Predict Car Price

Data Science Capstone Project Jan. 2023 Cohort

Si Steve Li

Thanks to Springboard mentor AJ Sanchez

The Problem

- In 2023, light vehicle sales in USA is around 15 millions.
- Geely Auto want to understand the features on which the pricing of cars depends so that it can sell competitive cars to gain market share in the American market.
- What features affect car price?
- What features influence car price the most?
-

Data Information

- 26 columns and 206 rows of data
- Data columns include car_ID, symboling, CarName, fuel type, aspiration, door number, carbody, drive wheel, engine location, wheelbase, car length, car width, car height, curb weight, engine type, cylinder number, engine size, fuel system, bore ratio, stroke, compression ratio, horsepower, peak rpm, city mpg, highway mpg, price

Pairplot

- Positive correlation

  carlength, carwidth, curbweigh, wheelbase

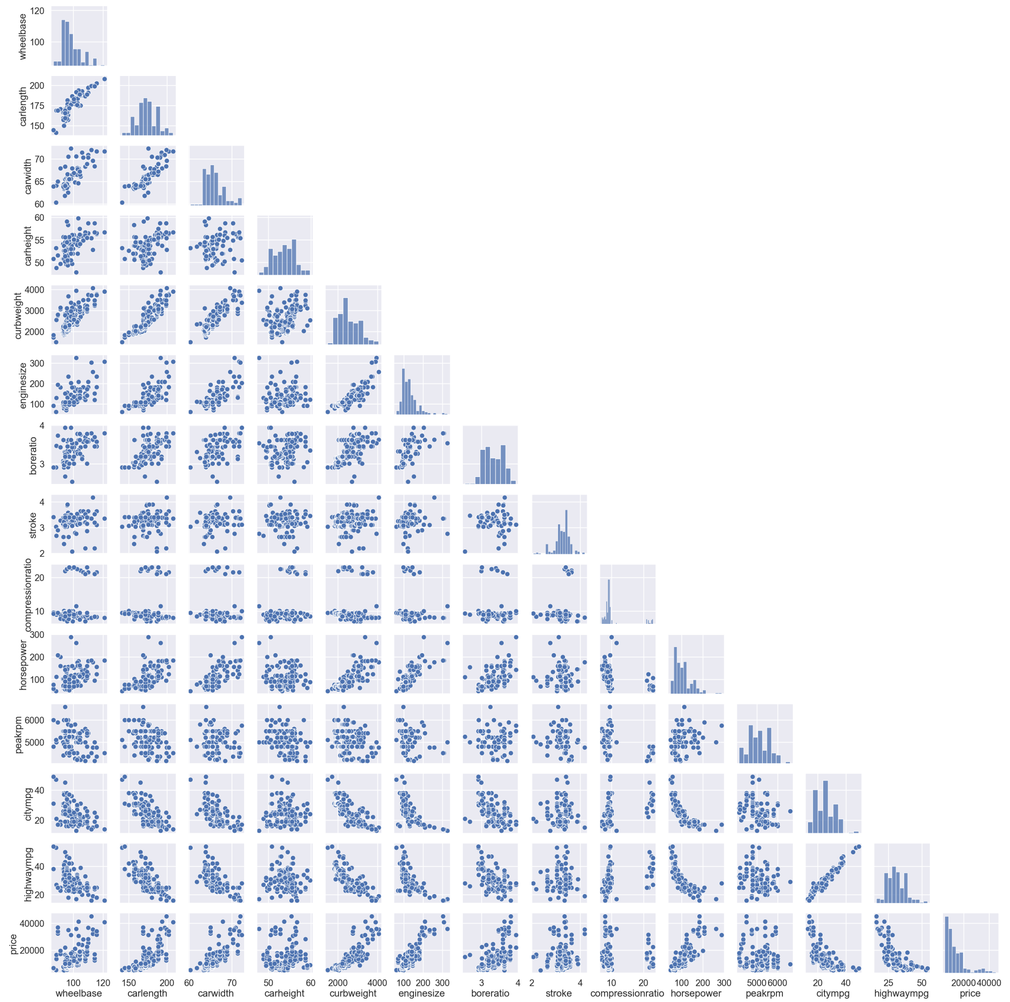
  highwaympg, citympg

  price, curbweight, engine size, horsepower
- Negative correlation

  carlength - highwaympg, citympg

  horsepower - highwaympg, citympg

  price - highwaympg, citympg



Heatmap

- Heatmap - features with over 0.8 correlation include

  Wheelbase, carlength

  Wheelbase, carwidth,

  Carlength, curbweight

  Carlength, carwidth

  Curbweight, enginesize

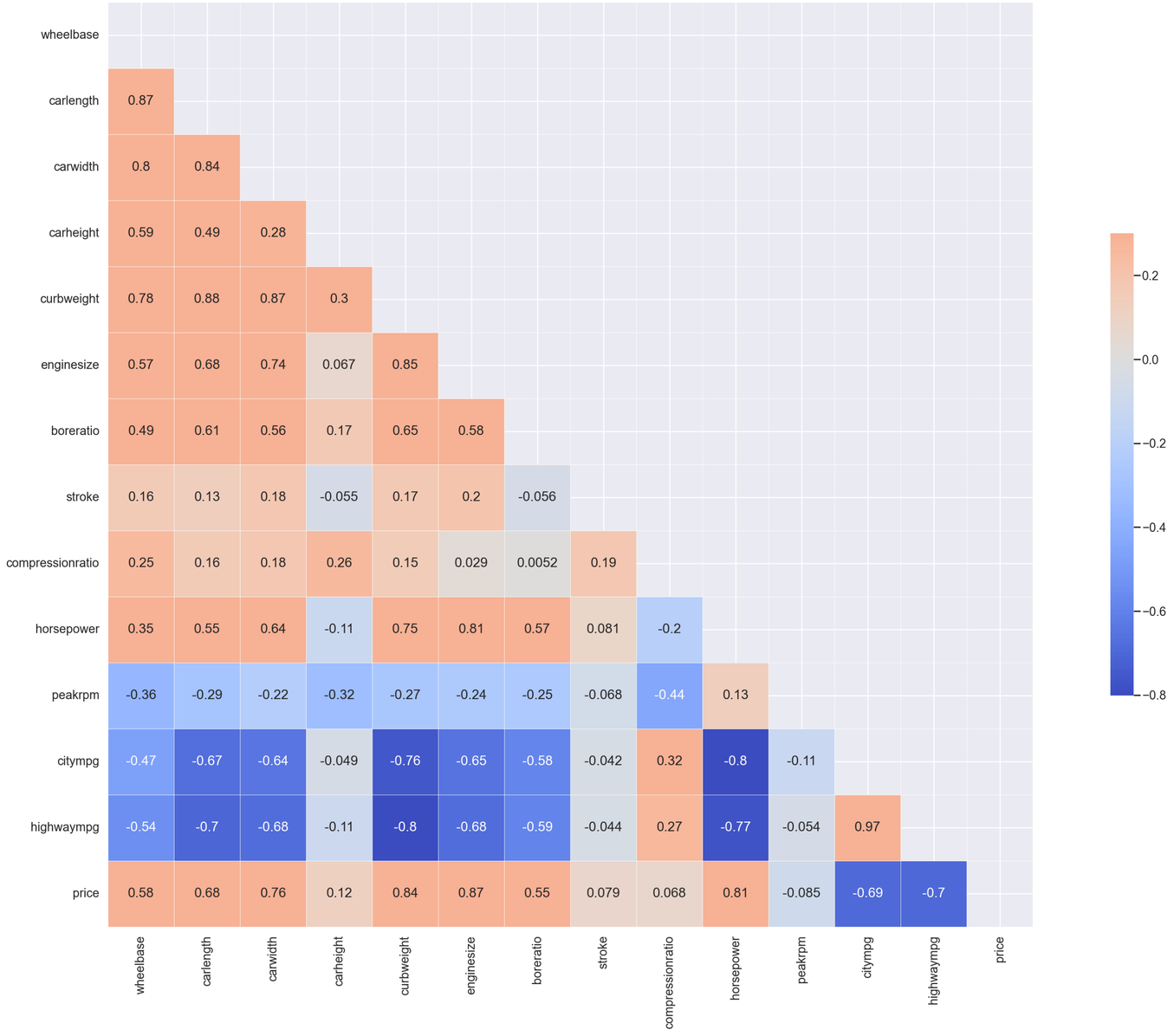
  Curb weight, highwaympg

  Enginesize, horsepower

  Horsepower, citympg

  Citympg, highwaympg

- Reconstruct new features
  - carsize=(wheelbase+carlength+carwidth)/3
  - fuelefficiency=(citympg+highwaympg)/curbweight
  - carpower=(enginesize+horsepower)/2



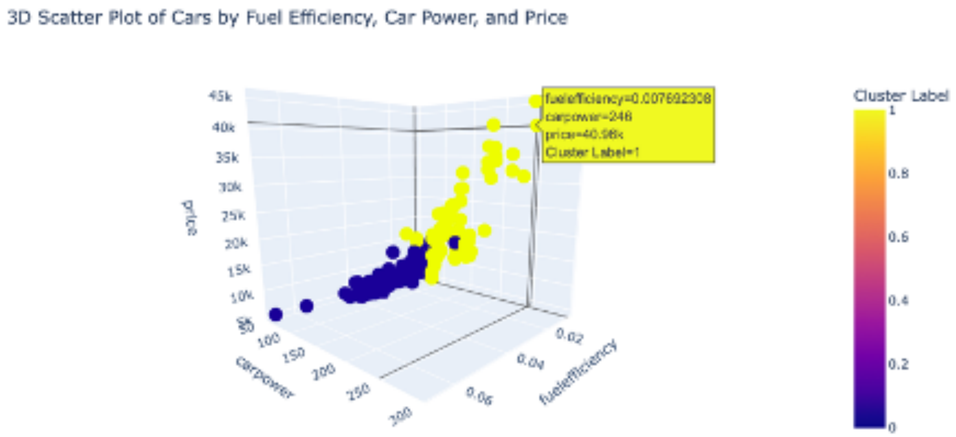
Brand Segmentation

Cluster 0 vs. Cluster 1

22 car brands were grouped into two cluster via k-means

Selected features include price, carsize, fuel efficiency, carpower, boreratio from regenerated heatmap

- Cluster 0
  - Affordability and accessibility
  - Wide range of models
  - Focus on practicality
  - Global market presence
  - Toyota, Honda, etc.



- Cluster 1
  - Luxury and High-End
  - Performance and design
  - Innovative and advanced
  - Brand identity and prestige
  - BMW, Porsche, etc.

Baseline Modeling

Linear Regression

- Train/test data split (test_size=0.25)
- Apply one-hot encoding on categorical columns
- Apply standard scaling on numerical columns
- Combine encoded and scaled columns
- Perform 5-fold cross-validation on training data
- Average RMSE test - 1.2e15
- Average RMSE train - 1236.7

Large difference between RMSE test and train

- Overfitting
- Outliers

Extended Modeling

- Lasso and Ridge Models

  Alphas space - np.logspace (-4, -4, 20)

  5 fold cross validation, optimal_alpha
- Random Forests, XGBoost, LightGBM Models

  Define parameter in rf_param_grid, xgb_param_grid, lgb_param_grid

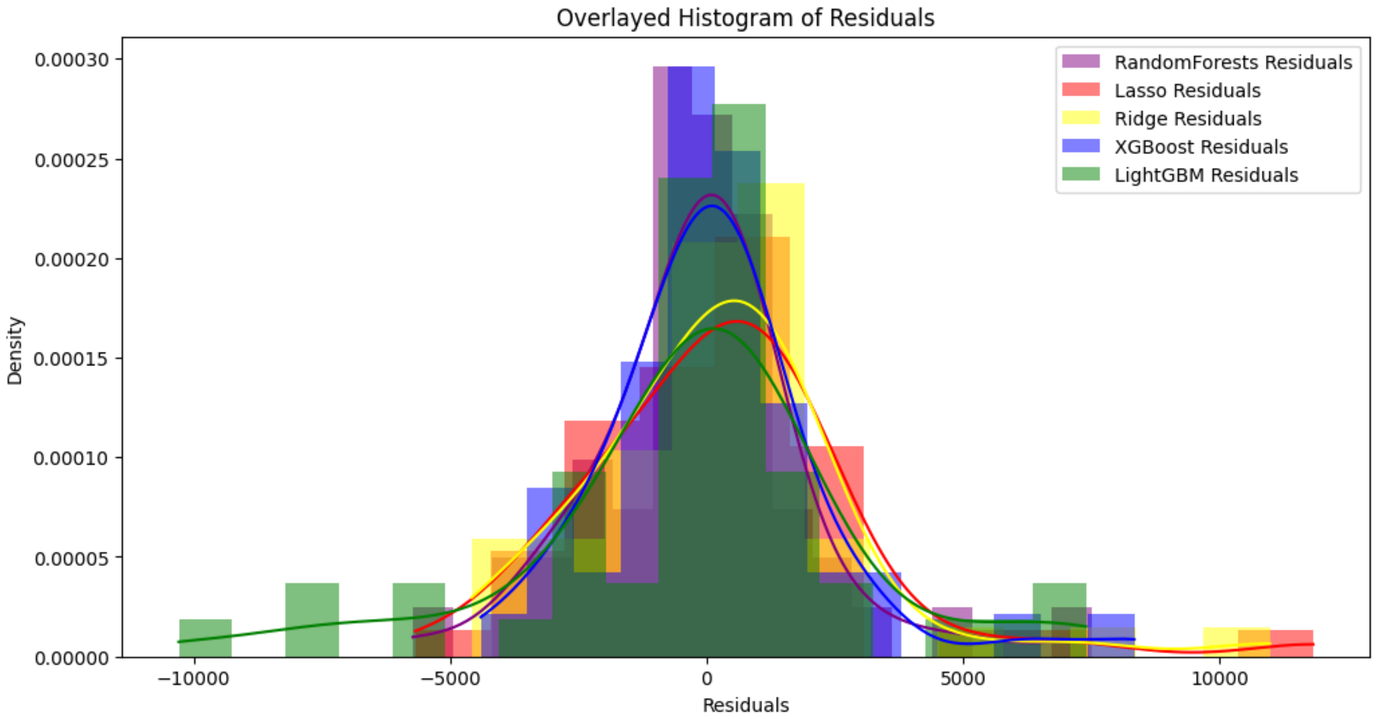
  Create a pipeline

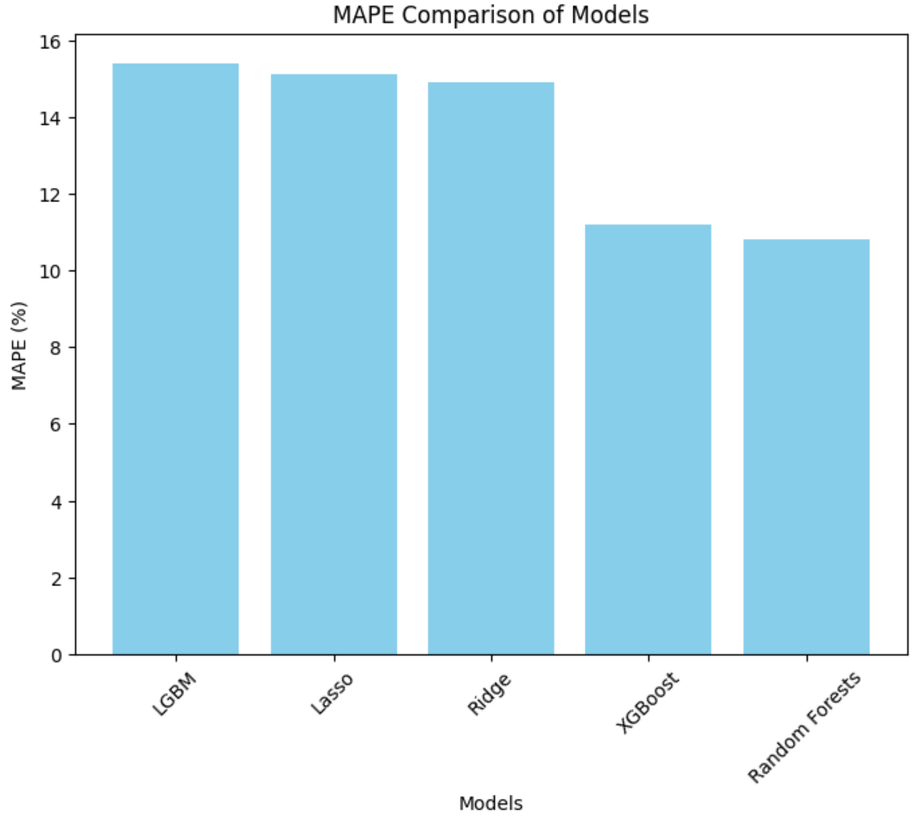
  Create GridSearchCV

  Make prediction using the best model

  Calculate RMSE, Adjusted R-squared, MAPE

Histogram of Residuals, MAPE Plot





- Residual histogram - Random Forests has narrow width and high peak value.
- MAPE plot - Random Forests model has the lowest MAPE

Evaluation Metrics

RMSE, Adjusted R-squared, MAPE

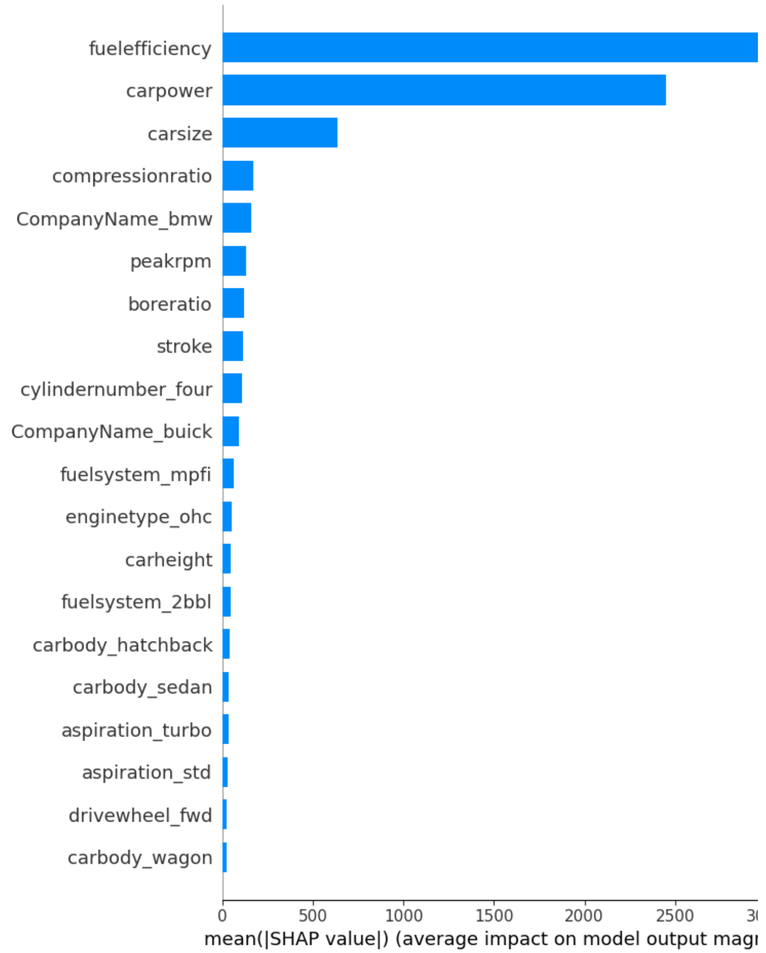
- Random Forests model has the best score across all three evaluation metrics.
- XGboost model comes close to Random Forest model.
- LightGBM model has the worst performance.

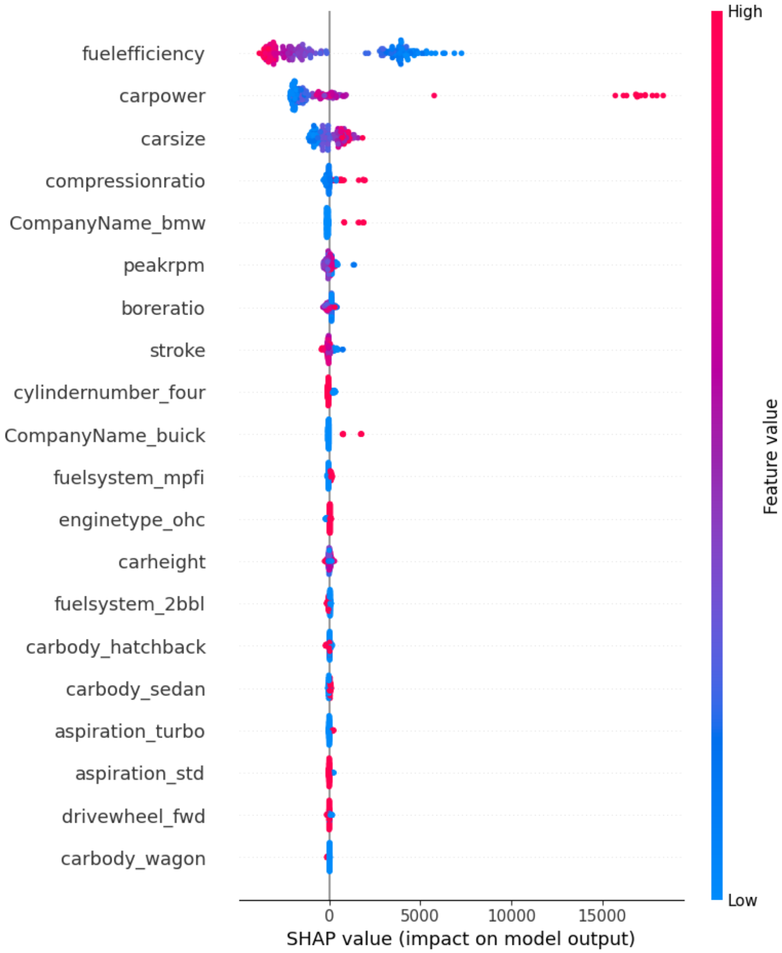
	RMSE	Adjusted R-squared	MAPE
Lasso	2756	1.38	15.1
Ridge	2613	1.34	14.9
Random Forests	2087	1.22	10.8
XGboost	2110	1.22	11.2
LightGBM	3215	1.52	15.4

SHAP

Impact on model output

- The top three features that impact model output are fuelefficiency, carpower, carsize
- Lower value of fuelefficiency tends to push car price up; higher value of fuelefficiency tends to push car price down
- Higher value of carpower tends to push car price up; lower value of carpower has moderate impact on car price
- Carsize has moderate impact on car price

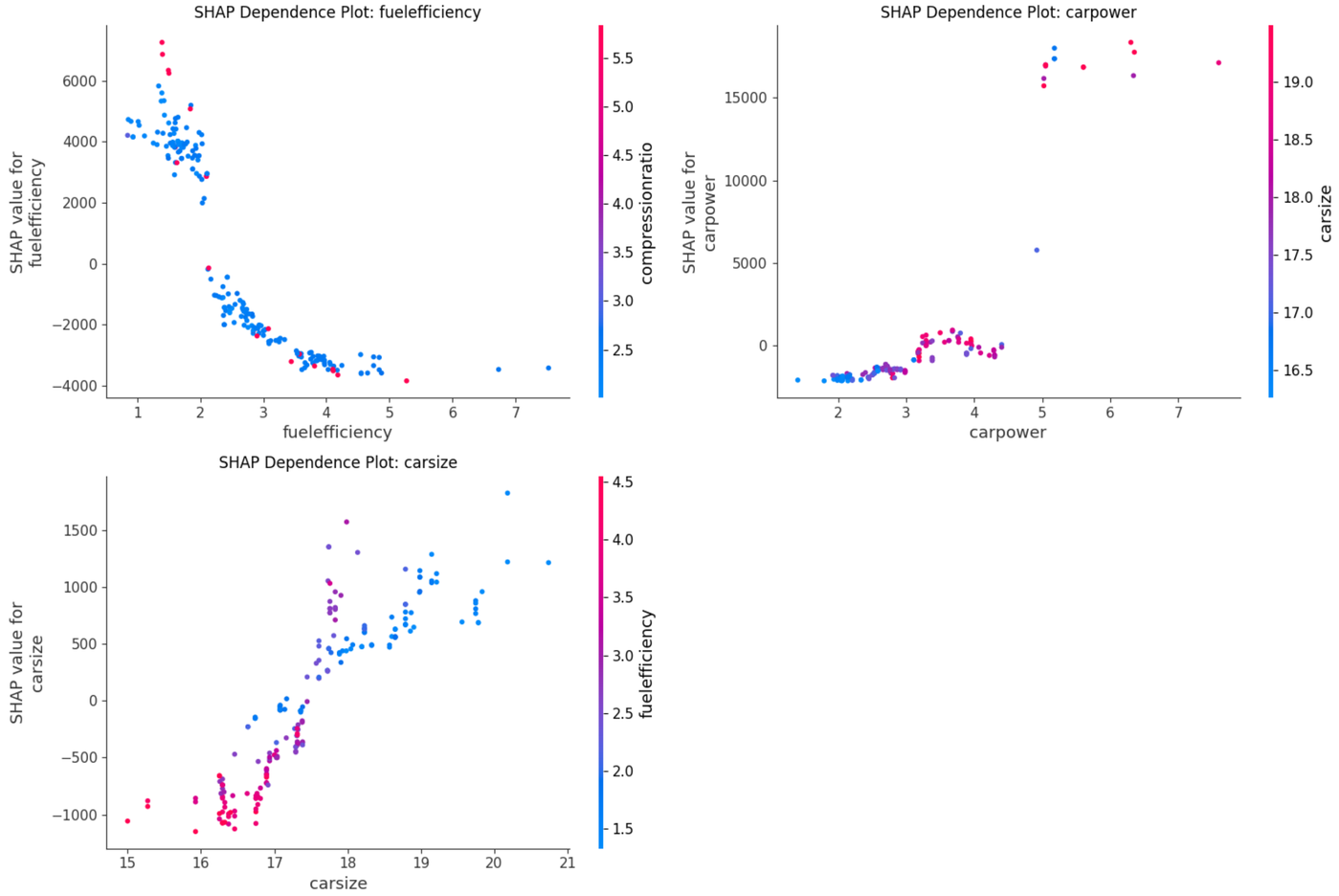




SHAP Dependence Plot

Fuelefficiency, carpower, carsize

- Fuelefficiency has negative correlation with carprice
- Compressioratio spreads out across all fuelefficiency
- Majority of low carpower has moderate impact on carprice while some extreme high carpower has very strong positive impact on carprice
- Small carsize has low carpower and large carsize typically has high carpower
- Carsize has positive correlation with carprice
- Small carsize has high fueleffiency



Conclusions and Future Work

- Six models were built to identify and understand the significant variables that influence car prices in the American market from engineered features
- Random Forests model has the best performance, evaluated by the metrics (RMSE, Adjusted R-squared, MAPE)
- SHAP analysis on Random Forests model indicates the top three features impacting car prices - fuelefficiency, carpower, carsize.
- Recommendation are provided based upon top features impact on car price

Future Work

- Explore models including GBM, Neural Network
- Apply ensemble methods
- Gather more data such as location, age, occupancy, etc.
- Expand additional hyper parameter values with advanced optimization techniques
- Create an API that is integrated with website or management system to allow dynamic pricing recommendations

Recommendations for Clients

- Fuelefficiency and carpower are the most impactful features on carprice.
- To sell expensive car, manufacturer needs to produce car with very high power, low efficiency, and big size.
- To sell cost-effective car, manufacturer needs to product car with high fuel efficiency, low car power, and small car size.

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