TUNIS BUSINESS SCHOOL

Car Price Prediction
Econometrics Project Report

SUBMITTED BY:

AYA SAADAOUI

MAJOR: INFORMATION TECHNOLOGY

MINOR: BUSINESS ANALYTICS

JANUARY, 2025

Abstract

This study focuses on predicting car selling prices using regression models, including Linear Regression, Polynomial Regression, and regularized models such as Ridge and Lasso. The goal was to select the most effective model based on performance metrics like MAE, MSE, RMSE, and R2 score. The dataset was cleaned and preprocessed, and various assumptions related to linearity, homoscedasticity, and normality were tested. The results revealed that Polynomial Regression (Degree 2) provided the best balance between model complexity and predictive accuracy. This report details the methodology, analysis, and key findings from the study, offering actionable insights for future improvements.

TABLE OF CONTENTS

1. INTRODUCTION

The automotive industry plays a crucial role in the global economy, and accurate pricing is vital for both buyers and sellers. Predicting car selling prices can be a complex task due to the various features influencing the price, such as car brand, age, mileage, and features. The objective of this project was to predict car selling prices using regression models and assess their performance. The models tested included Linear Regression, Polynomial Regression, and regularized models (Ridge and Lasso). The aim was to identify the model that best fits the data and provides reliable predictions.

2. Methodology

2.1. Dataset

The dataset consists of several attributes related to used cars, including features such as brand, model, year, mileage, engine size, and more. The target variable for this regression task was the Selling Price.

2.2. Analytical Methods

The following steps were taken to ensure accurate model development:

- **Data Cleaning:** Missing values were handled using the `SimpleImputer` from scikit-learn, ensuring no gaps in the dataset. Multicollinearity was checked using the Variance Inflation Factor (VIF), and features with high VIF were removed.
- **Exploratory Data Analysis (EDA):** Various visualizations were used to identify trends and relationships between the features and the target variable. Scatter plots, histograms, and box plots were employed to inspect distributions, correlations, and detect any potential outliers.
- **Assumption Testing:** The assumptions of linear regression were validated through:
  - **Linearity:** Scatter plots were examined.

  - **Homoscedasticity:** Residual plots were analyzed.
  - **Normality of Residuals:** Q-Q plots and the Shapiro-Wilk test were used.
  - **No Autocorrelation:** The Durbin-Watson statistic was calculated.
- **Modeling:** Several regression models were tested:
  - Linear Regression
  - Polynomial Regression (with degrees 2 and 3)
  - Ridge Regression
  - Lasso Regression
- **Performance Metrics:** Models were evaluated based on Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R2 Score.

2.3. Tools Used

- **Programming Language:** Python
- **Libraries:** NumPy, Pandas, scikit-learn, Matplotlib, Seaborn, Statsmodels

3. Analysis and Results

3.1. Key Findings

Polynomial Regression (Degree 2) outperformed the other models with the lowest MAE (0.72), MSE (2.84), and RMSE (1.69). It also achieved the highest R2 Score (0.87), indicating that the model explained 87% of the variance in the target variable.

Linear Regression and Multiple Linear Regression produced similar results, with R2 Score of 0.85, making them less effective than Polynomial Regression (Degree 2).

Polynomial Regression (Degree 3), Ridge, and Lasso Regression suffered from overfitting, resulting in poor performance and negative R2 scores, indicating poor generalization.

3.2. Visualizations

The following visualizations were used to gain deeper insights into the model performance and data characteristics:

- **Scatter Plots:** Used to check the relationships between features and the target variable.
- **Residual Plots:** Visualized the spread of errors, confirming the homoscedasticity assumption.
- **Q-Q Plots:** Examined the normality of residuals.
- **Box Plots:** Identified potential outliers.

3.3. Correlation Matrix Analysis

The correlation matrix below shows the relationships between car features. Strong positive correlations exist between:

- **Selling Price & Present Price:** Higher present price means higher selling price.
- **Age & Kilometers Driven:** Older cars tend to have more kilometers.

There are also strong negative correlations between:

- **Selling/Present Price & Seller Type:** Seller type seems to influence price negatively.

Many other relationships are weak. The diagonal shows perfect correlation (a variable with itself). Red indicates positive correlation, and blue indicates negative correlation.

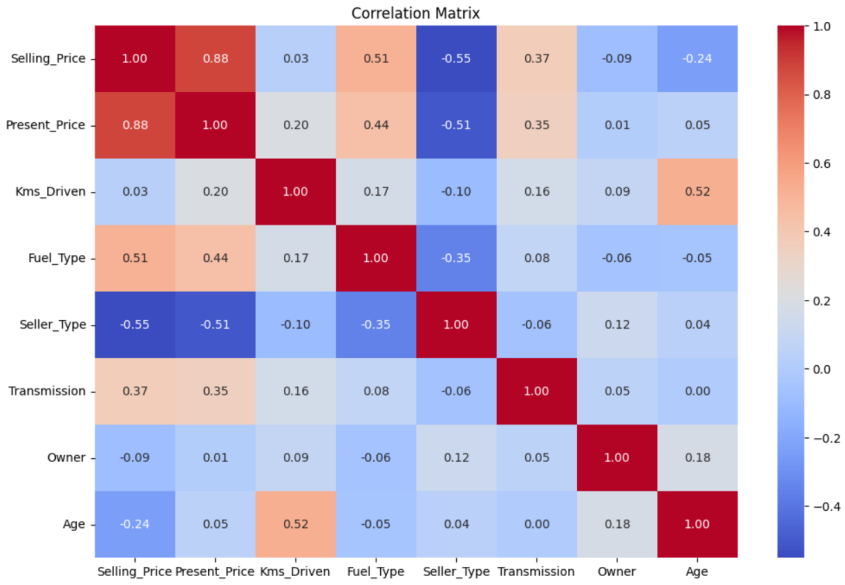

Figure 3.1: Correlation Matrix of Car Features

3.4. Analysis of Visualizations

Histogram of Popularity (Figure 3.2): This histogram shows the distribution of a variable called ”Popularity.” The x-axis represents the popularity values ranging from 0 to 80, and the y-axis represents the count or frequency of these values. The counts range from 0 to 3000, indicating how frequently each popularity value occurs. The histogram reveals varying levels of popularity, with some values being more frequent than others, suggesting a multimodal distribution.

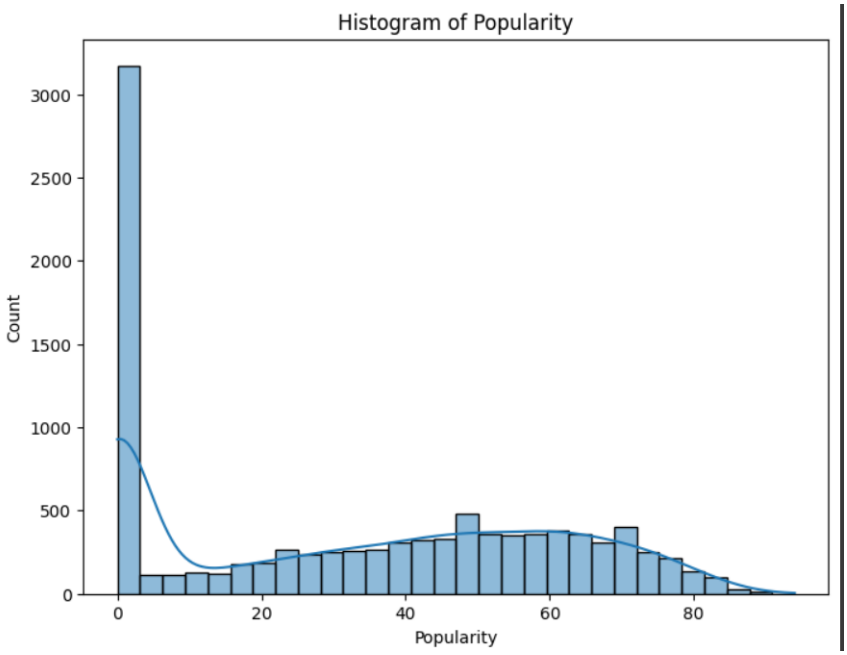

Figure 3.2: Histogram of Popularity

Homoscedasticity Check (Figure 3.3): The mean of residuals is -0.25046345939087045, which is close to zero, indicating that the model is unbiased. The plot visualizes the residuals (y-axis) versus predicted values (x-axis). The spread of residuals is fairly uniform, confirming the homoscedasticity assumption, which is ideal for regression analysis.

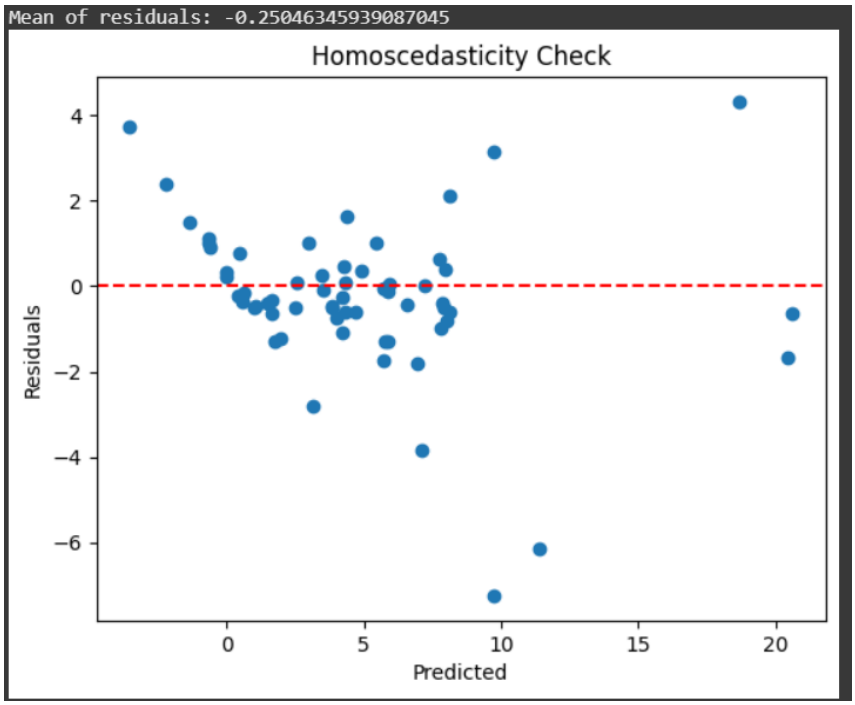

Figure 3.3: Homoscedasticity Check

Q-Q Plot (Figure 3.4): The Q-Q plot compares the sample quantiles to the theoretical quantiles of a normal distribution. The plot shows that the sample quantiles deviate from the theoretical quantiles, especially at the tails, which indicates the data is not normally distributed. This result is supported by the Shapiro-Wilk test with a p-value of 7.82e-06, rejecting the null hypothesis of normality.

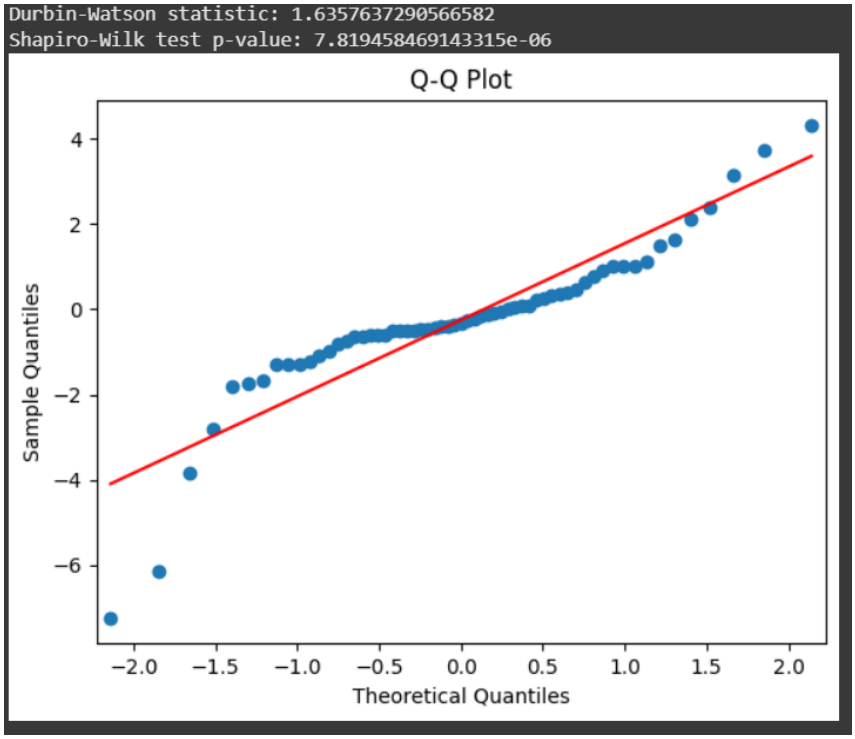

Figure 3.4: Q-Q Plot

Box Plot of Selling Price (Figure 3.5): The box plot shows the distribution of selling prices, ranging from 0 to 35. It highlights the central tendency and potential outliers. The spread of prices suggests that the dataset includes a wide variety of selling prices, with some extreme values that may need to be considered during model training.

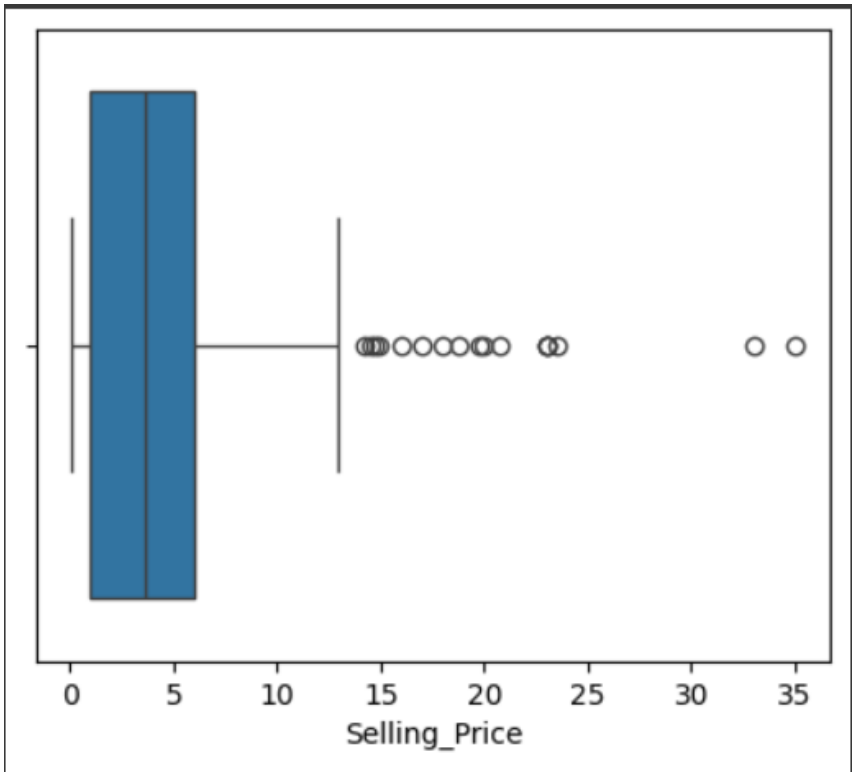

Figure 3.5: Box Plot of Selling Price

Distribution Plot of Selling Price (Figure 3.6): This plot shows the frequency of different selling prices. The distribution appears to be skewed, with a concentration of data points towards the lower end of the price range. This skewness indicates that most cars in the dataset are priced lower, which may affect model performance and should be taken into account during model evaluation.

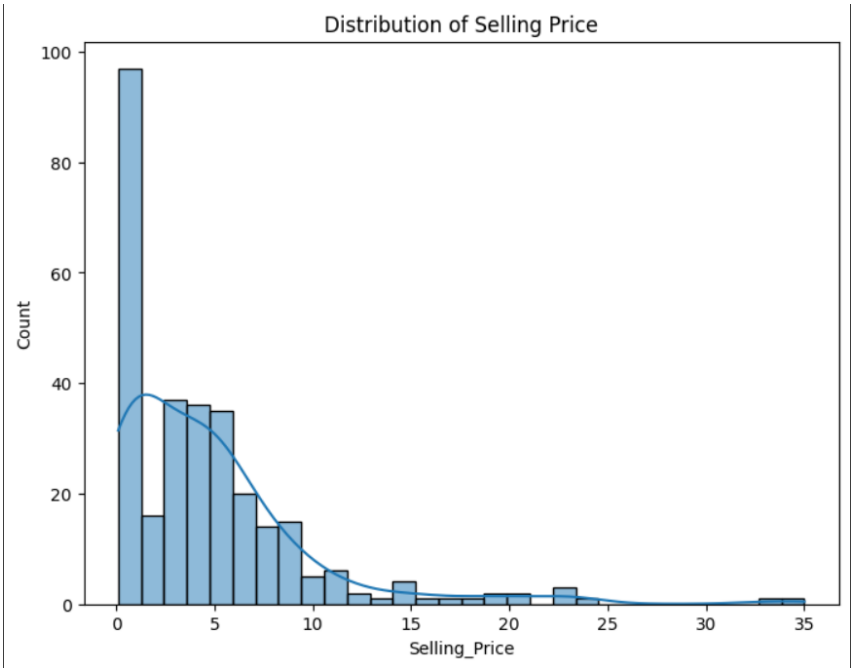

Figure 3.6: Distribution Plot of Selling Price

3.5. Model Comparison and Evaluation

Model Performance Comparison:

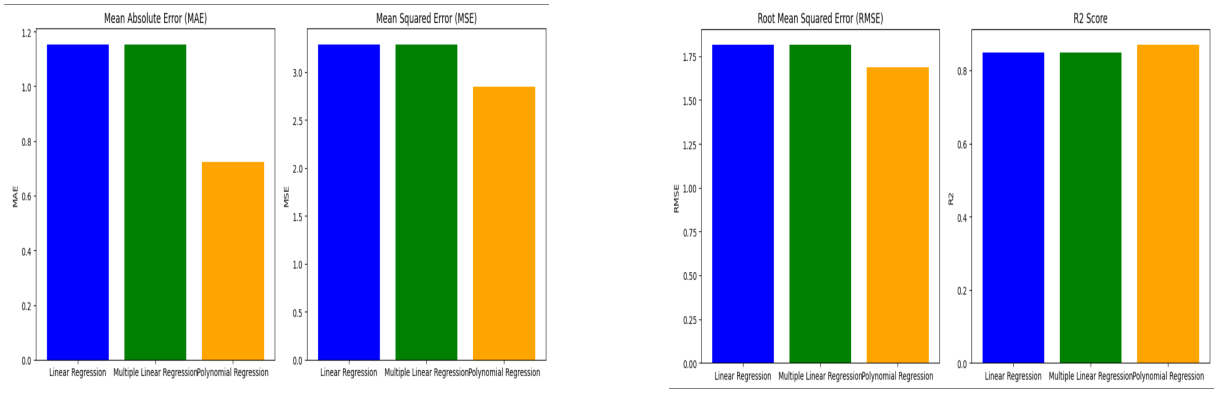

Figure 3.7: Linear, multiple and polynomial regression performance comparison

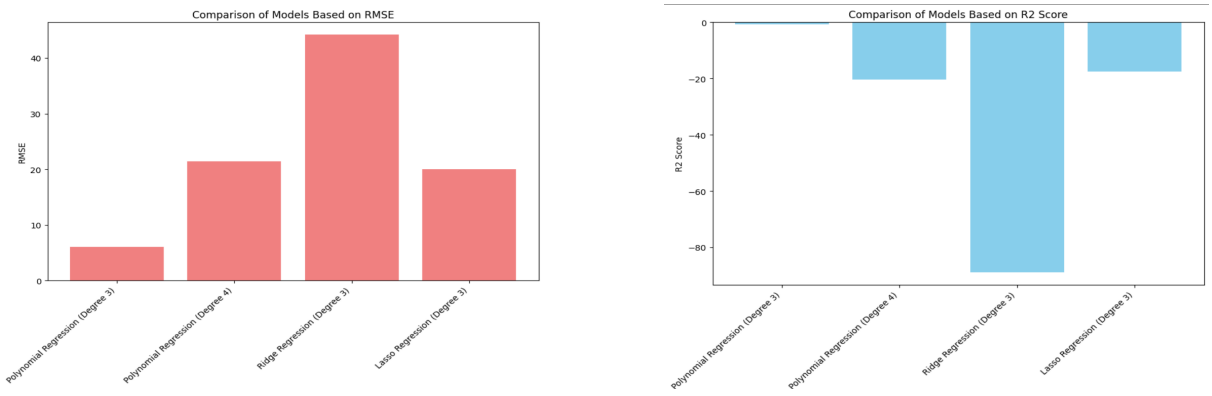

Figure 3.8: Polynomial levels comparison

- **Linear Regression & Multiple Linear Regression: MAE:** 1.153 (fairly high error on average) **MSE:** 3.288 **RMSE:** 1.813 (a bit high, indicating the presence of significant error) **R² Score:** 0.848 (good, indicating that the model explains about 84.8% of the variance)
- **Polynomial Regression (Degree 2): MAE:** 0.723 (lower, which means better performance in terms of average error) **MSE:** 2.843 (lower, which means the errors are smaller on average) **RMSE:** 1.686 (lower, indicating better fit) **R² Score:** 0.869 (better than linear regression, explaining 86.9% of the variance)
- **Polynomial Regression (Degree 3): MAE:** 1.87 **MSE:** 36.69 **RMSE:** 6.06 **R² Score:** -0.69 (Overfitting issue)
- **Polynomial Regression (Degree 4): MAE:** 4.76 **MSE:** 461.00 **RMSE:** 21.47 **R² Score:** -20.25 (Overfitting issue)
- **Ridge Regression (Degree 3): MAE:** 6.15 **MSE:** 1952.06 **RMSE:** 44.18 **R² Score:** -88.98 (Severe overfitting)
- **Lasso Regression (Degree 3): MAE:** 3.09 **MSE:** 401.64 **RMSE:** 20.04 **R² Score:** -17.51 (Overfitting issue)

Model Performance Interpretation: The **Polynomial Regression (Degree 2)** model provides a low Mean Absolute Error (MAE) of 0.72, meaning that the predicted values are very close to the actual values on average. The R² Score of 0.87 indicates that the model explains 87% of the variance in the target variable, which is excellent. This suggests that the model fits the data well and can make accurate predictions. The higher-degree Polynomial Regression (Degree 3) has significantly worse performance with a negative R² and higher errors (MAE, MSE, RMSE), indicating overfitting and poor generalization.

Conclusion: After evaluating the performance of different models, it is clear that **Polynomial Regression (Degree 2)** is the most effective model for this dataset. It balances predictive accuracy and complexity without overfitting.

Discussion

The results show that Polynomial Regression (Degree 2) provides the best balance between model complexity and predictive accuracy. It was able to fit the data without overfitting, unlike higher-degree polynomial models and regularized models such as Ridge and Lasso, which were too aggressive and underperformed. The positive R2 score of 0.87 suggests that Polynomial Regression (Degree 2) can explain a significant portion of the variance in the target variable, making it a suitable choice for car price prediction.

However, the fact that the residuals deviate slightly from normality indicates that there may still be room for improvement. Further exploration into feature engineering and transformations could potentially improve model performance.

4. Actionable Insights

4.1. Feature Engineering

It is essential to explore additional feature transformations or interactions between variables, such as encoding categorical variables in more sophisticated ways, or using logarithmic transformations to handle skewed data.

4.2. Model Tuning

Although Polynomial Regression (Degree 2) performed well, experimenting with more complex models like Gradient Boosting or XGBoost might yield even better results, especially with feature interactions.

4.3. Regularization

Ridge and Lasso performed poorly in this task, likely due to over-regularization. Fine-tuning their hyperparameters or considering different regularization techniques might improve their performance.

5. Conclusion

This study successfully predicted car selling prices using regression models. Polynomial Regression (Degree 2) emerged as the best model based on performance metrics, striking a good balance between bias and variance. Although further refinement of the model is possible, the current approach provides a solid foundation for accurate car price predictions. Future work could involve feature engineering, experimenting with more advanced models, and further validation using real-world data.

6. References

Python documentation: https://python.org Scikit-learn documentation: https://scikit-learn.org
Statsmodels documentation: https://statsmodels.org Matplotlib documentation: https://matplotlib.org
Seaborn documentation: https://seaborn.pydata.org Kaggle dataset: https://www.kaggle.com/code/nesanabipou
price-prediction-regression/input