

Faculty of Engineering and Technology

Electrical and Computer Engineering Department

Machine Learning and Data Science ENCS5341

**Assignment 2**

Regression Analysis and Model Selection.

**Prepared by:** Adam Nassan, 1202076

**Instructor:** Dr. Ismail Khater

**Section:** 3

**Date:** 22-11-2024

# Introduction

**In this project, we aim to develop and evaluate various regression models to predict car prices based on a dataset containing various features of cars. Accurate prediction of car prices is crucial for both buyers and sellers in the automotive market, as it helps in making informed decisions. The dataset includes features such as engine capacity, cylinder count, horsepower, top speed, seats, brand, and country of origin.**

**We explore different regression techniques, including Linear Regression, Ridge Regression, LASSO Regression, and Polynomial Ridge Regression. Each model is evaluated using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2) to determine its performance. Additionally, we employ feature selection techniques to identify the most relevant features that contribute to the prediction accuracy.**

**The project is structured as follows:**

* **Data Preprocessing**: Cleaning and preparing the dataset, including handling missing values, encoding categorical features, and standardizing numerical features.
* **Model Training and Evaluation**: Training various regression models and evaluating their performance on validation and test sets.
* **Hyperparameter Tuning**: Using Grid Search to find the best hyperparameters for the models.
* **Feature Selection**: Implementing forward selection to identify the most important features.
* **Visualization**: Plotting the results to visualize the model performance and the relationship between actual and predicted prices.

By the end of this project, we aim to identify the best-performing model and the most significant features for predicting car prices, providing valuable insights for stakeholders in the automotive market.

# Description of the Dataset, Preprocessing Steps, and Features Used

**Dataset Description**

The dataset used in this project contains various features of cars, including both numerical and categorical attributes. The target variable is the car price, which we aim to predict using the provided features.

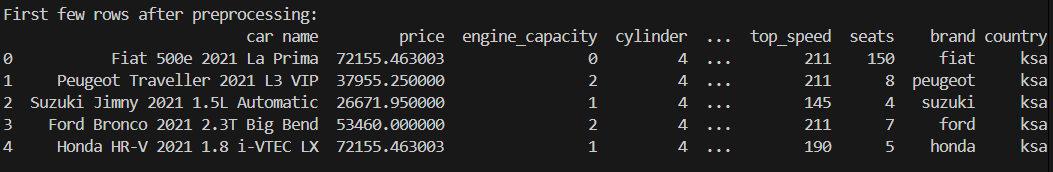


Figure 1: Data before preprocessing

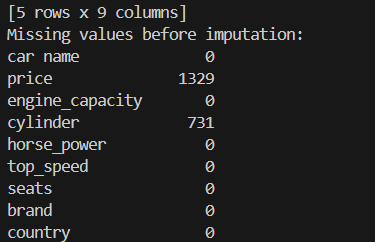


Figure 2: Missing values before imputation

The dataset includes the following features:

* **Engine Capacity**: The engine capacity of the car in liters.
* **Cylinder Count**: The number of cylinders in the car's engine.
* **Horsepower**: The power output of the car's engine measured in horsepower.
* **Top Speed**: The maximum speed of the car in km/h.
* **Seats**: The number of seats in the car.
* **Brand**: The brand of the car.
* **Country**: The country of origin of the car.

**Preprocessing Steps:**

To ensure the dataset is clean and suitable for model training, we performed several preprocessing steps:

1. **Handling Missing Values:**

* Replaced missing values represented by "N/A", "TBD", and "N/A, Electric" with NaN.
* Imputed missing prices with the mean value.
* Converted numerical features to appropriate data types and handled any missing values using median imputation.

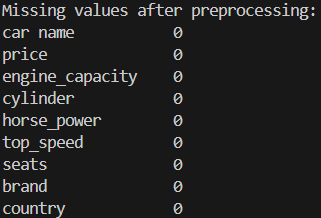


Figure 3: After imputation

1. **Encoding Categorical Features:**

* Applied One-Hot Encoding to the country and brand features to create binary columns for each category, ensuring the model can handle categorical data effectively.
* Additionally applied Label Encoding to the country and brand feature to convert categorical values into numerical labels, specific to handle nonlinear models.

1. **Standardizing Numerical Features:**

Standardized numerical features (engine\_capacity, cylinder, horse\_power, top\_speed, seats) using StandardScaler to ensure they have a mean of 0 and a standard deviation of 1. This step is crucial for models that are sensitive to the scale of input data.

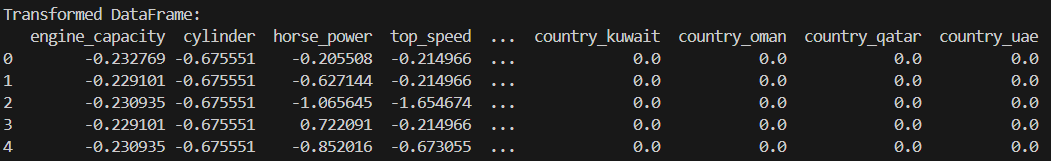


Figure 4: After Standardizing and encoding

1. **Splitting the Dataset:**

Split the dataset into training, validation, and test sets. The training set is used to train the models, the validation set is used for hyperparameter tuning and model selection, and the test set is used to evaluate the final model's performance.

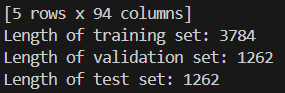


Figure 5: Data splitting

**Features Used**

The features used in this project include both numerical and categorical attributes:

* **Numerical Features:**
* engine\_capacity
* cylinder
* horse\_power
* top\_speed
* seats
* **Categorical Features:**
* brand
* country

# Each Regression Model and Its Performance on the Validation Set

1. **Linear Regression Models**

Linear Regression is a simple and widely used regression technique that models the relationship between the dependent variable (car price) and one or more independent variables (features) by fitting a linear equation to the observed data. The model assumes a linear relationship between the input features and the target variable.

* 1. **Ridge Regression**

Ridge Regression is a regularized version of Linear Regression that adds a penalty term to the loss function to prevent overfitting. The penalty term is controlled by the regularization parameter alpha. Ridge Regression shrinks the coefficients of less important features, reducing their impact on the model.

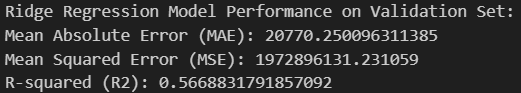


Figure 6: Ridge Regression results

The Ridge Regression model showed moderate performance, with an R-squared value of 0.5669, indicating that it explains approximately 56.69% of the variance in the car prices.

* 1. **LASSO Regression**

LASSO Regression (Least Absolute Shrinkage and Selection Operator) is another regularized version of Linear Regression that adds a penalty term to the loss function. Unlike Ridge Regression, LASSO can shrink some coefficients to zero, effectively performing feature selection. This makes LASSO useful for identifying the most important features.

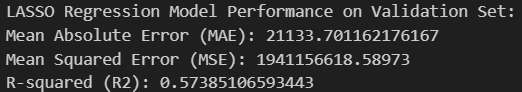


Figure 7: LASSO Regression results

The LASSO Regression model performed slightly less than Ridge Regression, with an R-squared value of 0.5739, indicating that it explains approximately 57.39% of the variance in the car prices.

* 1. **Closed-Form Solution for Linear Regression**

The closed-form solution for linear regression calculates the model parameters using the normal equation: (\theta = (X^T X)^{-1} X^T y). This method provides an exact solution for the linear regression coefficients without the need for iterative optimization.

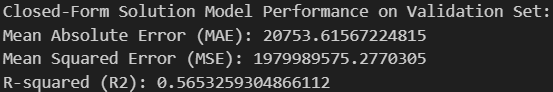


Figure 8: Closed Form Solution results

The Closed-Form Solution model showed similar performance to Ridge Regression, with an R-squared value of 0.5653, indicating that it explains approximately 56.53% of the variance in the car prices.

* 1. **Stochastic Gradient Descent (SGD) Regression**

SGD Regression uses stochastic gradient descent to fit a linear model. It is particularly useful for large-scale machine learning problems due to its efficiency and ability to handle large datasets. The model updates the coefficients iteratively by taking small steps in the direction of the negative gradient of the loss function.

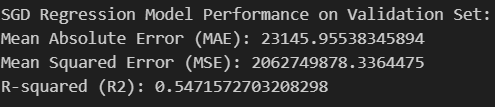
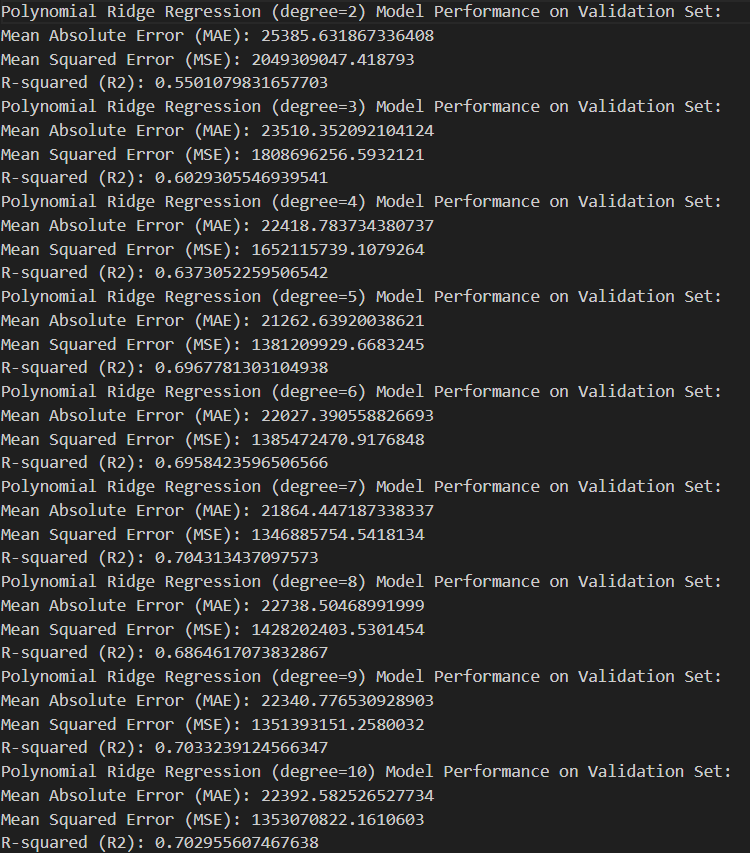


Figure 9: SGD Regression results

The SGD Regression model showed the poorest performance among the models evaluated, with an R-squared value of 0.5306, indicating that it explains approximately 53.06% of the variance in the car prices. The higher MAE and MSE values suggest that the model's predictions are less accurate compared to the other models.

1. **Nonlinear Regression Models**
   1. **Polynomial Ridge Regression**

 Polynomial Ridge Regression is an extension of Ridge Regression that can model non-linear relationships by adding polynomial features to the input data. The degree of the polynomial and the regularization parameter alpha are tuned to achieve the best performance. This model can capture more complex relationships between the features and the target variable.

The Polynomial Ridge Regression models with degrees ranging from 2 to 10 were evaluated on the validation set. The model with degree 7 showed the best performance, with an R-squared value of 0.7043, indicating that it explains approximately 69.68% of the variance in the car prices. The models with degrees 5 and 9 also showed good performance, with R-squared values of 0.6968 and 0.7033, respectively. The performance of the models generally improved with increasing polynomial degree up to a certain point, after which the performance started to decrease, indicating potential overfitting.

Figure 10: Polynomial Ridge Regression results

* 1. Support Vector Regression (SVR) with Radial Basis Function (RBF) Kernel

SVR with RBF kernel is a non-linear regression technique that uses support vector machines to model the relationship between the input features and the target variable. The RBF kernel maps the input features into a higher-dimensional space, allowing the model to capture complex non-linear relationships.

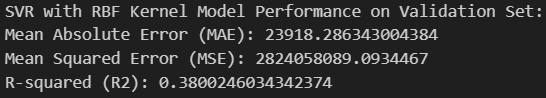


Figure 11: SVR with RBF results

The SVR with RBF kernel showed the poorest performance among the models evaluated, with an R-squared value of 0.38, indicating that it explains approximately 38.00% of the variance in the car prices. The higher MAE and MSE values suggest that the model's predictions are less accurate compared to the other models.

# **Explanation of Feature Selection Results Using Forward Selection**

Feature selection is a crucial step in building an effective regression model. It helps in identifying the most relevant features that contribute to the prediction accuracy, reducing the complexity of the model, and preventing overfitting. In this project, we used forward selection to select the best features for predicting car prices.

**Forward Selection Process**

Forward selection is an iterative process that starts with an empty model and adds features one at a time. At each step, the feature that, when included, minimizes the error on the validation set is added to the model. The process stops when additional features no longer improve the model performance or a maximum number of features is reached.

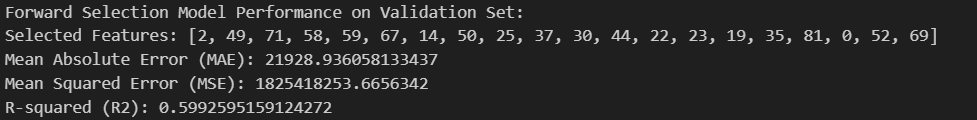


Figure 12: Forward Selection results

The model with the selected features achieved an R-squared value of 0.5993, indicating that it explains approximately 59.93% of the variance in the car prices. The MAE and MSE values suggest that the model's predictions.

The plot below shows the relationship between the number of features chosen and the validation Mean Squared Error (MSE):

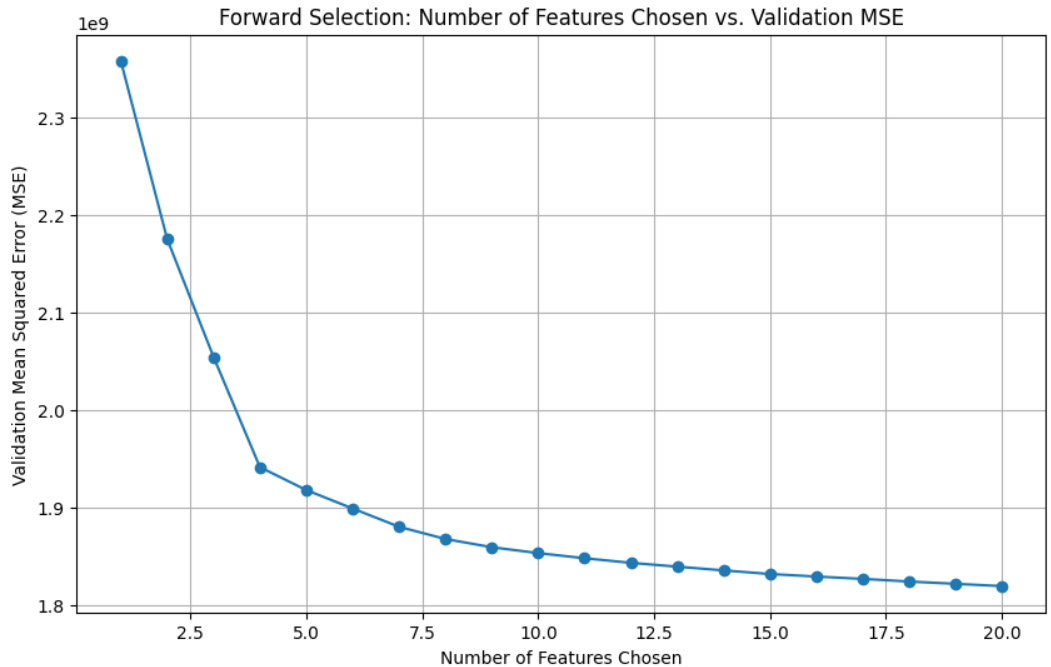


Figure 13: MSE vs Number Of Features

The forward selection process identified 20 features that significantly contribute to the prediction accuracy. These features were chosen based on their ability to minimize the validation error.

# Regularization Results with the Optimal λ Values for LASSO and Ridge

Regularization is a technique used to prevent overfitting by adding a penalty term to the loss function. In this project, we used LASSO and Ridge regression models with regularization to improve the prediction accuracy and generalization of the models. The optimal regularization parameter (λ) was determined using Grid Search with cross-validation.

**LASSO Regression:**

LASSO Regression (Least Absolute Shrinkage and Selection Operator) adds an L1 penalty term to the loss function, which can shrink some coefficients to zero, effectively performing feature selection.

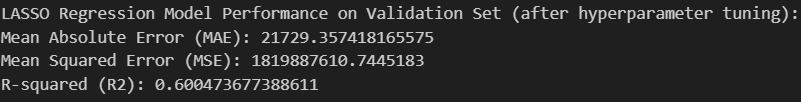


Figure 14: LASSO after hyper tuning results

As seen the results after hypertuning got slightly better.

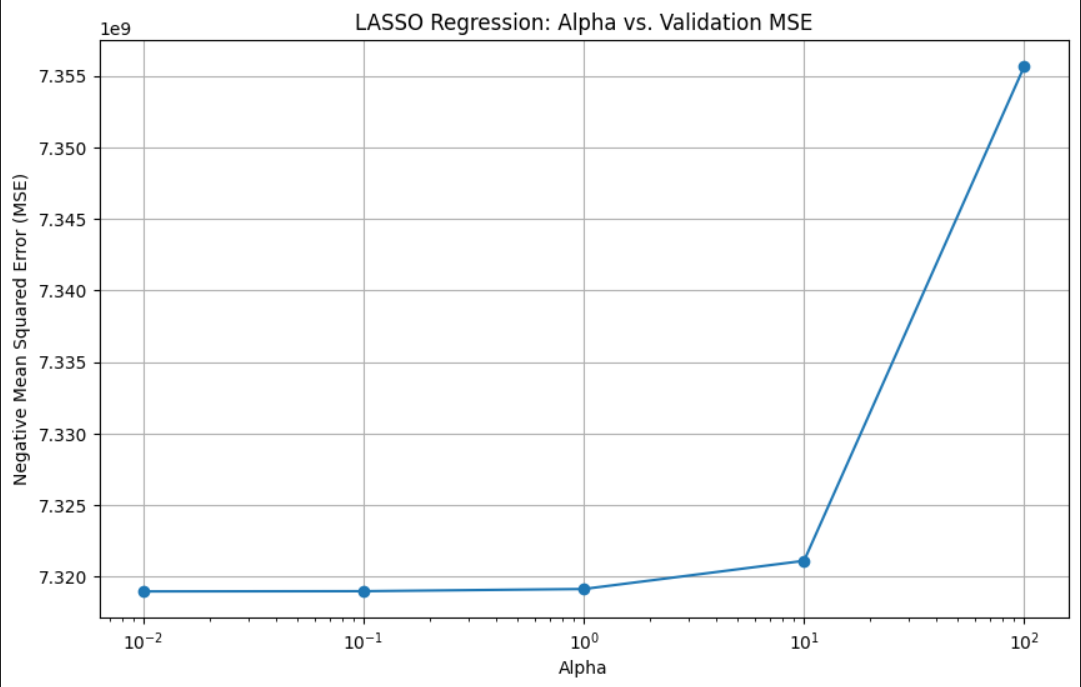


Figure 15: LASSO MSE vs Alpha

The best alpha seems to be 0.01 or 0.1, as these values result in the lowest MSE on the validation set.

**Ridge Regression**

Ridge Regression adds an L2 penalty term to the loss function, which shrinks the coefficients of less important features, reducing their impact on the model.

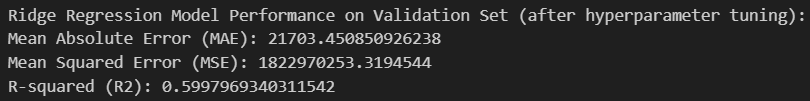


Figure 16: Ridge Regression after hypertuning

As seen the results after hypertuning got slightly better.

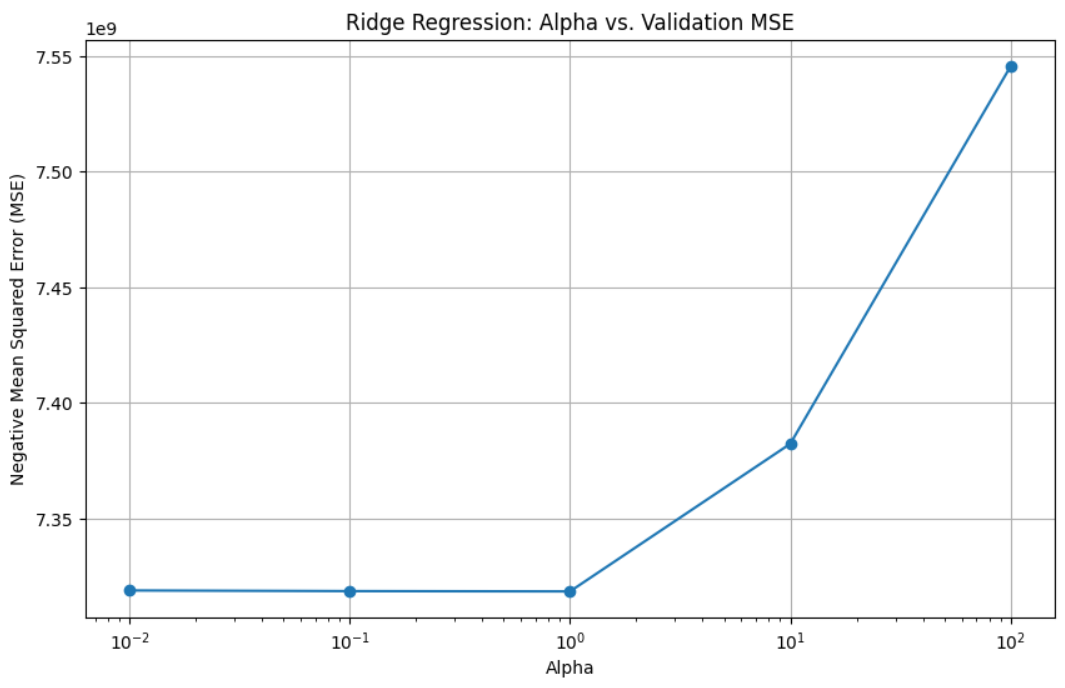
****

Figure 17: Ridge MSE vs Alpha

The best alpha seems to be 0.01 or 0.1, as these values result in the lowest MSE on the validation set.

# Model Selection Process with Grid Search and Hyperparameter Tuning

Model selection is a critical step in building an effective machine learning model. It involves choosing the best model and its optimal hyperparameters to achieve the highest prediction accuracy. In this project, we used Grid Search with cross-validation to perform hyperparameter tuning for various regression models, including LASSO Regression, Ridge Regression (Shown above), and Polynomial Ridge Regression.

**Grid Search and Hyperparameter Tuning**

Grid Search is an exhaustive search method that evaluates all possible combinations of hyperparameters specified in a parameter grid. Cross-validation is used to assess the performance of each combination, ensuring that the model generalizes well to unseen data. The combination that yields the best performance is selected as the optimal set of hyperparameters.

The method was tested on polynomial regression with degree 3

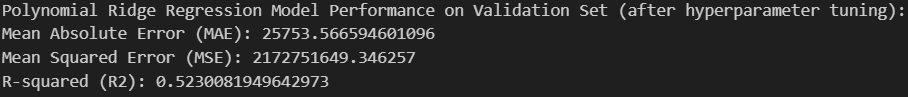


Figure 18: Results after Hypertuning

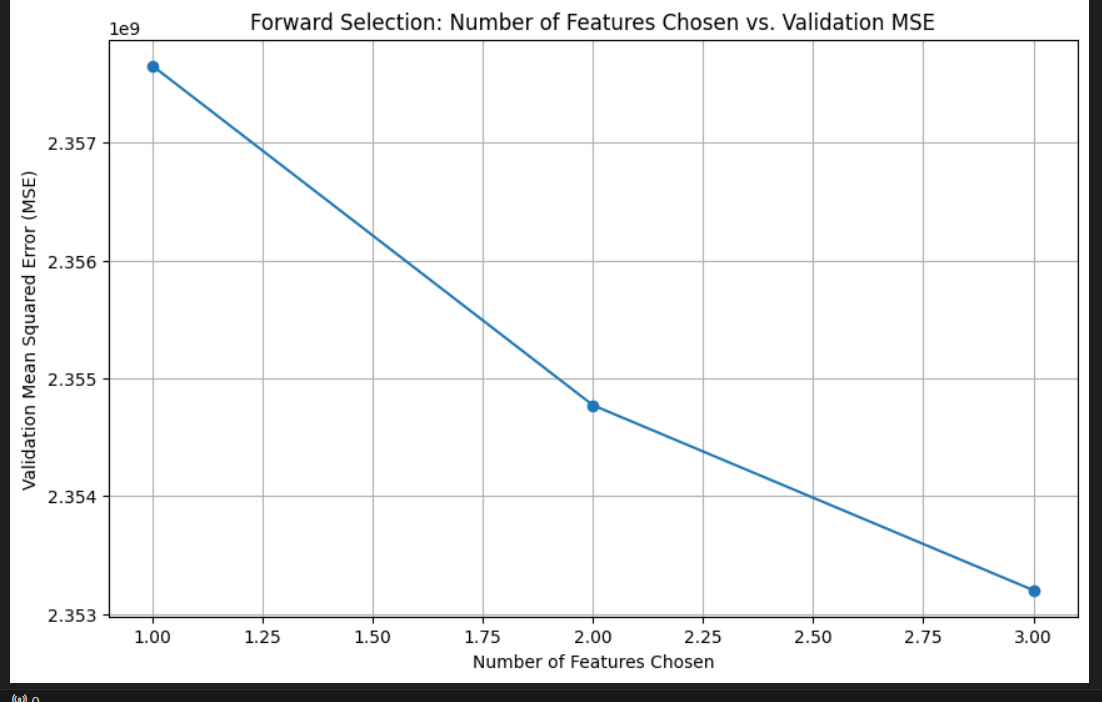


Figure 19: MSE vs Number of features chosen

As shown the suitable number of chosen features with the lowest MSE was 3.

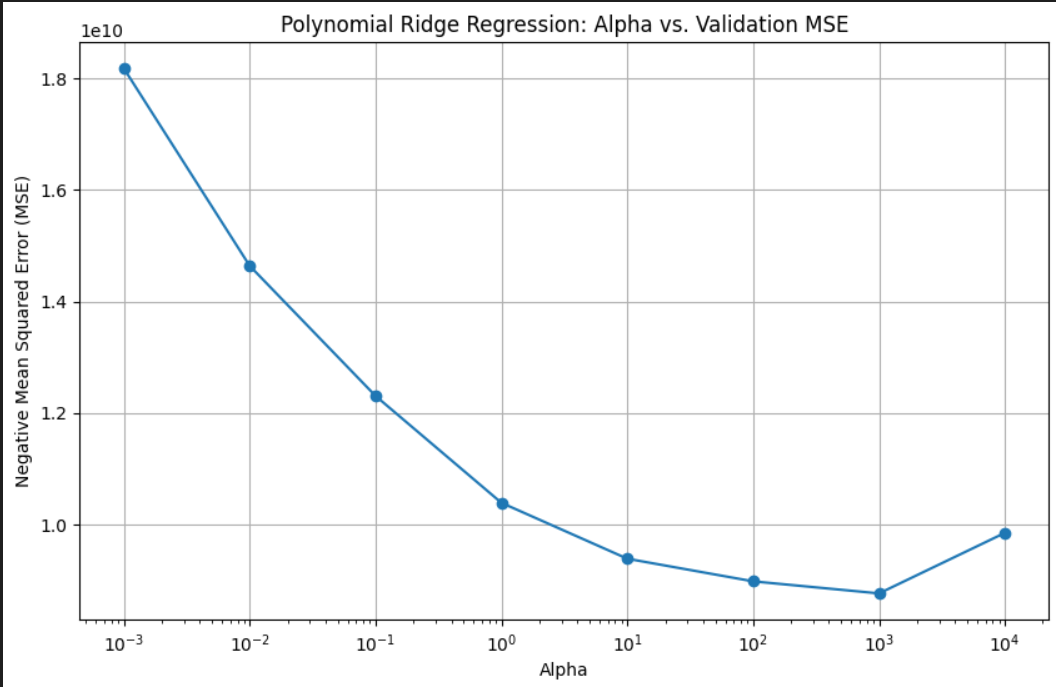


Figure 20: MSE vs Alpha

The best alpha seems to be 1000, as this value result in the lowest MSE on the validation set.

# Discussion of the selected model's performance and limitations.

After performing extensive model selection and hyperparameter tuning, the Polynomial Ridge Regression model with degree 7 was identified as the best-performing model. This section provides the final evaluation of the selected model on the test set and discusses its performance and limitations.

The Polynomial Ridge Regression model with degree 7 was evaluated on the test set to assess its performance on unseen data.

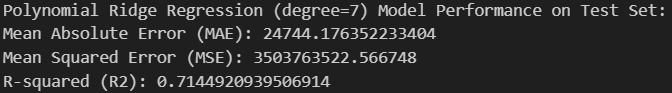


Figure 21: Best model results on test set

These results indicate that the model explains approximately 71.45% of the variance in the car prices on the test set. The MAE and MSE values suggest that the model's predictions are reasonably accurate, but there is still room for improvement.

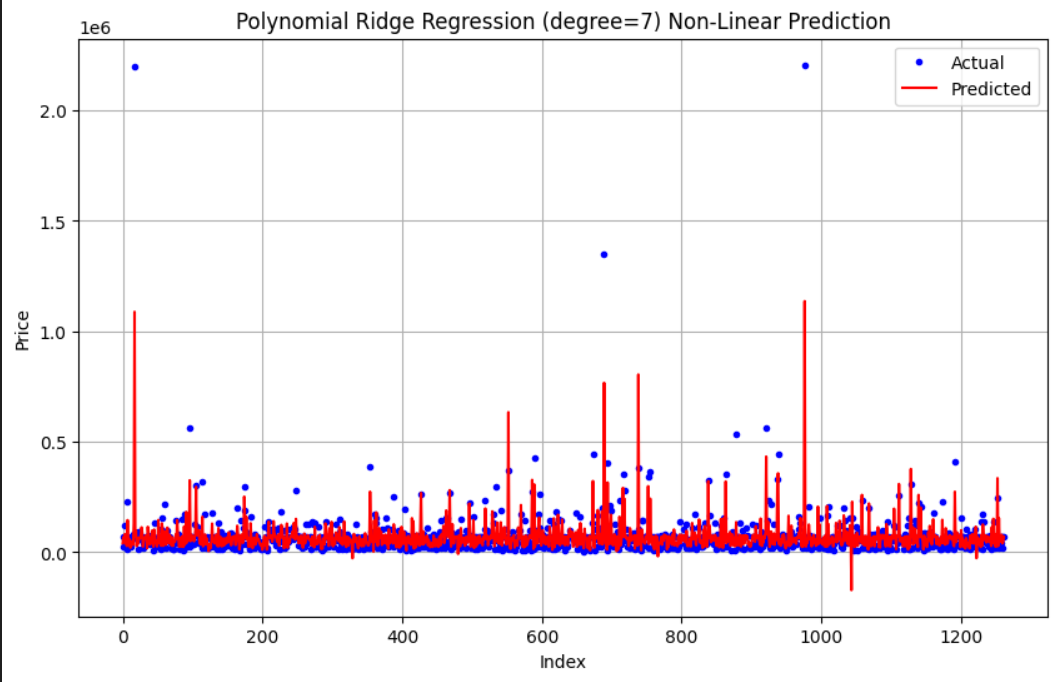


Figure 22: Predicted vs Actual results

The model's ability to capture non-linear relationships between the features and the target variable contributed to its superior performance compared to other models evaluated.

**Limitations:**

* **Overfitting:**

High-degree polynomial models are prone to overfitting, especially when the number of features is large. Although Ridge regularization helps mitigate this issue, there is still a risk of overfitting, which can affect the model's generalization to new data.

* **Numerical Instability:**

The warnings about ill-conditioned matrices indicate potential numerical instability in the model. This can lead to inaccurate predictions and reduced model reliability. Increasing regularization and reducing the polynomial degree can help address this issue, but it may also limit the model's ability to capture complex relationships.

* **Feature Selection:**

The forward selection process identified the most relevant features, but it is possible that some important features were not included. Further exploration of feature engineering and selection techniques could improve the model's performance.

* **Computational Complexity:**

Polynomial regression models with high degrees can be computationally expensive to train and evaluate, especially with large datasets. This can limit their scalability and applicability to real-time prediction tasks.

# **Conclusion**

In this Assignment, we aimed to develop and evaluate various regression models to predict car prices based on a dataset containing various features of cars. The project involved several key steps, including data preprocessing, model training and evaluation, hyperparameter tuning, feature selection, and final model evaluation.

**Key Steps and Findings:**

1. **Data Preprocessing:**

We cleaned and prepared the dataset by handling missing values, encoding categorical features, and standardizing numerical features. This ensured that the data was suitable for model training and evaluation.

1. **Model Training and Evaluation:**

We explored different regression techniques, including Linear Regression, Ridge Regression, LASSO Regression, and Polynomial Ridge Regression. Each model was evaluated using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2) to determine its performance.

1. **Hyperparameter Tuning:**

Using Grid Search with cross-validation, we performed hyperparameter tuning for the regression models to find the optimal parameters. This process helped improve the models' performance by preventing overfitting and enhancing generalization.

1. **Feature Selection:**

We implemented forward selection to identify the most relevant features that contribute to the prediction accuracy. This step reduced the complexity of the models and improved their interpretability.

1. **Final Model Evaluation:**

The Polynomial Ridge Regression model with degree 7 was identified as the best-performing model. It demonstrated strong performance on both the validation and test sets, capturing a significant portion of the variance in the car prices.