Car Price Prediction Using Linear and Lasso Regression: A Comparative Analysis of Model Performance

Abstract

Predicting car prices is a crucial task for both buyers and sellers in the automotive market. Accurate price predictions can help inform decision-making and enhance market efficiency. This study presents a comparative analysis of two regression techniques—Linear Regression and Lasso Regression—applied to car price prediction. The models were evaluated based on their R-squared error, with Linear Regression achieving 0.8799 on the training set and 0.8365 on the test set, while Lasso Regression obtained 0.8427 on the training set and 0.8709 on the test set. The results indicate that Lasso Regression outperforms Linear Regression on the test data, suggesting that Lasso's feature selection capability contributes to better generalization.

Introduction

Car price prediction is a challenging task that requires accurate modeling of the various factors influencing the price. Factors such as brand, model, year of manufacture, engine size, mileage, and additional features all contribute to the final market value of a car. Traditional methods of price estimation often rely on expert judgment, which can be subjective and inconsistent. With the rise of machine learning, more objective and data-driven approaches have become viable. This study focuses on two widely used regression techniques—Linear Regression and Lasso Regression—and compares their performance in predicting car prices. The objective is to determine which method provides more accurate and generalizable predictions.

Related Works

Numerous studies have applied machine learning algorithms to predict car prices, using various features and techniques. Linear Regression has been a popular choice due to its simplicity and interpretability. However, it often suffers from overfitting, especially when dealing with high-dimensional data. To address this, researchers have explored regularization techniques such as Ridge and Lasso Regression. Lasso, in particular, has gained attention for its ability to perform feature selection by shrinking less important coefficients to zero. Previous studies have shown that Lasso can improve prediction accuracy by reducing model complexity, making it an ideal candidate for comparison against standard Linear Regression.

Algorithm

1. Linear Regression

Linear Regression is a fundamental statistical method for modeling the relationship between a dependent variable and one or more independent variables. The model assumes a linear relationship between the input features and the output, represented by the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Where:

y is the dependent variable (car price).

β_0 is the intercept.

RHS terms are the coefficients for the independent variables x1,x2,…,xnx_1, x_2, \dots, x_nx1,x2,…,xn.

ϵ is the error term.

The goal of Linear Regression is to find the coefficient values that minimize the sum of squared errors (SSE) between the predicted and actual values.

2. Lasso Regression

Lasso Regression is a regularized version of Linear Regression that adds a penalty term to the cost function. This penalty term is the sum of the absolute values of the coefficients, controlled by a hyperparameter λ:

$$\text{Cost Function} = \text{SSE} + \lambda \sum_{j=1}^{n} |\beta_j|$$

The Lasso algorithm performs both parameter shrinkage and variable selection, making it effective in scenarios with many features. As λ increases, Lasso forces some coefficients to become exactly zero, effectively removing irrelevant features from the model.

Methodology

The methodology for car price prediction using Linear and Lasso Regression involved the following steps:

Data Collection: A dataset of car prices and associated features was obtained. The features included make, model, year, engine size, mileage, and additional specifications.

Data Preprocessing: The data was cleaned to remove missing values and outliers. Categorical variables were encoded using one-hot encoding, and numerical variables were scaled to ensure uniformity. The dataset was then split into training (80%) and testing (20%) sets.

Model Training: Both Linear Regression and Lasso Regression models were trained on the training set. For Lasso Regression, the λ parameter was tuned using cross-validation to identify the optimal value.

Model Evaluation: The performance of both models was evaluated using the R-squared error on the training and testing sets. This metric indicates how well the model explains the variance in the data, with higher values indicating better performance.

Experimental Work

The experimental work involved implementing both Linear and Lasso Regression models on the prepared dataset. The models were trained on the training set, and their performance was evaluated on both the training and testing sets. The R-squared error was used as the primary evaluation metric. Additionally, the Lasso Regression model's ability to perform feature selection was analyzed by examining the coefficients of the final model.

Results

The results of the experiments are as follows:

Linear Regression:

- R-squared error on training data: 0.8799
- R-squared error on testing data: 0.8365

Lasso Regression:

- R-squared error on training data: 0.8427
- R-squared error on testing data: 0.8709

The results show that while Linear Regression performed slightly better on the training data, Lasso Regression outperformed it on the testing data. This suggests that Lasso's regularization technique helped the model generalize better to unseen data by reducing overfitting.

Conclusion

This study compared the performance of Linear Regression and Lasso Regression in predicting car prices. The findings indicate that Lasso Regression offers better generalization capabilities due to its regularization and feature selection properties, as evidenced by its superior performance on the test set. These results highlight the importance of using regularization techniques in regression models, particularly when dealing with high-dimensional datasets. Future work could explore the impact of different regularization parameters and the inclusion of additional features, such as market trends, to further improve prediction accuracy.

References

- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis. Wiley.
- Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.