Final Project

Car Price Prediction and Value Analysis Using Machine Learning

Juan Morales-Vargas
jdm161@students.uwf.edu

Data Mining, CAP 4770

[Fall, 2024]
The University of West Florida

December 2, 2024

Dr. Shashi Bhushan Jha

Table of Content

Car Price Prediction and Value Analysis Using Machine Learning

I. Introduction (problem statement and background)

The automotive industry is always changing and the buyers need to be informed to be able to make the right decisions. When the buyer wants to buy a new car, it is very important they have an informed idea about the car prices and how sellers and manufacturers work. This project is going over the prediction of car prices based on various features using the Kaggle data set. The task is to build a predictive model that can give reliable price predictions and try to explain which cars have the most value for price. This project will show an analysis of data while displaying the relationship between the different features and the car prices. The results obtained are supposed to help the buyers make informed choices on the purchase of cars and help in understanding the market trend.

The Linear Regression model was chosen because it is simple, interpretable, and gives a clear baseline by which all other models must be compared. Ridge Regression is used because the coefficient estimates would get stabilized due to multicollinearity among features. The Decision Tree model is used to capture the non-linear relation and interactions among the features. The objective of the project is to develop an efficient machine learning model for car price prediction, considering the feature variables presented in the data set. Linear Regression, Ridge Regression, and Decision Tree, will be applied in this project to model the relationship between different features of cars and their prices

In order to evaluate the performance of these models, various algorithms in machine learning have been used, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R-square). For this purpose, MAE gives a simple measure of the accuracy of predictions, showing the mean magnitude of errors, while RMSE emphasizes larger errors by squaring the differences and sensitivity to outliers. R-squared will be used in establishing how the model fits the variance in car prices.

This analysis will provide insights on what directs the car prices and what will make the model interpretable for stakeholders or the person making the decisions. The project will analyse a car data set of 11,915 entries, this data set includes new and old cars up to the year of 2017 and also sixteen other features like year, make, model, price, miles per gallon, horsepower, and transmission.

II. Problem Description (Data collection, preprocessing, data exploring)

This data set was collected from Kaggle, and this data will have all features, which included the car brand, model, price, horsepower (HP), and miles per gallon (MPG). The data was collected from the kaggle dataset using an sql database that I set up. Data reliability and quality can be provided by multiple preprocessing steps. The missing values will be taken care of through removal of the incomplete data records using dropna(). Duplicate rows were also removed. This will help keep the analysis without any biases or flaws. Only the key features relevant to car price prediction were retained, such as HP, MPG, price, make, model, and more while the unnecessary columns were excluded. The values were also renamed for readability and ease of editing. For the purpose of modeling, the features went through a standardization using StandardScaler from the sklearn library to ensure that all variables were on the same scale. This preprocessing approach was crucial for the prediction because it removed potential biases that

can arise from different measures of the features that will help to lay a strong foundation for the predictive modeling phase.

A variety of visualizations were made so we can have a better understanding of the data set and find meaningful patterns in it. In figure one (Histogram of Car Prices), we see a histogram of the distribution of the car prices which showed that most cars are under $50,000. This distribution indicates the range of affordability, from budget friendly vehicles to high-end luxury cars. This histogram showed a number of outliers in the higher range of prices, reflecting more expensive models that may need special attention in the analysis. This also shows that the market for affordable to mid priced cars is bigger than any other sector.

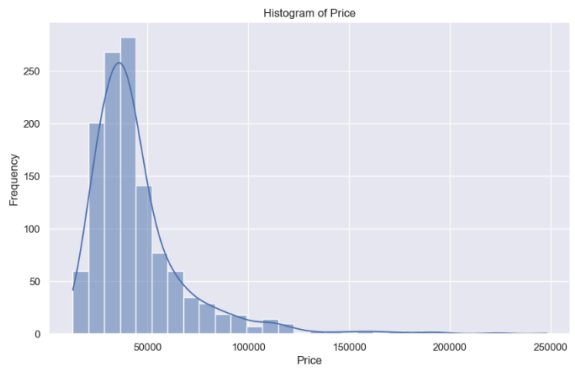

Fig. 1: Histogram of Car Prices

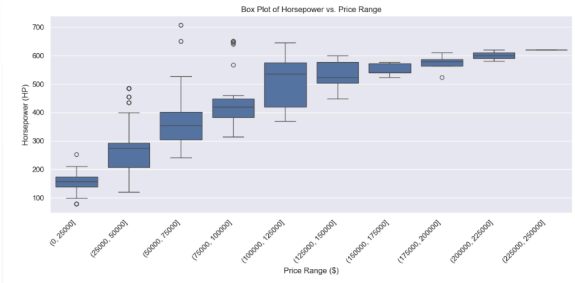

Fig. 2: Box Plot of Horsepower vs. Price

In figure two (Box Plot of Horsepower vs. Price) a box plot of horsepower against price gave an insight into the relationship between the power of an engine and the valuation of a car. The plot showed that cars with higher horsepower generally are more expensive. Outliers in this plot often represent the sports cars, which could have a better performance while maintaining a lower price. In figure three (Scatter Plot of MPG vs. Price), the scatter plot was MPG versus

price, with higher MPG cars generally having more affordable prices. That reflects that fuel efficiency is normally emphasized in economy cars while in high-performance (or high end) vehicles, a lot of focus is set on power rather than efficiency. In figure four we show a scatter plot for horsepower vs price. This graph shows how big the variance is in price once you reach the higher horsepower levels. This shows that horsepower is not directly correlated with the price of the car which makes the price of the car harder to predict.

Fig. 3: Scatter Plot of MPG vs. Price

Fig. 4: Scatter Plot of HP vs Price

To understand the correlations between different features, the correlation matrix was plotted as a heat map (figure five). It represents a strong dependency of horsepower and the number of cylinders with the price, the engine performance is a key influence on the value of an automobile. The MPG shared a negative correlation with horsepower and price and solidified the fact that the fuel efficiency and power generally come as trade offs.

A scatter plot of horsepower vs. MPG (figure six) underlined the trade off between power and fuel economy. The cars with high values of horsepower had a low MPG, reinforcing that high performance comes at the price of fuel consumption. These figures help to provide a broad

view of the data set and help to gain an understanding of what variables are the most important that affect the car prices. This analysis provided the basis for developing the more accurate predictive models.

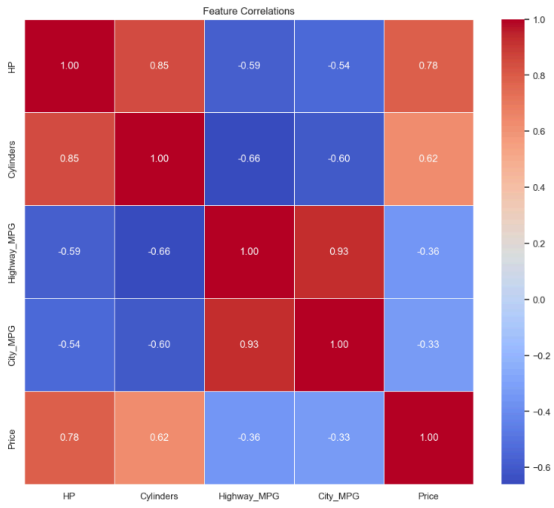

Fig. 5: Heatmap of Correlation Matrix

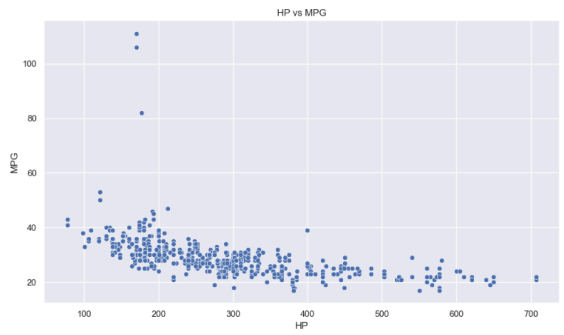

Fig. 6: Scatter Plot of Horsepower vs. MPG

III. Methodology (Data modeling)

The several modeling techniques were each designed to understand feature relationships between horsepower, MPG, and the price of the car. We then gathered the data on how they performed. This is necessary because comparing the effectiveness of the algorithms can be used to find the best prediction overall.

Linear Regression

Linear Regression was used because it is easier to interpret the data and provide a clear understanding of the relations between the variables. A linear relationship is an excellent baseline model because it facilitates a deep understanding between features and the target variable of price. It provided insight into how each feature, especially horsepower and MPG, affects car prices. The evaluation metrics developed for this model, Mean Squared Error (MSE) and R-squared (R-square), showed that the overall trend was captured by the model, and the

areas where it was not performing that well on more complex patterns. Especially in the extremes, where high-performance or luxurious cars showed up.

Ridge Regression

Ridge regression was used to analyze the predictive performance with the regression data. This method is suitable for a higher number of variables and emphasizes other important predictors in the data set. This model performed better than standard Linear Regression especially when there was suspected multicollinearity among some features, such as between horsepower and the number of cylinders. These metrics evidenced better generalization, lower MSE and higher R-square score, than the baseline linear regression model.

Decision Tree Regression

A nonlinear relationship in the data was modeled using the Decision Tree Regression. Unlike linear models, a Decision Tree can split the data into segments based on different feature values. This will make it easier to capture more difficult interactions between the features. The Decision Tree model performed better than linear and ridge regression models with a much higher R-square score and lower MSE. This model became more accurate in predicting the car prices across various segments of the best value cars (and under different ranges of prices).

Model Evaluation and Validation

All models were evaluated using cross-validation to ensure that they were reliable and capable of performing well with the new data to avoid biases in evaluation. The data set was split multiple times into several training and testing subsets. Among these different models developed, the Decision Tree Regression performs best with an R-squared value of 0.77, explaining 77% of the variance in car prices. The ridge regression and linear regression models resulted in lower

R-square values at around 0.61, illustrating how the two models were not very powerful in modeling the complexity of the data.

Modeling Insights

The modeling results showed that horsepower and the number of cylinders were strong predictors of car price, while MPG affects the pricing of a car. The decision tree regression had the capability to capture more complication patterns, and it also had the best predictions in identifying cars that offered the best value. These findings emphasize the need to choose appropriate modeling techniques based on the nature of the data set and the problem at hand. When we combined all the modeling approaches (and the possibility to analyze car price prediction from all sides was present) the most effective technique for this type of problem was the Decision Tree Regression. The knowledge that comes from such models has not only brought a more improved accuracy in such predictions but also helped the car buyer to select those cars which would offer the maximum value for their investment.

IV. Conclusion (results and interpretation)

Results and Interpretation

The performance of the car price prediction models was represented using some key evaluation metrics, namely Mean Squared Error, Root Mean Squared Error, and R-squared. These metrics give a full overview of how each model represents the relationship between features of cars and their prices and their predictive accuracy.

Linear Regression Results

The Linear Regression baseline gave an R-square score of about 0.44, meaning it explains 44% of the car price variance. It turned out that this model produced reasonable predictions for mid-segment cars but ran into problems with either luxury or high-performance cars. As expected, this resulted in an MSE of over 381 million, showing the limitation of this model to capture non-linear relationships inherent in the data.

Ending

The Linear Regression results show that simple features like horsepower and MPG have a significant linear effect on the price of cars, but the model was too simple to catch the more complex patterns, especially at the extremes. The result agrees with the expectations that sophisticated models are needed for better predictions in such a diverse data set.

Ridge Regression Results

It turned out that Ridge Regression with cross validation did improve the over fitting and slightly gave better prediction accuracy, recording a better MSE with a bit higher R-square score of 0.68 when compared to Linear Regression, yet it still provided poor overall performance in estimating prices for cars with extreme horsepower or extremely high values of MPG. Ridge Regression is generally better when it comes to performance, regularizing less influential features while preventing overfitting. On the contrary, this is very restricted when considering the capturing of non-linear patterns, and more complicated interaction models are necessary.

Results Using Decision Tree Regression

The Decision Tree Regression model with cross validation did the best, having an R-square score of 0.86 and a mean squared error of about 109 million. This model did a very good job in predicting car prices across various categories, most especially in identifying cars

offering the best value under different price thresholds. It segmented the data well to capture the relationships between horsepower, MPG, and price, leading to superior predictive performance. The Decision Tree results bring out the strength of handling non-linear data. It gives an insight into how a couple of high-horsepower cars give very fine value for their price-the Ford Mustang and the Dodge Challenger-being the best-valued cars below $45,000. Besides this, its efficiency in predicting lower-priced fuel-efficient cars like the Honda Civic reinforces its robustness across a wide range of different segments.

General Insights

Best Value Cars: The Ford Mustang-V8, Dodge Challenger, and Chevrolet Camaro provided the best value for the cars because of the high horsepower relative to price and mpg.

Feature Impact: Horsepower was the strongest predictor of price, and higher MPG meant lower prices, reflecting a trade-off between performance and efficiency.

Comparing models: Decision Tree Regression became the best model for this problem, since the capturing of non-linear patterns and interaction became necessary to improve prediction.

General insights from the results gave a great understanding of what affects car prices and which one offered better value. These findings provide very practical guidance to buyers regarding fuel efficiency, balancing performance, and cost, allowing informed decisions on car purchases.

Works cited

Cooper Union. (n.d.). *Car dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/CooperUnion/cardataset/data

Scikit-learn developers. (n.d.). *Scikit-learn: Machine learning in Python*. https://scikit-learn.org/stable/

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to data mining* (2nd ed.). Pearson.