Predicting Car Prices
BUS 351 (006)
Cary Tang, Viktoriia Tkachenko, Surabhi Metpally

Executive Summary

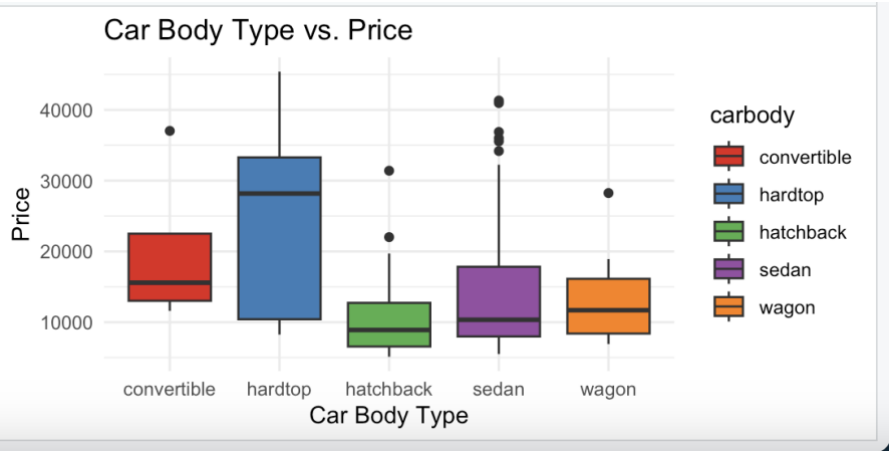
Goal of the Project : This project aims to develop a predictive model for car prices, helping Geely Auto understand key factors influencing U.S. pricing, such as horsepower, dimensions, and engine characteristics. Our team of three collaborated to create the most accurate predicting model for strategic insights.

Key Factors Affecting Car Prices:

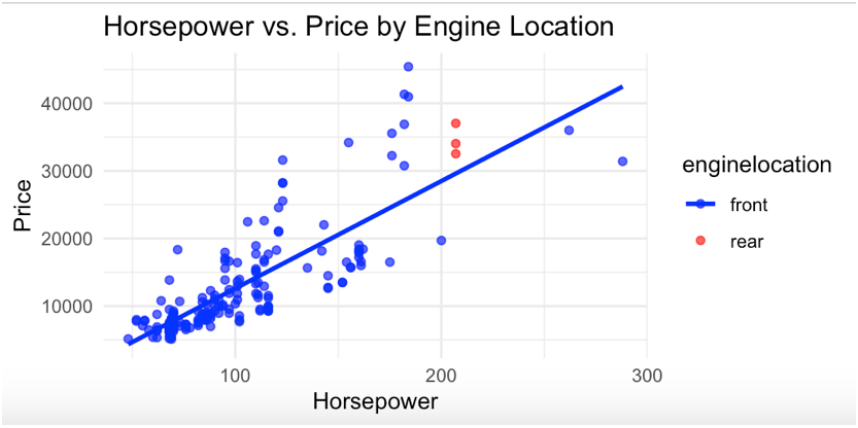
1. **Engine Location:** with front-engine cars generally being less expensive compared to rare rear-engine vehicles, which tend to be associated with high-end, luxury models.
2. **Horsepower and Curb Weight:** incorporated a combine feature called hp_weight to show that higher horsepower and heavier curb weight together often correspond to higher prices
3. **Engine Size and Power:** Larger engines (enginesize and enginesize2) often correlate with higher prices, likely due to the increased performance and manufacturing costs
4. **Car Dimensions:** Length, width, and wheelbase are comfort factors
5. **Fuel Type:** Cars with diesel engines may be priced higher
6. **Car Body Type:** SUVs and Hardtop are priced higher than compact styles

Kaggle Performance: The inclusion of additional features and the switch to Random Forest allowed the model to better capture the underlying patterns in the data, bringing the model to #2 on the Kaggle leaderboard.

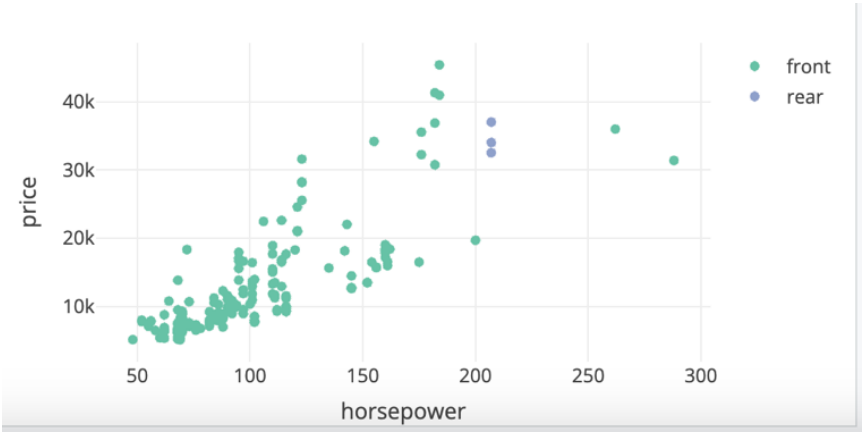
Visualization



The boxplot illustrates that there is a significant difference in price among different car body types. Hardtop cars appear to have the highest median price, followed by convertibles and sedans. The large spread in the hardtop boxplot suggests a wider range of prices for this body type. This visualization suggests that car body type is a factor influencing car prices, with hardtops being generally more expensive than other types.



The scatter plot shows the relationship between horsepower and price, with points colored by engine location. Blue points (front-engine) display a positive correlation, as higher horsepower corresponds to higher prices. Red points (rear-engine) cluster at higher prices, partially overlapping with front-engine cars. The trend line reinforces the positive correlation for front-engine cars.



The scatter plot shows the relationship between horsepower and price, with points colored by engine location. Most cars have a front-engine setup, displaying a trend of higher prices with increased horsepower. This suggests horsepower influences car prices, with rear-engine cars generally being more expensive and powerful. But due to less dominant data, Engine location appears to be a secondary factor, with rear-engine cars generally being more expensive and powerful.

Modelling and Summary

We began with a baseline linear regression model using basic variables. To improve predictive performance, we applied feature engineering and switched to a Random Forest model, adding interaction and polynomial terms like hp_weight (horsepower × curbweight) and enginesize² (square of enginesize) to capture non-linear relationships. Random Forest offered a more flexible model to handle complex data patterns. This iterative, data-driven approach, focused on refining the model, led to meaningful gains in predictive accuracy, supporting better business decisions.