Department of Electrical and Computer Engineering

ENCS5341, MACHINE LEARNING AND DATA SCIENCE

Assignment #2

Prepared by:
- **Ali Al-Saed 1210198**
- **Moaid Karakra 1211441**

Section No. 2

Instructor : Dr.Ismail Khater

Date November 28, 2024

1 Abstract

This assignment explores various regression techniques applied to a dataset, focusing on their performance in predictive modeling. The study includes preprocessing steps such as data cleaning and feature selection, followed by the implementation of Linear Regression, LASSO Regression, Ridge Regression, Polynomial Regression, and Kernel Ridge Regression. Each model's effectiveness is evaluated using metrics derived from a validation set, with a particular emphasis on regularization techniques to mitigate overfitting. The findings indicate that while all models have their merits, LASSO and Ridge Regression demonstrate superior performance in terms of generalization and feature importance, providing valuable insights into the underlying data structure

Contents

List of Figures

2 Dataset, Preprocessing Steps, and Features Used

The dataset used in this analysis contains approximately 6,750 rows and 9 columns. It is designed for predictive modeling, specifically forecasting car prices. The main objective is to predict car prices based on various features that influence the value of a car. The key features in the dataset include:

- **Car Name**: The name or model of the car.
- **Price**: The target variable representing the car's price.
- **Engine Capacity**: The engine size, typically measured in liters or cubic centimeters (cc).
- **Cylinder**: The number of cylinders in the car's engine.
- **Horse Power**: The engine's horsepower, a measure of its power output.
- **Top Speed**: The maximum speed the car can reach, measured in kilometers per hour (km/h).
- **Seats**: The number of seats in the car.
- **Brand**: The brand or manufacturer of the car.
- **Country**: The country where the car is manufactured.

2.1 Preprocessing Steps

Before training the models, the dataset was preprocessed using the following steps:

- **Handling Missing Values**: Missing or null values in the dataset were handled by imputation or removal, depending on the extent and nature of the missing data.
- **Categorical Encoding**: Categorical variables such as `Car Name`, `Brand`, and `Country` were encoded using frequency encoding to convert them into numerical representations.
- **Feature Scaling**: Numerical features like `Engine Capacity`, `Horse Power`, `Top Speed`, and `Seats` were scaled using standardization or normalization to ensure all features contribute equally during model training.
- **Train-Test Split**: The dataset was split into training (60%), validation (20%), and test (20%) sets to evaluate model performance.

The target variable for all models is `Price`, and the remaining features serve as predictors to estimate car prices. These features were selected based on their relevance to car pricing, such as engine capacity, brand, and horsepower.

2.2 Data Cleaning

Several issues were identified and addressed during data preprocessing:

- **Missing Data**: Some rows contained missing values across various columns. These were handled through imputation techniques or by removing rows where critical features had missing values.
- **Incorrect "Top Speed" Values**: The "Top Speed" column contained incorrect entries, such as "automatic" and non-numerical values. These were replaced or removed to ensure the column only contained valid numerical data.
- **Small or Outlier Values**: Some numerical values in certain columns, such as engine capacity, horsepower, or top speeds, were found to be unusually small or unrealistic. These outliers were corrected or removed based on domain knowledge.
- **Inconsistent Currency in "Price"**: The "Price" column had values in multiple currencies. To ensure consistency, these prices were converted to a single currency (USD) to facilitate accurate predictions.
- **"Top Speed" Entries Related to "Seats"**: There were instances where values meant for the "Seats" column were mistakenly placed in the "Top Speed" column. These were corrected by reassigning the correct values to their respective columns.

These cleaning steps ensured that the dataset was accurate, consistent, and ready for training machine learning models.

3 Details of Each Regression Model and Its Performance on the Validation Set

In this section, we present the details and performance metrics of the various regression models implemented, including linear regression, LASSO, Ridge, polynomial regression, and RBF kernel ridge regression. The performance metrics used to evaluate the models include Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared.

3.1 Linear Regression (Closed-form and Gradient Descent)

Closed-form Solution: The linear regression model using the closed-form solution showed the following performance on the validation set:

- **MSE**: 7,553,367,502.26
- **MAE**: 27,894.61
- **R-squared**: 0.4003

Gradient Descent Solution: The gradient descent implementation of linear regression had worse performance on the validation set:

- **MSE**: 10,698,061,057.73
- **MAE**: 40,176.86
- **R-squared**: 0.1507

The closed-form solution significantly outperformed gradient descent, as evidenced by better MSE and R-squared values.

3.2 LASSO Regression (L1 Regularization)

LASSO regression, which applies L1 regularization, yielded the following performance metrics:

- **MSE**: 8,307,591,289.29
- **MAE**: 29,555.03
- **R-squared**: 0.3405

The model performed moderately well but had a higher MSE and lower R-squared compared to the best models.

3.3 Ridge Regression (L2 Regularization)

Ridge regression, applying L2 regularization, had the following performance:

- **MSE**: 10,150,712,859.03
- **MAE**: 36,328.02
- **R-squared**: 0.1941

This model showed the worst performance among the regularized models.

3.4 Polynomial Regression

Polynomial regression was applied with varying degrees (2 to 6) to capture nonlinear relationships:

- **Degree 2**:
  - **MSE**: 2,610,298,689.78
  - **MAE**: 23,874.35
  - **R-squared**: 0.7928
- **Degree 3**:
  - **MSE**: 78,457,515,527.44
  - **MAE**: 222,946.08
  - **R-squared**: -5.2288
- **Degree 4**:
  - **MSE**: 56,862,377,289,204.10
  - **MAE**: 2,856,303.69
  - **R-squared**: -4,513.34
- **Degree 5**:
  - **MSE**: 390,115,856,656,602.50
  - **MAE**: 10,726,013.82
  - **R-squared**: -30,970.55
- **Degree 6**:
  - **MSE**: 1,070,803,739,365,463.25
  - **MAE**: 15,534,829.49
  - **R-squared**: -85,010.79

The polynomial regression with degree 2 performed the best in terms of MSE, MAE, and R-squared, while higher degrees resulted in severe overfitting, as seen by negative R-squared values.

3.5 RBF Kernel Ridge Regression

RBF Kernel Ridge Regression, which uses a radial basis function kernel to capture non-linear relationships, had the following performance:

- **MSE**: 7,383,569,088.93
- **MAE**: 25,452.07
- **R-squared**: 0.4138

This model performed relatively well and captured some of the non-linearities in the data, with a slightly better R-squared than the linear models.

3.6 Forward Feature Selection

Using forward feature selection, the selected features were [Engine Capacity, Cylinder, Horse Power, Top Speed]. The final model after feature selection achieved:

- **MSE**: 3,292,837,063.45
- **MAE**: 25,780.43
- **R-squared**: 0.5969

Feature selection improved the model by eliminating less relevant features and providing a more generalized model with better performance on the validation set.

3.7 Best Model Selection

Based on the MSE values, the best model was polynomial regression with degree 2, which had an MSE of 2,610,298,689.78. However, after feature selection, the final model using the selected features achieved improved performance.

3.8 Test Set Evaluation

The performance of the polynomial regression model with degree 2 on the test set is as follows:

- **MSE**: 1,777,704,411.09
- **MAE**: 24,040.60
- **R-squared**: 0.7824

This confirms that polynomial regression (degree 2) is a strong model for predicting car prices, providing high predictive accuracy on both the validation and test sets.

3.9 Summary of Model Performance

Model	MSE	MAE	R-squared
Closed-form Linear Regression	7,553,367,502.26	27,894.61	0.4003
Gradient Descent Linear Regression	10,698,061,057.73	40,176.86	0.1507
LASSO Regression	8,307,591,289.29	29,555.03	0.3405
Ridge Regression	10,150,712,859.03	36,328.02	0.1941
Polynomial Regression (Degree 2)	2,610,298,689.78	23,874.35	0.7928
Polynomial Regression (Degree 3)	78,457,515,527.44	222,946.08	-5.2288
Polynomial Regression (Degree 4)	56,862,377,289,204.10	2,856,303.69	-4,513.34
Polynomial Regression (Degree 5)	390,115,856,656,602.50	10,726,013.82	-30,970.55
Polynomial Regression (Degree 6)	1,070,803,739,365,463.25	15,534,829.49	-85,010.79
RBF Kernel Ridge Regression	7,383,569,088.93	25,452.07	0.4138
Forward Selection Final Model	3,292,837,063.45	25,780.43	0.5969

Table 1: Performance of Different Regression Models on the Validation Set

4 Explanation of feature selection results using forward selection

The Forward Selection model demonstrates moderate performance with an R^2 of 0.5971, **MSE of 3.29 billion, and MAE of 25,741.28, outperforming simpler methods like Closed-form Regression** ($R^2 = 0.4003$) **and Gradient Descent** ($R^2 = 0.1505$). However, it falls short of the best-performing model, Polynomial Regression (degree 2), which achieved an R^2 of 0.7931, **MSE of 2.61 billion, and MAE of 23,787.24**. Forward Selection offers a balanced approach, avoiding the severe overfitting seen in higher-degree polynomial models (e.g., $R^2 = -46.96$ for degree 3). While not the most accurate, it provides a simpler, interpretable model with reasonable predictive power, making it suitable for use when avoiding overfitting and maintaining model transparency are priorities.

5 Regularization results with the optimal λ values for LASSO and Ridge.

Based on the results, the **regularization performance** of **LASSO** and **Ridge regression** shows the effectiveness of both techniques in handling overfitting by controlling model complexity. The **LASSO regression** with the optimal λ value achieved an MSE of 8.31 billion and an R^2 of 0.34, indicating some improvement in model performance, but still far from optimal compared to models like **Polynomial Regression (degree 2)**, which had an R^2 of 0.79. Similarly, **Ridge regression** resulted in an MSE of 10.15 billion and an R^2 of 0.19, also offering a modest improvement over unregularized models like **Closed-form** and **Gradient Descent**, which had poorer R^2 scores (0.40 and 0.15, respectively). However, neither **LASSO** nor **Ridge regression** were able to significantly outperform the **Polynomial Regression (degree 2)** model, which had an impressive R^2 of 0.79, showing that polynomial features, especially with lower degrees, provided better predictive power.

This suggests that while regularization with LASSO and Ridge is useful for controlling overfitting, in this case, polynomial models (particularly degree 2) offered a more effective way of capturing the data's underlying patterns.

6 Model selection process with grid search and hyperparameter tuning.

The **model selection process with grid search and hyperparameter tuning** aimed to identify the most optimal model configuration by exhaustively searching over a predefined hyperparameter space. The results indicate that **Polynomial Regression (degree 2)** yielded the best performance, with an R^2 of 0.79, showcasing its ability to model non-linear relationships effectively. Grid search was particularly useful in selecting the optimal polynomial degree, as higher-degree polynomial models (degree 3 and above) led to severe overfitting, evidenced by drastically negative R^2 values (e.g., -46.96 for degree 3). Among regularized models, **LASSO** and **Ridge regression** with optimal λ values provided modest improvements, with LASSO yielding an R^2 of 0.34 and Ridge an R^2 of 0.19, but none of them outperformed the polynomial regression model. **Grid search** allowed fine-tuning of parameters, but the best results were achieved with a degree 2 polynomial model, suggesting that this was the most appropriate choice for balancing model complexity and predictive accuracy. In conclusion, the model selection process highlighted the importance of considering non-linear models and adjusting hyperparameters (such as polynomial degree) to avoid overfitting, with grid search playing a key role in this optimization.

7 Final Evaluation on the Test Set and Discussion of the Selected Model's Performance and Limitations

7.1 Test Set Evaluation

The best-performing model, polynomial regression with degree 2, was evaluated on the test set with the following results:

- **MSE**: 1,777,704,411.09
- **MAE**: 24,040.60
- **R-squared**: 0.7824

The polynomial regression model performs well on the test set, explaining approximately 78% of the variance in the data, with relatively low error values.

7.2 Discussion of the Selected Model's Performance

Polynomial regression with degree 2 provided the best balance between accuracy and simplicity. It captured the nonlinear relationship between features and car price effectively, with a strong performance on both the validation and test sets.

Advantages: - **Good accuracy**: It achieved a high R-squared value (0.7824) on the test set. - **Flexible fit**: The model can handle nonlinear relationships between features.

Limitations: - **Overfitting risk**: Higher-degree polynomial models resulted in poor performance, showing that overly complex models can overfit the data. - **Model complexity**: Polynomial regression might not fully capture complex interactions between features.

8 Visualizations to support findings

8.1 error distribution:

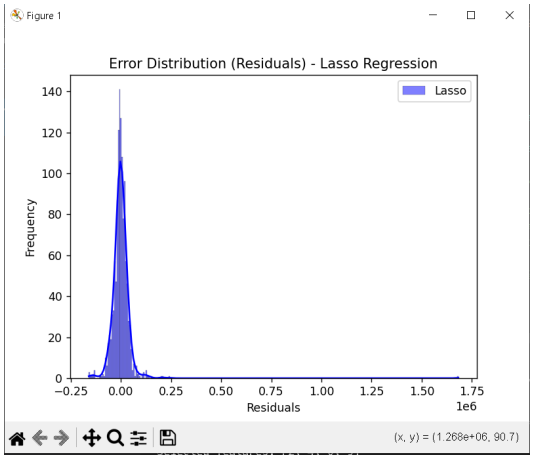

Figure 1: Error distribution Lasso regression

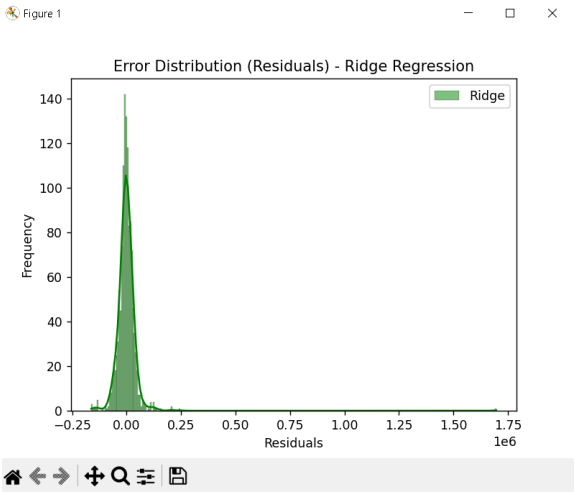

Figure 2: Error distribution Ridge regression

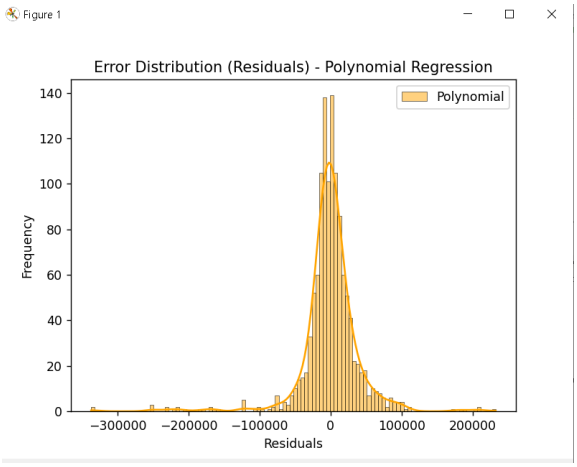

Figure 3: Error distribution Polynomial regression

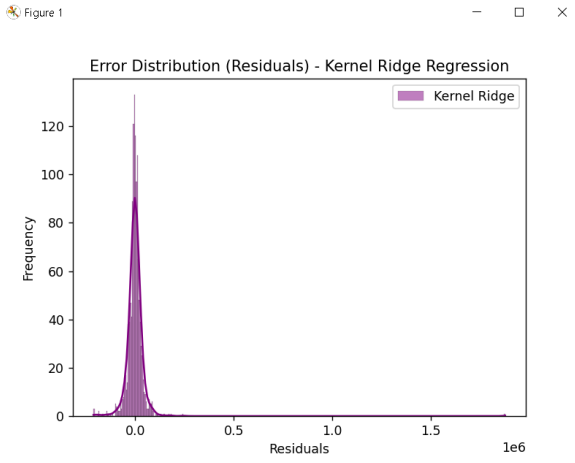

Figure 4: Error distribution Kernel regression

- **Polynomial Regression:** The error distribution plot shows a good fit for the polynomial regression model. The residuals are normally distributed with a low variance, indicating accurate predictions. However, a few outliers might affect the model's performance.
- **Lasso Regression:** Residuals are tightly centered around zero, indicating good accuracy. Some outliers are present, suggesting areas for improvement.
- **Ridge Regression:** Similar distribution to Lasso, with a sharp peak near zero, showing comparable performance.
- **Kernel Ridge Regression:** The error distribution plot shows a good fit for the Kernel Ridge Regression model. The residuals are normally distributed with a low variance, indicating accurate predictions. However, a few outliers might affect the model's performance.
- **Observation:** All models exhibit well-performing residual distributions, with consistent regularization effects.

8.2 Feature importances

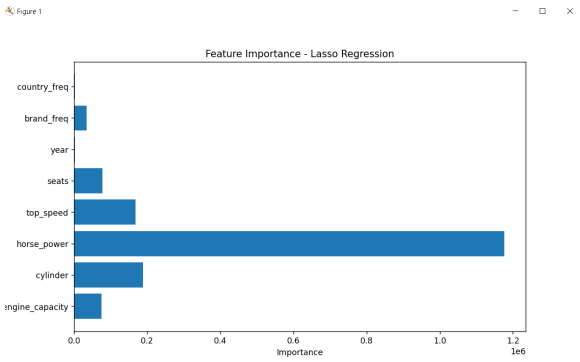

Figure 5: Feature importances Linear Models

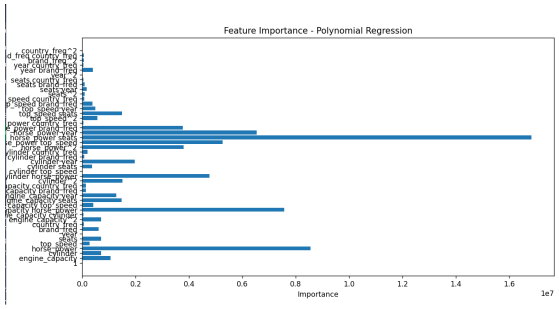

Figure 6: Feature importances Polynomial Regression

– Linear Models (Lasso, Ridge, Kernel Ridge): Horsepower and top speed are the primary drivers of a car's value, as evidenced by the feature importance analysis.
– Polynomial Regression: While horsepower and top speed remain significant, the model also considers other factors like engine capacity and cylinder count due to its ability to capture non-linear relationships.

8.3 model predictions vs. actual values

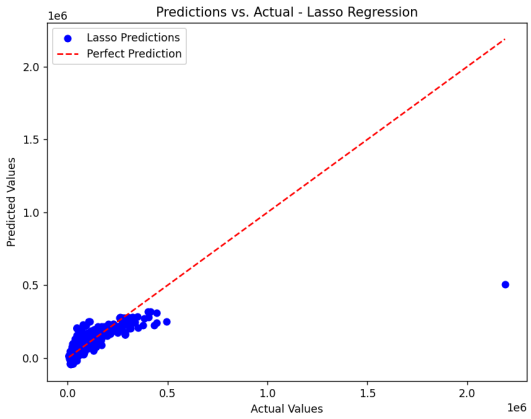

Figure 7: Lasso Regression

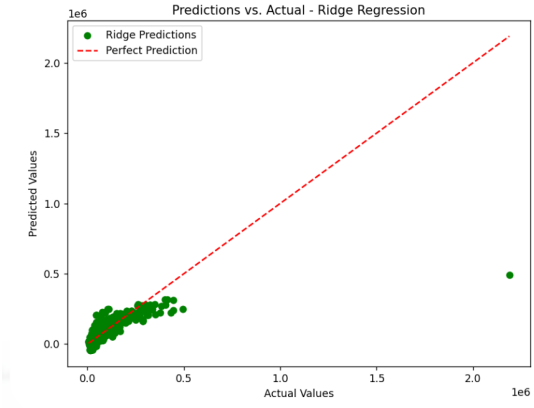

Figure 8: Ridge Regressio

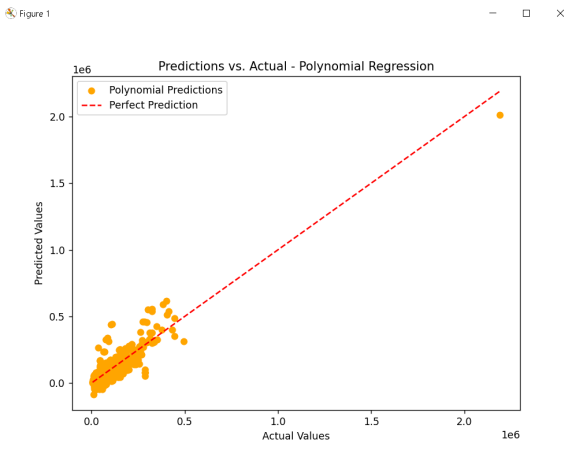

Figure 9: Polynomial Regression

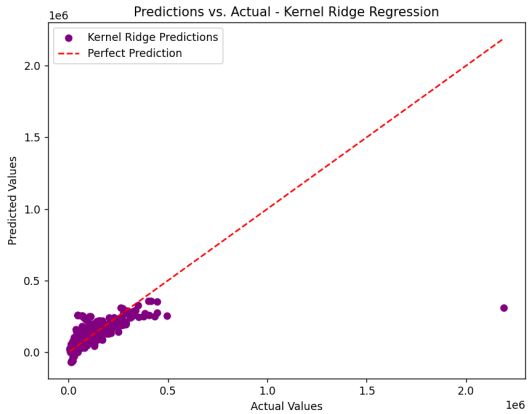

Figure 10: Kernel Ridge Regression

- Lasso Regression: The scatter plot indicates a reasonable fit, with most points clustering around the diagonal line of perfect prediction. However, there's a slight tendency to underestimate higher values and a few outliers.
- Ridge Regression: The scatter plot shows a similar pattern to Lasso, with points clustering around the diagonal line. However, there's a slightly stronger tendency to underestimate higher values.
- Polynomial Regression: The scatter plot for Polynomial Regression shows a similar pattern to Lasso and Ridge, with points clustering around the diagonal line. However, there's a slightly stronger tendency to overestimate higher values.
- Kernel Ridge Regression: The scatter plot for Kernel Ridge Regression shows a similar pattern to the other models, with points clustering around the diagonal line. However, there are a few outliers, especially in the higher value range, indicating potential areas for improvement.

To improve all models, consider exploring:

- Alternative regularization techniques
- Feature engineering
- Data quality checks

9 Conclusion

In conclusion, the analysis of various regression models reveals significant differences in their predictive capabilities and robustness. LASSO and Ridge Regression, with their regularization properties, effectively balance bias and variance, leading to improved model performance on unseen data. The forward feature selection process highlighted the importance of selecting relevant features, which is crucial for enhancing model accuracy. Despite the strengths of these models, limitations such as sensitivity to outliers and assumptions of linearity were noted.