XYZ Car Trading Case Study

How XYZ Company can increase the profit by using Data ?

Manoj Kumar

The problem

Company

XYZ company was founded in 2012 to change the way cars are bought and sold by connecting buyers and sellers end-to-end. It essentially aims to build a stock exchange for used cars, using technology to connect buyers with sellers.

Context

XYZ company buys used cars from OEMs, dealerships, and its own business units and sells them into its global dealership network. It also buys cars from individual sellers based on its own inventory and that of dealerships and manufacturers that use its platform.

Problem statement

Maximize the profit by using data analytics to sell and buy more cars at the right price.

Challenges deep-dive

Challenge 1

Pricing the Car

Setting the right price both while buying as well as selling the car.

Challenge 2

Supply-Chain Planning

Buying the right cars.

Challenge 3

Getting to know the customers

Understand the requirements of both the seller and the buyer.

Use-Case

How to solve the challenges using Data Analytics?

Use-Cases

1. **Pricing the Car :** 

Pricing Analytics using the various properties of the car like make, horsepower, peak-rpm, curb-weight etc and as well as insurance related features like symboling and normalized losses. This will help the business by correctly pricing the car based on these features which in order makes the business to earn profits.

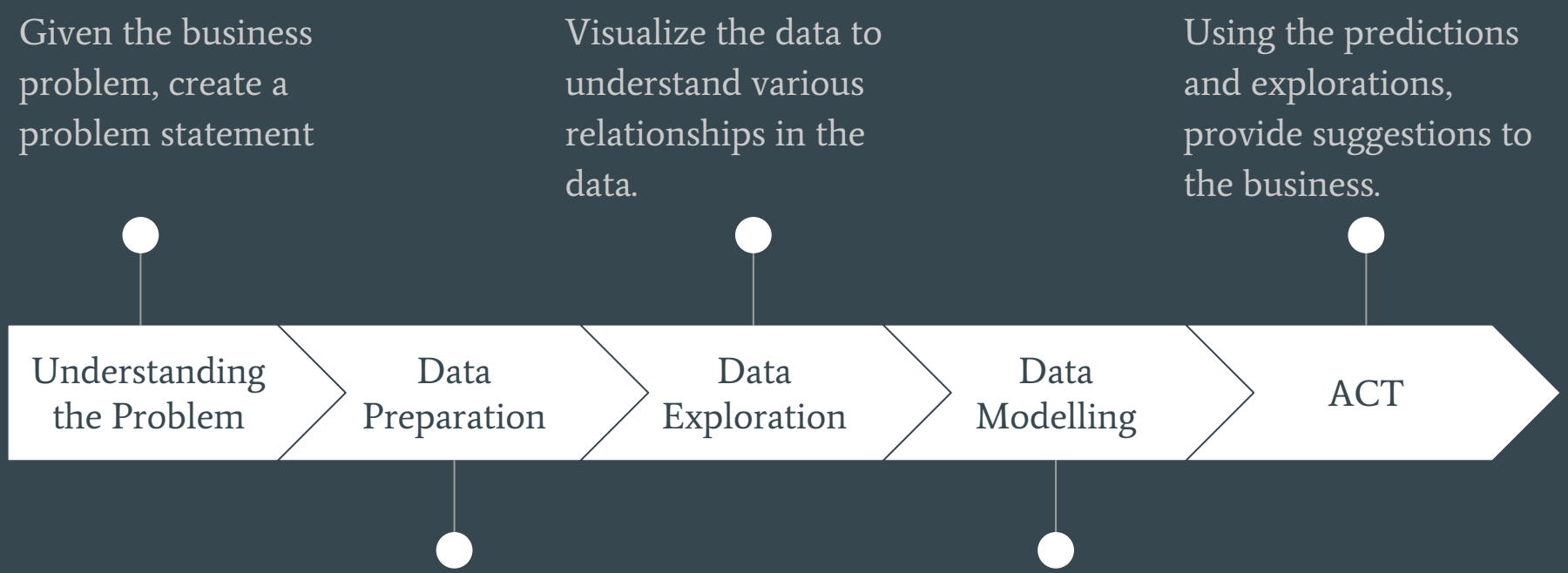
2. **Supply-Chain Planning :** 

Risk Modelling the normalized-losses will help to predict the normalized-losses given the features of a car. Based on the normalized-losses, the business can chose to buy a car, which in turn leads to a proper inventory planning and in-turn more conversion.

3. **Getting to know the customers :** 

Using the details of customers like previous car owned, their occupation etc, the right cars can be recommended (Recommendation Problem). This cannot be done from the given data.

Implementation


Given the business problem, create a problem statement
Visualize the data to understand various relationships in the data.
Using the predictions and explorations, provide suggestions to the business.
Understanding the Problem
Data Preparation
Data Exploration
Data Modelling
ACT
Clean the raw data to a format in which it can be explored and modelled.
Build a predictive model to address the problem statement.

Understanding the problem

- To build a Regression model to predict the price of a car by using the features of a car and insurance related features as the independent features.

- To build a Regression model to predict the normalized-losses using the features of a car.

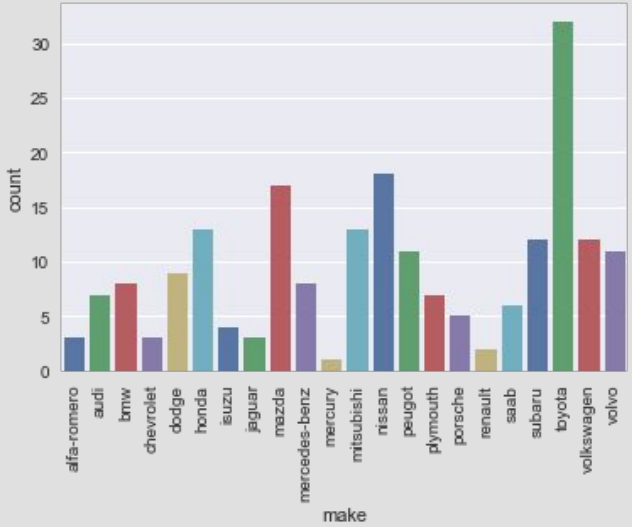
Data Available :

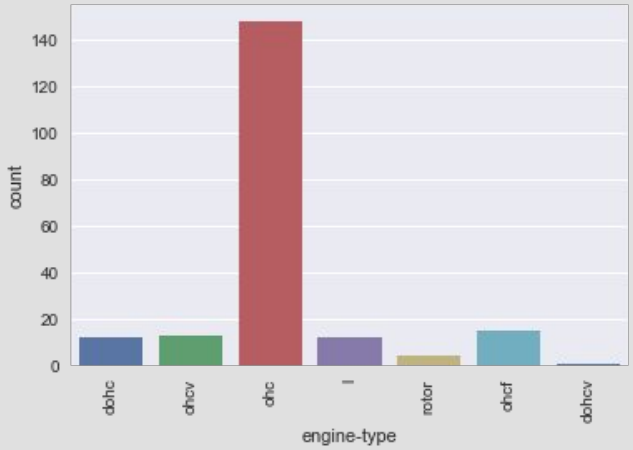
1. symboling: -3, -2, -1, 0, 1, 2, 3.
2. normalized-losses: continuous from 65 to 256.
3. make: alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
4. fuel-type: diesel, gas.
5. aspiration: std, turbo.
6. num-of-doors: four, two.
7. body-style: hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels: 4wd, fwd, rwd.
9. engine-location: front, rear.
10. wheel-base: continuous from 86.6 120.9.
11. length: continuous from 141.1 to 208.1.
12. width: continuous from 60.3 to 72.3.
13. height: continuous from 47.8 to 59.8.
14. curb-weight: continuous from 1488 to 4066.
15. engine-type: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. num-of-cylinders: eight, five, four, six, three, twelve, two.
17. engine-size: continuous from 61 to 326.
18. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore: continuous from 2.54 to 3.94.
20. stroke: continuous from 2.07 to 4.17.
21. compression-ratio: continuous from 7 to 23.
22. horsepower: continuous from 48 to 288.
23. peak-rpm: continuous from 4150 to 6600.
24. city-mpg: continuous from 13 to 49.
25. highway-mpg: continuous from 16 to 54.
26. price: continuous from 5118 to 45400.

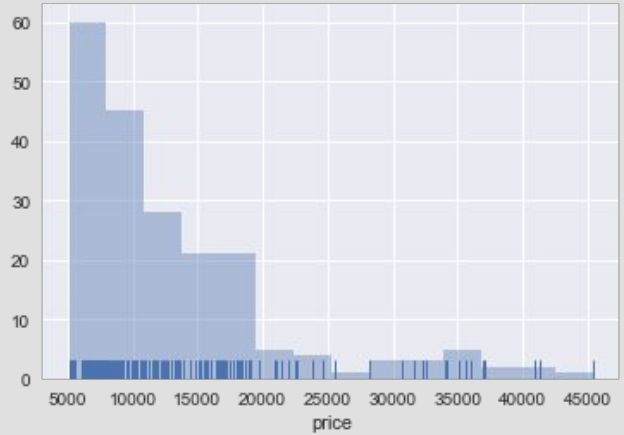
Data Preparation

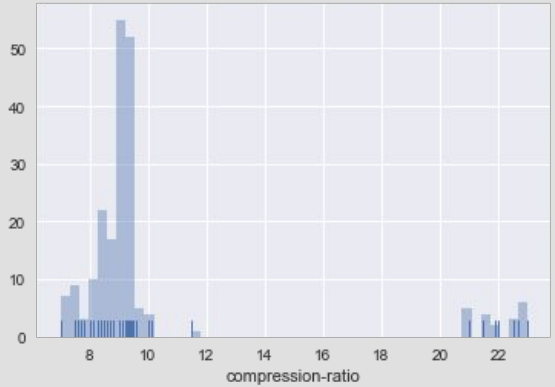
- Convert the attributes to its appropriate data types.
- Check for unnecessary characters like ‘?’.
- Check for nulls/NaN’s.
- Imputing the nulls/NaN’s using KNN Imputation.
- Scaling the data
- Split into Train and Test (70:30)

Data Exploration (Uni-Variate Analysis)

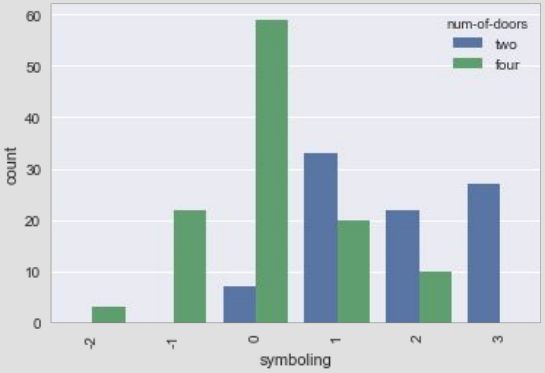


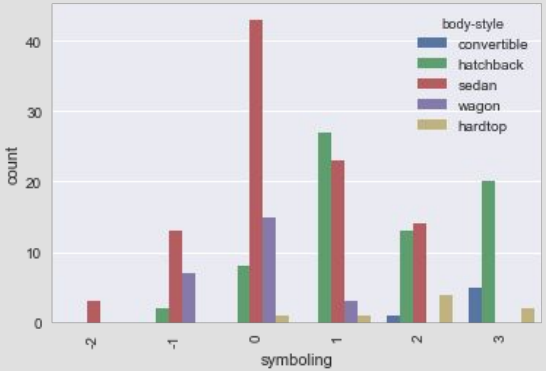


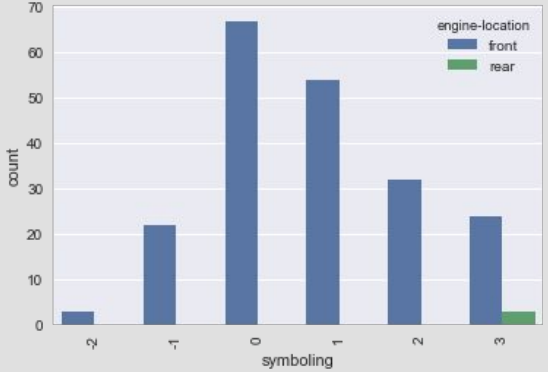


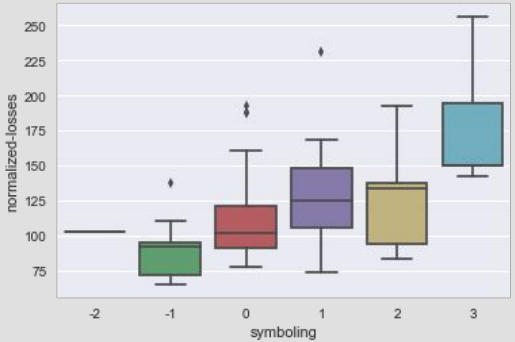


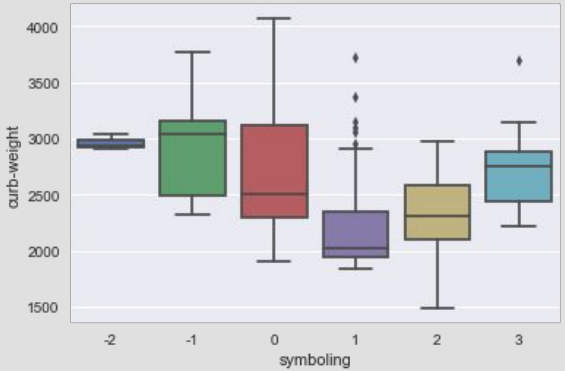
Data Exploration (Uni-Variate Analysis)

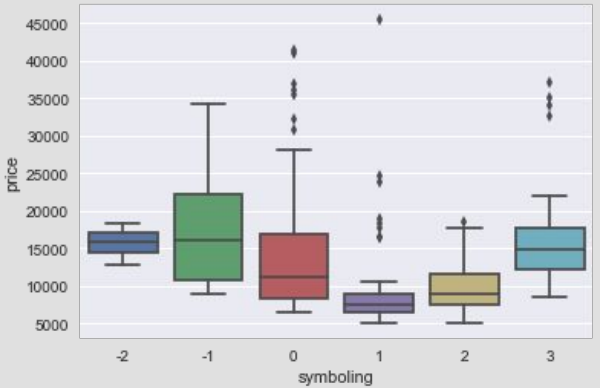












Analysis from Data Exploration

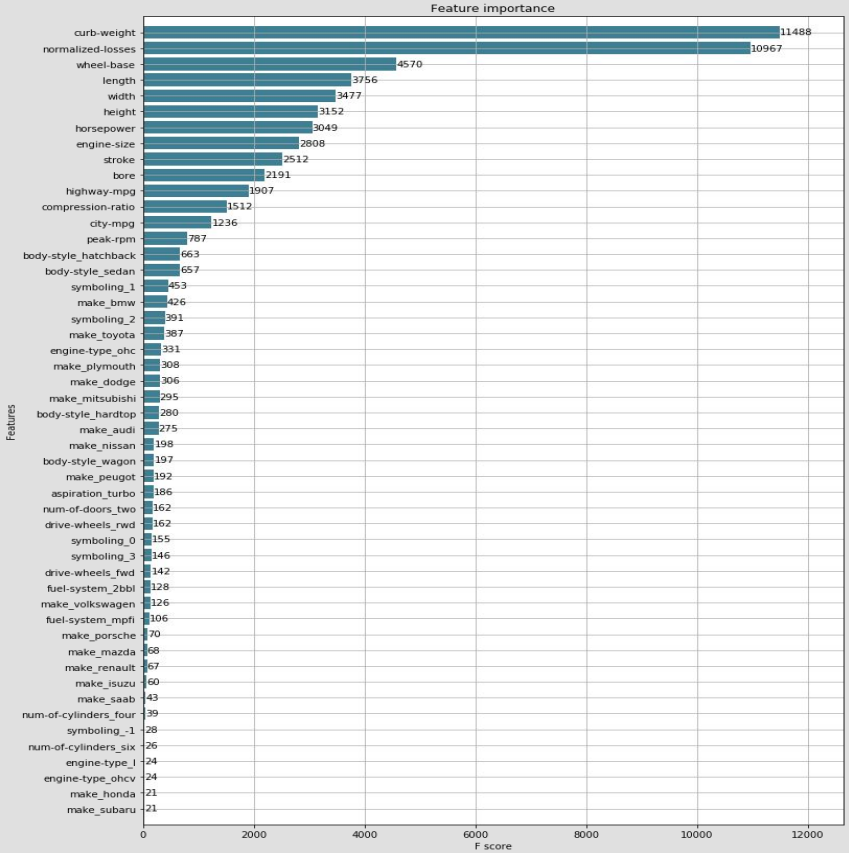
- Toyota has the most number of vehicles.
- Chc engine-type is the highest used engine-type.
- Price variable is left skewed.
- Compression-ratio seems to be bimodal.
- Two door cars seems to increase the insurance rating making it risky.
- Hard-top body-style makes it more risky.
- All the rear engine-location are risky.
- More the normalized-losses, more the risk rating.

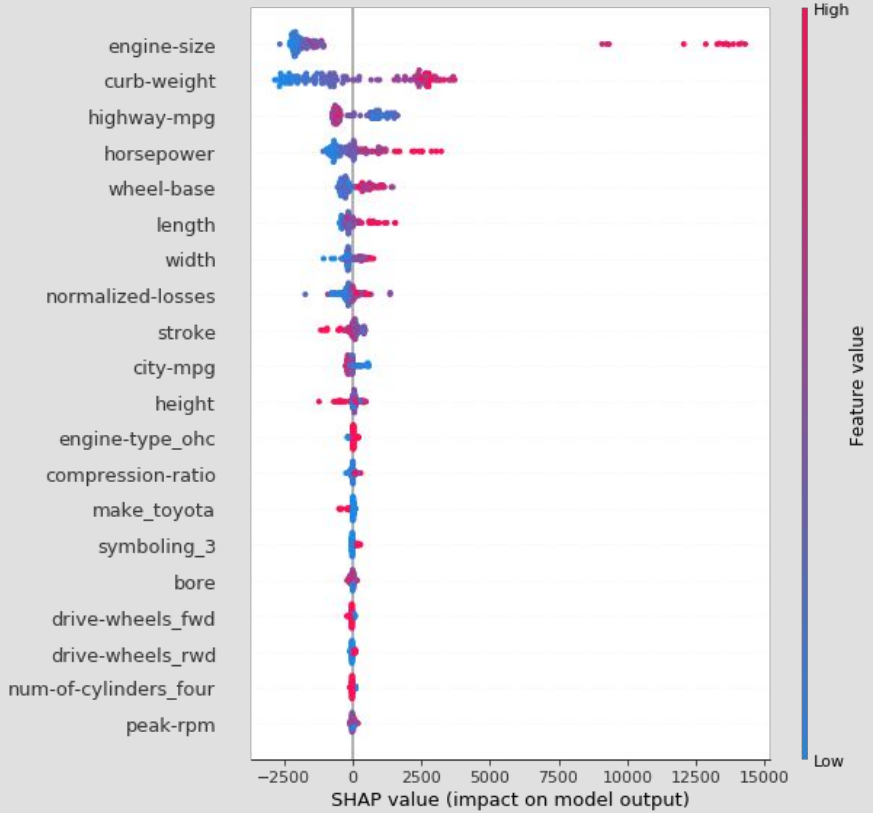
Data Modelling (For Predicting price of a car)

Model	MAE	MAPE	R2 Score
Linear Regression	1586790752436320.5	-	-1.1390190813029911e+24
Decision Tree	1540.42	13.14	0.88
Random Forest	1393.96	11.87	0.89
Xgboost	1261.50	10.75	0.91

* Metrics on Test Data

Data Modelling Plots

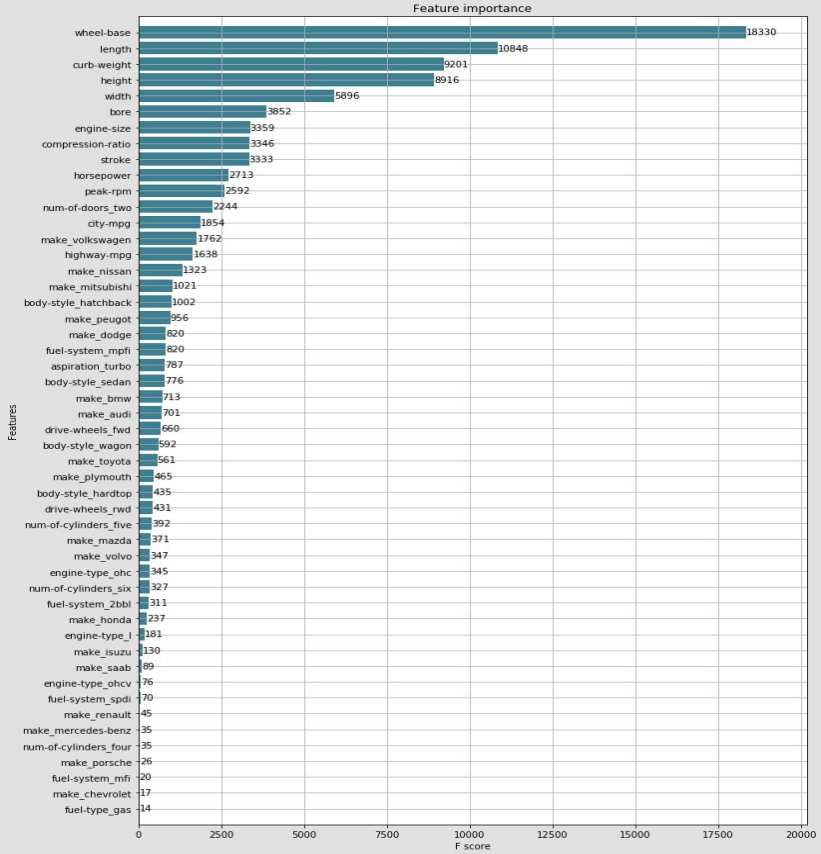




This reveals for example that a high curb-weight increases the predicted car price.

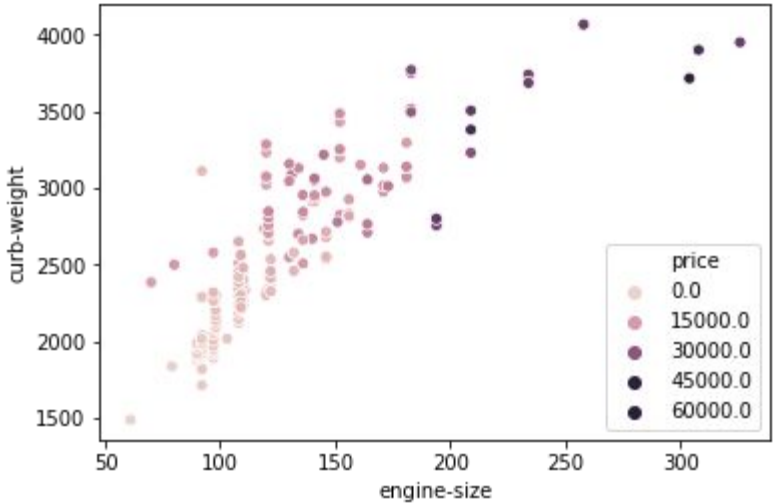
Data Modelling (For Predicting Normalized-losses)

Model	R2 Score
Xgboost	0.80



Impact

Accurately predict price based on curb-weight and engine-size



Conclusion

- Xgboost model came out as the best.
- We now have models to predict price as well as normalized-losses.
- Using these models, we can price the car correctly as well as buy the right cars.
- Curb-weight and Engine-size came out as the most important in predicting car prices.
- Wheel-base, length, curb-weight and height came out as important in predicting normalized losses.
- Now we can select cars with less normalized-losses based on just the features of a car.