data and can effectively forecast outcomes on unfamiliar data. The ultimate evaluation on the test data provides important information about the model's durability and its capacity to make dependable predictions in real-world situations.

→ Model Evaluation Result:

```
Best Model Performance on Test Set:
Model: RBF
Best Parameters: {'rbf__alpha': 0.01, 'rbf__gamma': 0.1}
  Model      Test_MSE       Test_MAE   Test_R2      Test_RMSE
0   RBF  3.769578e+08  11911.896162  0.869672  19415.399986
```

Figure 17: Model Evaluation Result

The RBF model achieved an R^2 score of 0.8697, accounting for 87% of data variance, using alpha=0.01 and gamma=0.1. The test mean squared error (MSE) was 3.769578e+08, mean absolute error (MAE) was 11911.90, and the root mean squared error (RMSE) was 19415.40, indicating precise predictions and strong generalization.



- **Top Left**: **Model R^2 Comparison**

This bar graph shows the R^2 scores for four regression models: LASSO, Ridge, Polynomial, and RBF, comparing their coefficient of determination. The R^2 score evaluates how effectively each model clarifies the variation in the target variable (Actual Price), with superior scores (nearer to 1) indicating enhanced performance. Out of the models, RBF attains the highest R^2 score, indicating its superior ability to capture and elucidate the variability present in the data.

- **Top Right: Actual vs Predicted Prices (RBF)**

The graph illustrates the relationship between the Actual Price (on the x-axis) and Predicted Price (on the y-axis) for the RBF model, with a red dashed line showing the perfect alignment of predicted and actual prices ($y=x$). The points closely group together on this line, showing that the RBF model has high predictive accuracy.

- **Bottom Left: Residuals Distribution**

The histogram shows how the residuals, which are the discrepancies between real and estimated prices, are distributed. The residuals cluster close to 0 and have a generally symmetrical form, indicating unbiased errors and a normal distribution, indicating a good model fit. Nevertheless, there are a few exceptions on each end that demonstrate when the model struggled to accurately predict.

8. Conclusion

Using a dataset from the Middle Eastern automotive sector, this study effectively built and assessed a number of regression models to forecast automobile prices. Both linear and nonlinear models were shown to be beneficial in the investigation; the best-performing models were the Radial Basis Function (RBF) kernel and Polynomial Regression (degree=3). With an R2 score of 0.8697 on the test set, RBF demonstrated a high degree of generalization and accurate prediction of unknown variables. Additionally helpful in striking a balance between model complexity and performance were regularization techniques such as Ridge regression and LASSO.

The importance of feature selection and hyperparameter adjustment in maximizing model accuracy are among the main conclusions. The most important characteristics were emphasized using forward feature selection, and regularization parameters were adjusted via grid search to enhance generalization. The study emphasizes how crucial it is to select the ideal feature and model combination for challenging tasks like price prediction.

Overall, the results emphasize the potential of advanced regression techniques in building reliable predictive models for real-world applications, providing a foundation for future improvements through deeper exploration of data and hyperparameter spaces.

Group Work Policy

Dana:

1. Data Cleaning and Preprocessing
3. Model Selection Using Validation Set
4. Feature Selection with Forward Selection
7. Model Evaluation on Test Set

And write these sections of the report.

Mohammad:

2. Building Regression Models
5. Applying Regularization Techniques
6. Hyperparameter Tuning with Grid Search

And write these sections of the report.

9. References

1 >>
https://www.imsl.com/blog/what-is-regression model#:~:text=A%20regression%20model%20provides%20a,by%20a%20linear%20regression%20modl

2 >>
https://en.wikipedia.org/wiki/Polynomial_regression#:~:text=In%20statistics%2C%20polynomial%20regression%20is,nth%20degree%20polynomial%20in%20x.

3>> https://en.wikipedia.org/wiki/Radial_basis_function