Springboard – DSC

Capstone Project 2

Car Price Prediction

By Steve Li

Jan, 2024

1. Introduction

The object is to identify and understand the significant variables that influence car prices in the American market to aid an auto company in strategically entering and competing in the US automobile industry.

The stakeholders include:
The automobile company, who is interested in understanding the American car market dynamics to make informed decisions about the pricing strategies.
The automobile consulting firm, hired by the automobile company to conduct the market analysis.

Among all factors, higher fuel efficiency negatively impacts car price predictions. Car power has a positive impact on car price predictions, especially for larger cars. Larger car size tends to positively affect the car price predictions, and this effect is somewhat more pronounced in cars with lower fuel efficiency.

Link to Github

2. Approach

2.1 Data Acquisition and Wrangling

After loading the data file into a data frame, shape, data type and semantics are shown to provide basic information of the data. A quality check was performed including the duplicated rows, missing values. In addition, a new 'company' column was created from the 'CarName' column. The correction on the company names was made to these new columns. Finally, categorical and numerical columns are selected.

2.2 Storytelling and Inferential Statistics

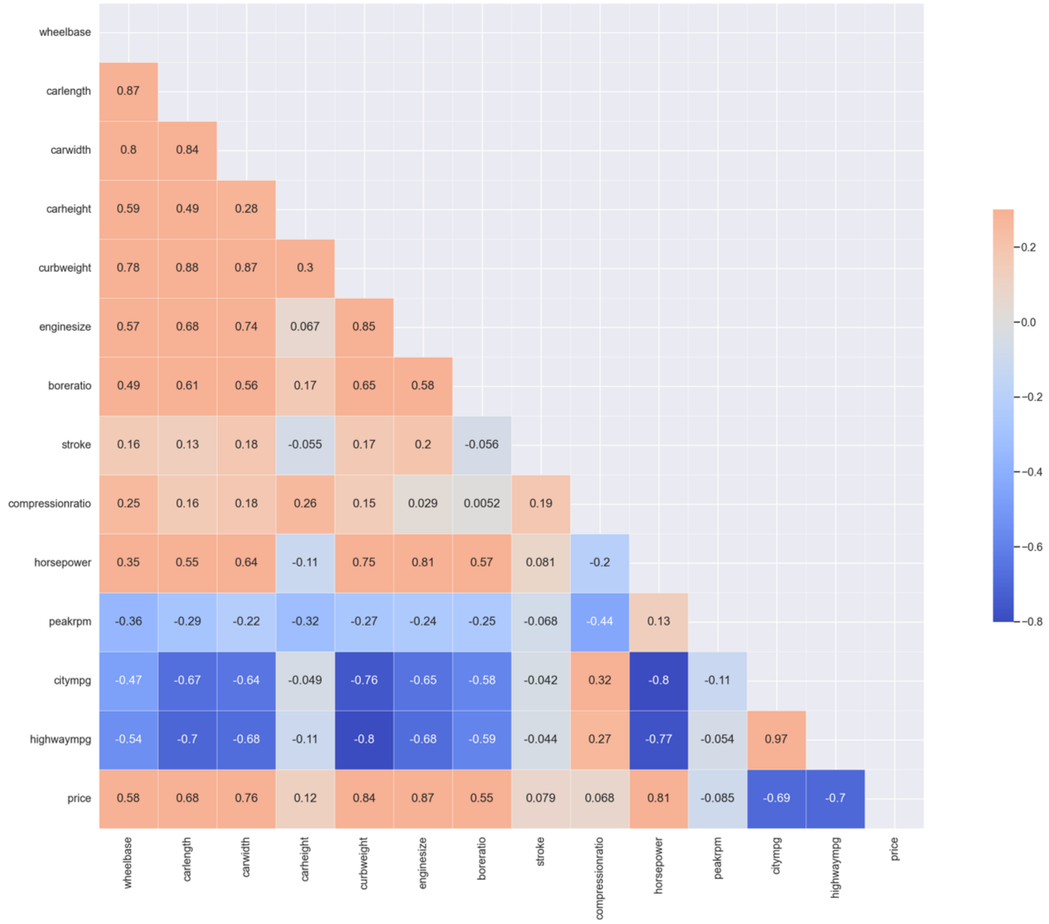
The initial finding from the heatmap indicates that features including carwidth, curbweight, enginesize, horsepower have clear positive impact on the target price . The correlation between features that is over 0.8 includes: wheelbase and carlength, wheelbase and carwidth, carlength and curbweight, carlength and carwidth, curbweight and enginesize, curbweight and highwaympg, enginesize and hoursepower, hoursepower and citympg, citympg and highwaympg. In order to reduce the multilinearity without losing much information, these features are combined in the following ways.

Creating a new feature carsize that is the average of wheelbase, carlength, carwidth. These three features individually represent different dimensions of a car's size. They are likely highly correlated as larger cars will generally have larger values for all three. Carsize captures the overall size of the car more effectively.

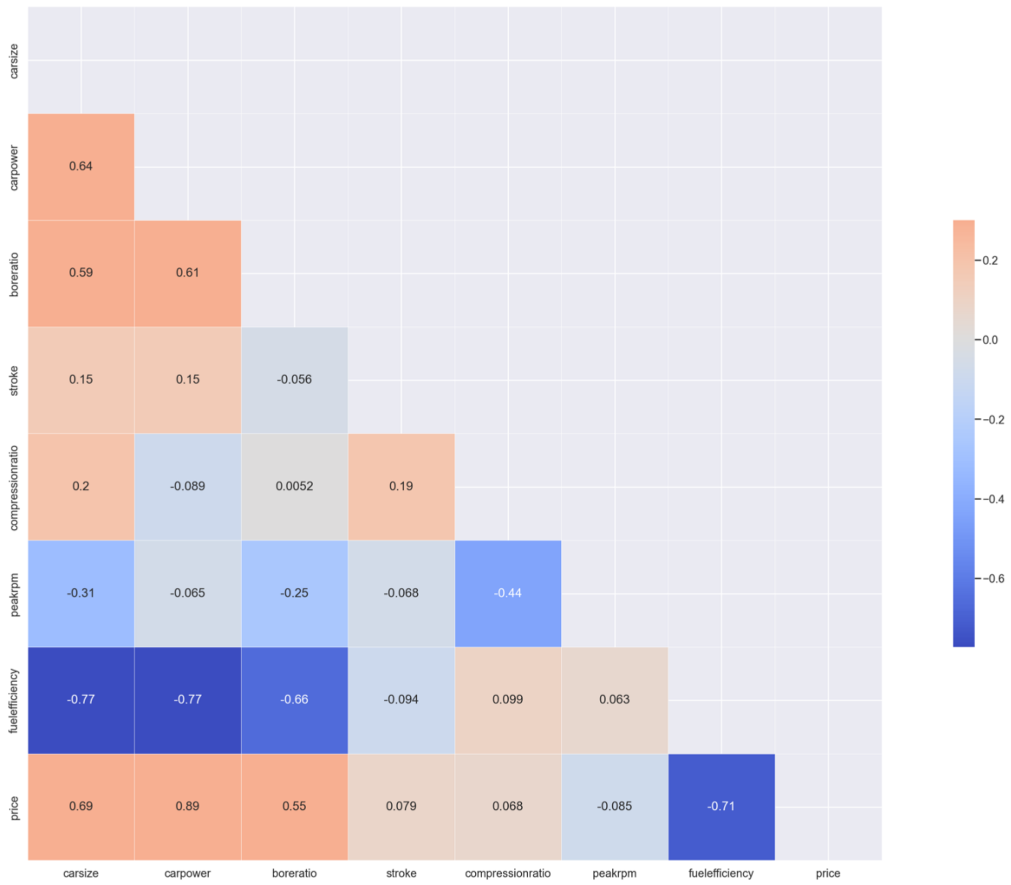
Similarly, creating a new feature fuelefficiency that is the addition of citympg and highmpg devided by the curbweight. Both citympg and highwaympg are measures of fuel efficiency. Curbweight is also related to fuel efficiency as heavier cars typically consume more fuel. The feature fuelefficiency creates a more comprehensive metric of fuel efficiency that accounts for both types of mileage and the car's weight. This singular metric might be more predictive and less noisy than considering each individually.

Lastly, creating a new feature carpower that is the average of enginesize and horsepower. Engine size and horsepower are measures of a car's power and performance. The feature carpower creates a more robust measure of the car's overall power. This could help in reducing the features while still capturing the essence of the car's performance capabilities.

The overall rationale behind combining these features is to reduce the number of variables (thus simplifying the model), reduce multicollinearity (which can skew results and make models unstable), and to create new features that might have a more straightforward or stronger relationship with the target variable (in this case, the price of the car). These engineered features are expected to retain most of the predictive power of the original set of features but in a more condensed form. This can lead to models that are easier to understand and work with, and potentially even improve performance by reducing overfitting and noise in the data.



Heatmap of features and target ‘price’ before feature engineering



Heatmap after feature engineering

From the regenerated heatmap, four features including price, carsize, fuelefficiency, carpower, and boreratio were selected for applying k-mean clustering algorithm. The purpose is to group car brands into clusters.

After the clustering algorithm, the 3D plot separate car brands into 2 clusters with one cluster representing the affordable and accessible vehicles to a broad range of consumers and the other cluster representing luxury and high-end models.

Specifically, they are

Cluster 0:

Alfa-Romeo, Chevrolet, Honda, Isuzu, Renault, Dodge, Mazda, Mitsubishi, Nissan, Plymouth, Subaru, Toyota, Volkswagen

Common Traits:

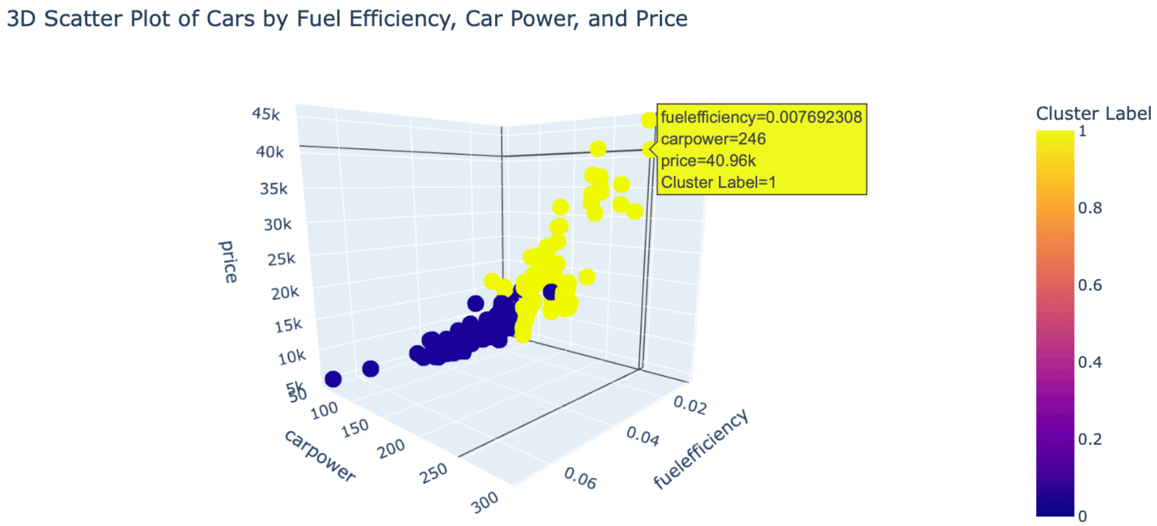
Affordability and Accessibility: These brands are typically known for producing vehicles that are affordable and accessible to a broad range of consumers.

Wide Range of Models: Offering a variety of models, including economy, family, and utility vehicles.

Focus on Practicality: Emphasis on practicality, reliability, and efficiency, catering to everyday use and needs.

Global Market Presence: These brands have a strong presence in various global markets, known for their mass-market appeal.

Cluster 1:
Audi, BMW, Buick, Jaguar, Mercury, Peugeot, Porsche, Saab, Volvo
Common Traits:
Luxury and High-End Models: This cluster includes brands known for luxury vehicles, often with higher price points and premium features.
Performance and Design: Brands like BMW, Porsche, and Jaguar are recognized for their focus on performance, advanced engineering, and distinctive design.
Innovative and Technologically Advanced: These brands often lead in incorporating advanced technology and innovative features in their vehicles.
Strong Brand Identity and Prestige: Brands in this cluster typically have a strong brand identity, often associated with prestige, status, and luxury.



3D Scatter Plot of carprice vs carpower and fuelefficiency

2.3 Baseline Modeling

Linear regression algorithm was used to build the first model. Columns including car_ID, symbolling, cluster are dropped because they do not contain useful information. Numerical and categorical columns are identified. Then one-hot encoding was applied on the categorical columns. Standard scaling was applied on the numerical columns. And the resulted transformed columns are combined as the dataset for modeling.
The result from linear regression with cross-validation came out to have very high RMSE on the test data compared to the train data. This indicates the potential overfit issue. Two off the scale points in the y_pred dataset can cause the enormous size of RMSE. Therefore, both Lasso and Ridge regression were applied. And they output reasonable results with Lasso RMSE being around 2755 and Ridge RMSE around 2613, which shows great improvement over the Linear Regression model.

2.4 Extended Modeling

Lasso and Ridge regression algorithms produce linear models. The dataset exhibits certain nonlinearity from observing the pairplot. Therefore, additional modeling methods such as Random Forests, XGboost, and LightGBM were explored in this section. Several techniques such as pipeline, gridsearch, were implemented. And RMSE, adjusted R-squared, MAPE were calculated across all three models. Residuals of Lasso, Ridge, Random Forests, XGBoost, LightGBM models were plot as the overlayed histogram. Finally, SHAP evaluation were applied on the best model-Random Forests to identify the features with the highest impact on car price. And the dependence of these features was plot as well.

2.5 Findings

	RMSE	Adjusted R-squared	MAPE
Lasso	2756	1.38	15.1
Ridge	2613	1.34	14.9
Random Forests	2087	1.22	10.8
XGboost	2110	1.22	11.2
LightGBM	3215	1.52	15.4

From the table above, the Random Forests model is the best one evaluated by all three metrics.

From the SHAP value plots, fuelefficiency, carpower, and carsize are the top three features that impact car price. In the lower range fuelefficiency, higher fuelefficiency increases car price. In the higher range fuelefficiency, higher fuelefficiency decreases car price. Fuelefficiency has a nonlinear impact on car price. In the higher range carpower, it has very strong impact on increasing car price. Feature carsize has moderate impact, with higher range carsize increases car price while lower range carsize decreases car price.

Therefore, if the goal of the automobile manufacturer is to produce high price car, the car needs to have high car power such as horse power and engine size. And there is no need to be concerned about the fuel efficiency of such car.

2.6 Conclusions and Future Work

The object is to identify and understand the significant variables that influence car prices in the American market. Six linear and nonlinear models including the baseline model were built to predict car price from engineered features. Among them, Random Forests model has the best performance by the evaluation metrics. SHAP analysis was applied on the Random Forests model trained on full data set to illustrate the impact of features on model output and dependency between the top features.

Future work can include experimenting with more models such as GBM, Neural Network, or applying ensemble methods; expand additional hyperparameter values and utilize more advanced optimization techniques; Collect more data; create an API for the model so that it can be integrated with websites or management systems, allowing for dynamic pricing recommendations.

2.7 Recommendations for the Clients

- Fuel efficiency and car power are the most impactful factors on car price.
- If the goal is to sell expensive car, manufacturer needs to product car with very high power, low efficiency, and big size.
- If the goal is to sell cost-effective car, manufacturer needs to product car with high fuel efficiency, low car power, and small car size.

2.8 Consulted Resources

- Data is sourced from https://www.kaggle.com/datasets/hellbuoy/car-price-prediction?select=CarPrice_Assignment.csv
- Libraries include Numpy, Pandas, Matplotlib, Seaborn, Scikit-learn, XGBoost, LightGBM, Plotly, Jupyter Notebook
- Springboard Data Science
- ChatGPT 4.0