*MSRP & Me: Finding the car that's right for you

Darren Lyles

image courtesy of:
http://www.cartoonswallpapers.net

*MSRP: Manufacturer's Suggested Retail Price

Agenda

- Introduction
- Data Set Overview
- Data Wrangling
- Initial Findings
- In Depth Exploratory Data Analysis
- Machine Learning Analysis
- Recommendations and Conclusion

image courtesy of: https://techflourish.com

Introduction

image courtesy of:
https://blog.dealersimplified.com

- Buying a car is a key milestone in one's life.
- Who is the main audience?
  - Car Buyers
  - Marketing Analysts
  - Car Manufacturers
- What are some questions to consider?
  - I want a car with these features. What is my price point?
    - Car Buyers
  - Between Kia and Toyota, which brand has the competitive price edge?
  - Is Mercedes more expensive than its competitor, BMW?
    - Marketing Analysts and Car Manufacturers
- I will be predicting a car's price based on its features to provide the answers.

Key Question: Why should you care?

- Customers can determine what price they expect to pay for their desired features, and decide if a car is priced reasonably.
    - Example: A customer wants a Kia with 150 HP, 4 cylinders, 4-door, automatic transmission, 35 city MPG, and 40 highway MPG. The predicted price is $20,000.
- Businesses and marketing analysts can price cars competitively for their respective markets.
    - Example: Using the previous example, Kia wants to compare a car from Toyota with the exact features, and it turns out that the car from Toyota would be $23,000.
        - Kia then has a competitive edge for the given car

Data Set Overview

- Original data set acquired from Kaggle.com
- Original contained 11,914 rows, 16 columns with the following features:

Make	*Engine HP*	**Number of Doors**	*Highway MPG*
Model	**Engine Cylinders**	**Market Category**	*City MPG*
Year	**Transmission Type**	**Vehicle Size**	*Popularity*
Engine Fuel Type	**Driven Wheels**	**Vehicle Style**	*MSRP*

Key: **Categorical Data**
Numeric Data

- To define the scope of the problem, I only considered the cars manufactured for 2017
  - Newly defined data set now contains 1,668 rows, 16 columns

Data Wrangling

- Cleaning the data to be ready for analysis
- What to look for?
  - Data Type Consistency (i.e. Need to make sure all data in the column are the same type)
  - NaN Values (i.e. missing values)
- Features which had missing or inconsistent values were Engine Horsepower and Market Category, with 0.96% and 23%, respectively. Since the Market Category feature had significant missing data, it had to be dropped.
- A row containing features for the 2017 Audi A6 Sedan was removed since it had a Highway MPG of 354, which was fact checked (fueleconomy.gov) and clearly incorrect.

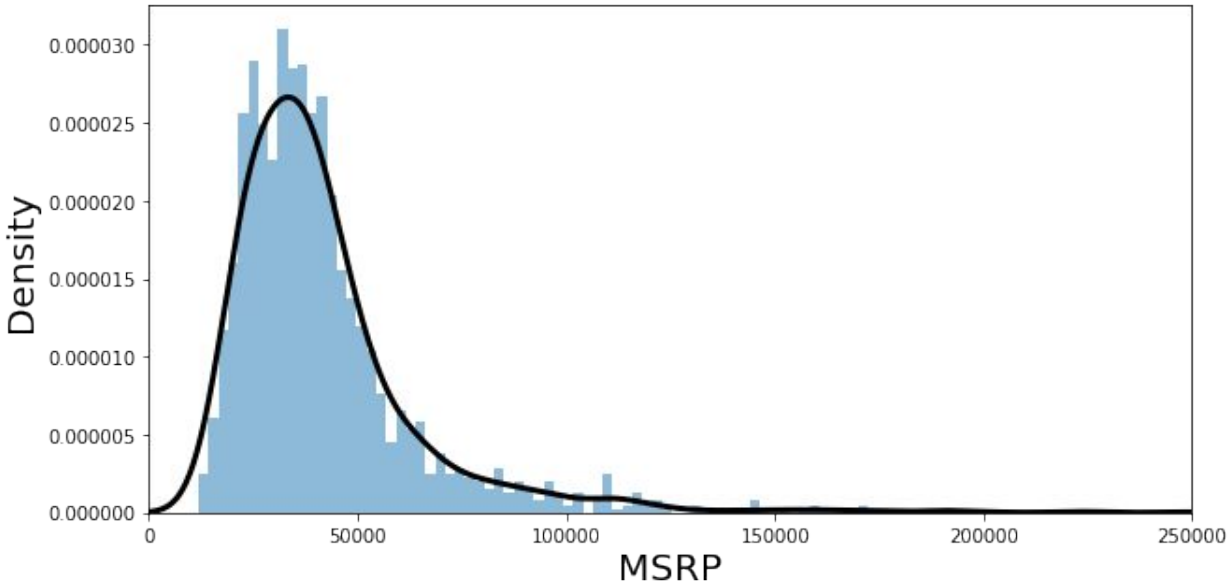
Initial Findings

Number of Auto Brands	30
Manual Transmission	205
Automatic Transmission	1356
Automated Manual Transmission	98
Direct Drive Transmission	9
2-Door Cars	276
3-Door Cars	16
4-Door Cars	1376

- The data set is limited on the variety of car brands, which could be a limiting factor in analyzing the data.
- There are an overwhelming majority of cars with automatic transmission, which could limit the buyer, if an alternative transmission (such as manual) is desired.
    - Manual transmission is usually cheaper to implement and could play a factor in a car's price.
      https://www.consumerreports.org/cro/2012/01/save-gas-and-money-with-a-manual-transmission/index.htm
- Interesting find: There are three-door cars that were all 2017 Ford Transit Wagons

In-depth Exploratory Data Analysis

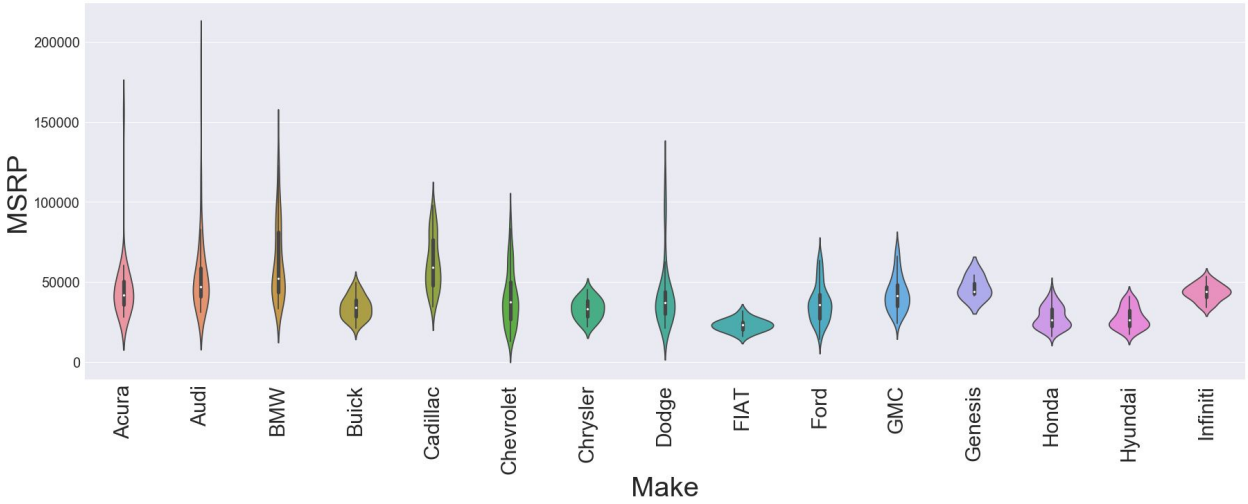
Exploratory Data Analysis (EDA)



The car pricing data we have is roughly centered around $30,000.

However, on the right end of the graph, we see that there are quite a few high end luxury cars in the market, which is definitely a niche for some car brands.

Pricing Distributions by Make

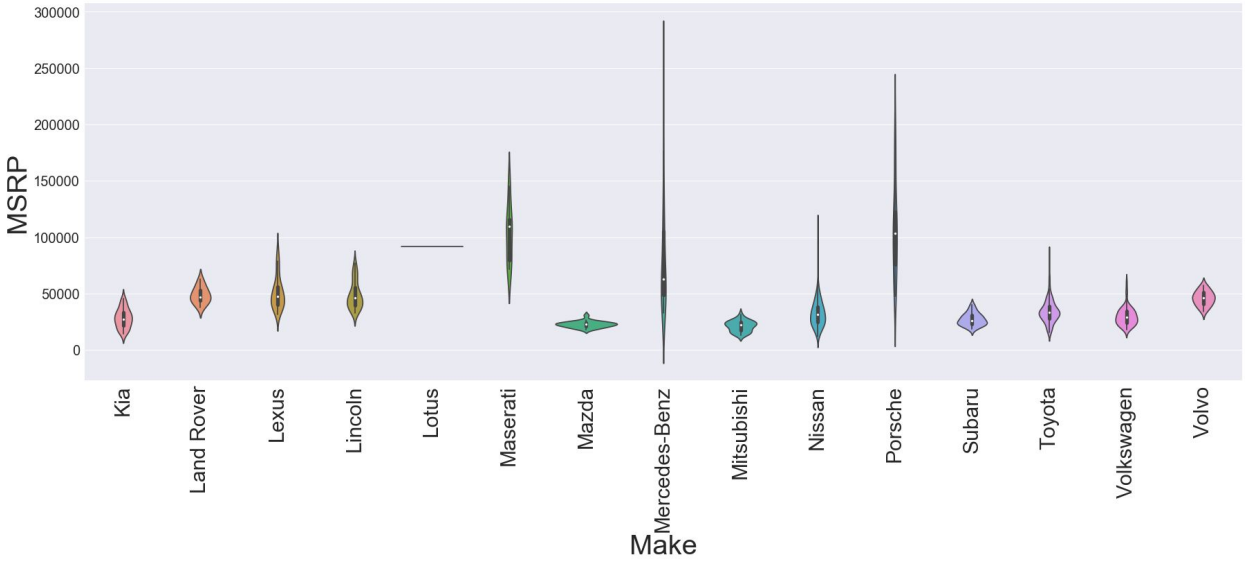


In the distributions by brand, we see there are brands that have small and large spreads in pricing.

Brands with small price spreads (i.e. Buick, Chrysler, FIAT, etc.) and a low median price cater to buyers interested in affordable options.

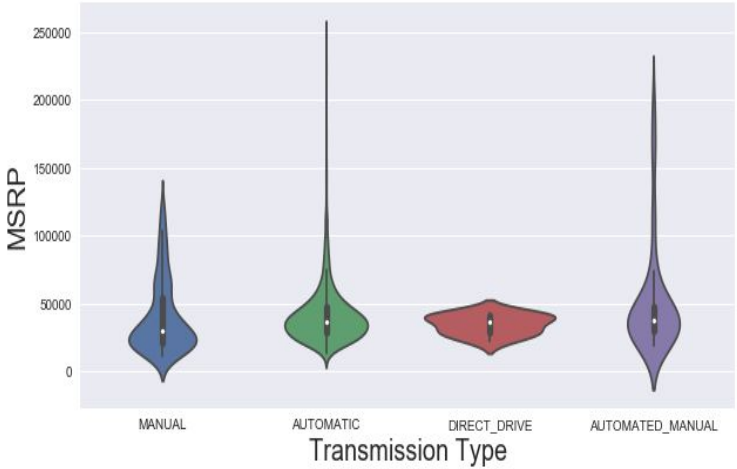
Brands with large price spreads (i.e. Audi and BMW) attract a wider range of consumers, especially with a luxury series for high-end buyers.

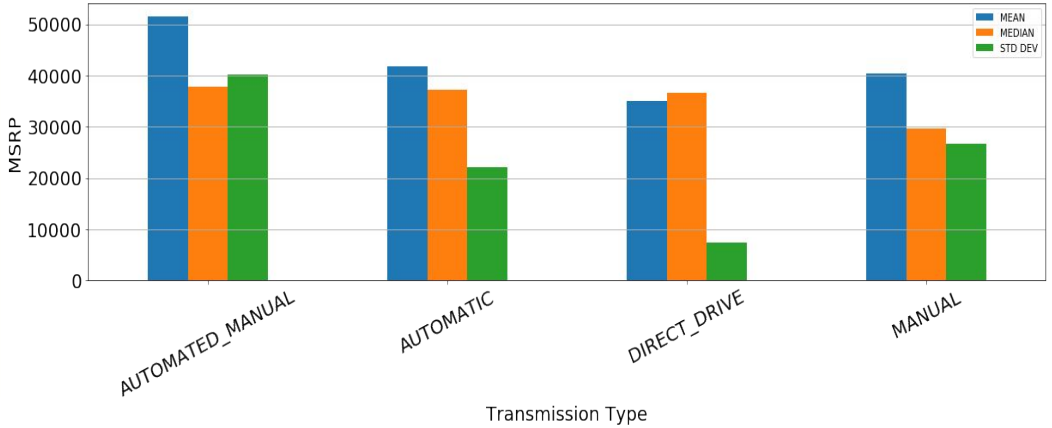
Pricing Distribution vs. Make (cont.)



Interesting note: Mercedes-Benz has the largest spread with a median price of about $60,000. The spread hints that there are some affordable options offered by the luxury brand.

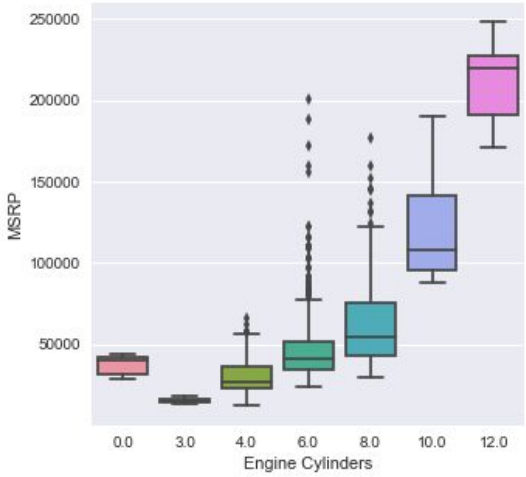
Pricing by Transmission





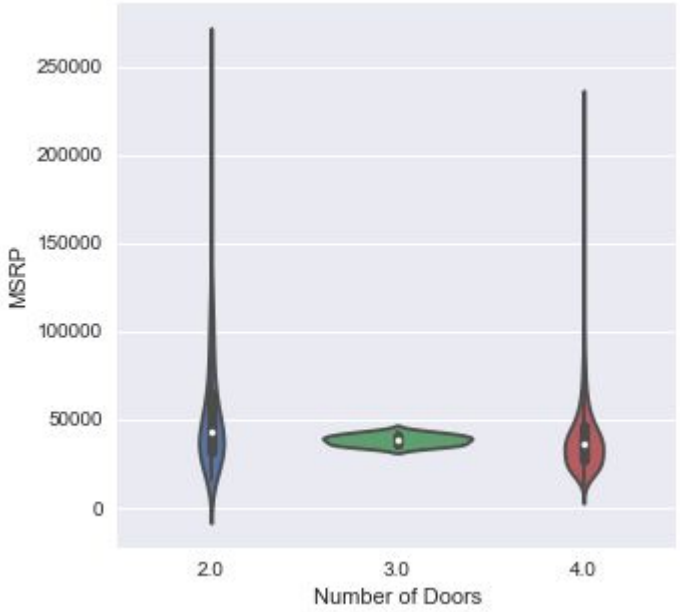
For any buyer who is looking at transmission options, cars with manual transmission are the cheapest options. This is mainly because the manual transmission is the cheapest to implement.

Pricing by Number of Cylinders



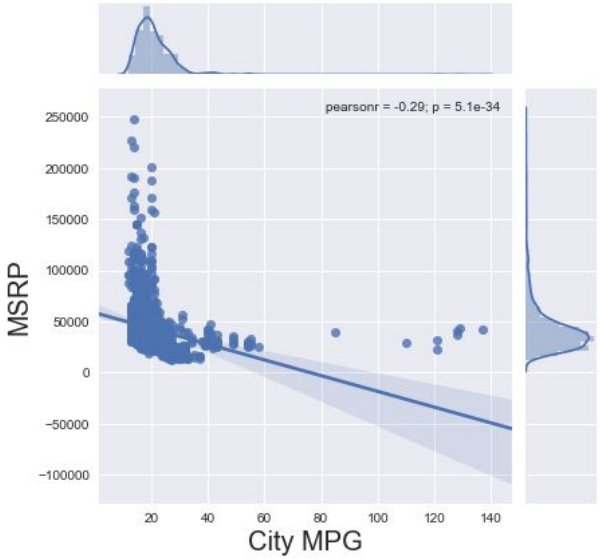
- 0 Cylinders implies the vehicles are electric.
- With the exception of electric cars, the higher the engine cylinder count, the higher the median price.
- Cars with a high number of engine cylinders tend to be luxury sports and performance cars.
- Examples:
  - 10 cylinder: Audi R8 and Dodge Viper, and the
  - 12 cylinder: Mercedes-Benz Maybach, S-Class, and SL-Class
  - $100k+

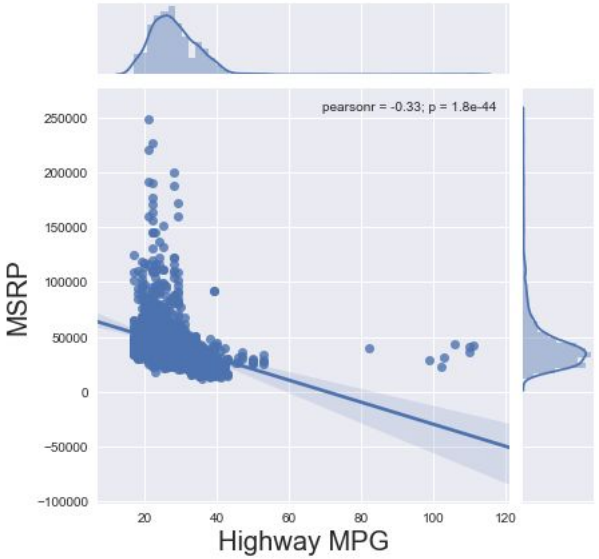
Pricing by Number of Doors



- Interesting category: 3-Doors
  - These are exclusively 2017 Ford Transit Wagons
- 2-Doors: Highest median and widest spread in pricing
  - Possibly because many 2-Door cars are likely high-performance sports cars
- 4-Doors: Lowest median and more narrow spread compared to 2-Doors
  - Recommend 4-Door cars for more affordable options

Relationship between Pricing and *MPG

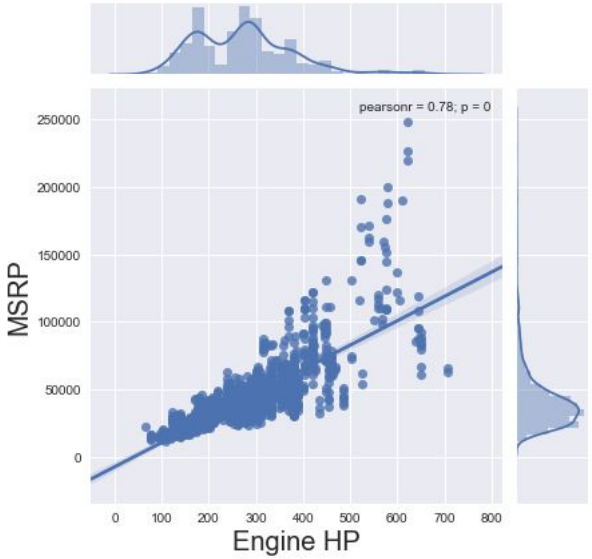




- MSRP has weak negative correlations with both city and highway MPG:
  - MSRP vs. City MPG
    R=-0.29 R^2=0.0841
  - MSRP vs. Highway MPG
    R=-0.33 R^2=0.1089
- City and Highway MPG has an effect on pricing, but not significant
- From a buyer's standpoint, MPG is not too important when determining car price

*MPG=Miles Per Gallon

Relationship between Pricing and HP



- Unlike the MPG features, MSRP has a strong positive correlation with horsepower.
  - R=0.78 R^2=0.6084.
- This implies that low horsepower cars will be more affordable than high horsepower cars.
- High horsepower cars are usually sports and performance vehicles, which, by nature, are more expensive.

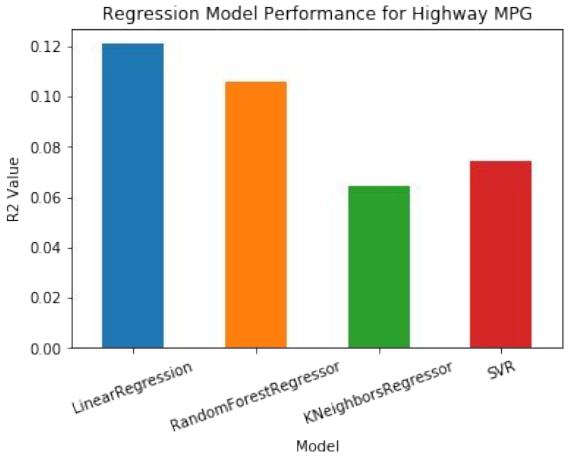
*HP=horsepower

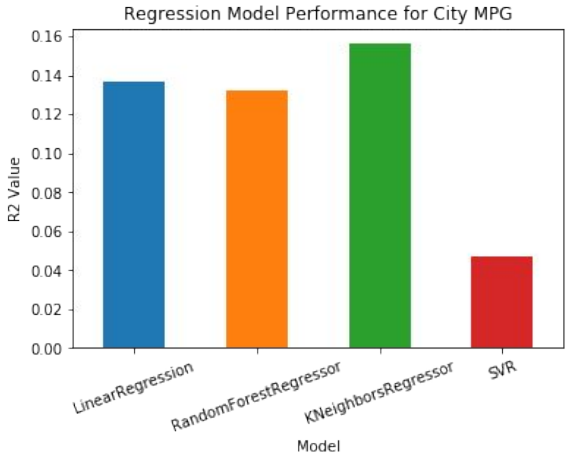
Machine Learning Analysis

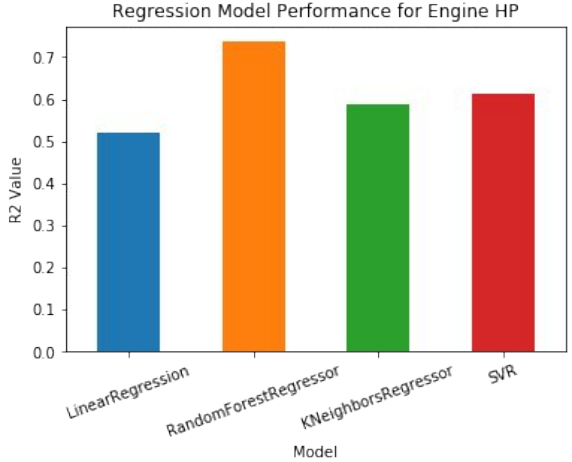
Overview

- This a supervised learning problem for predicting price with using three numerical features and four categorical features
  - Features: Make, Engine Horsepower, Number of Cylinders, Transmission Type, Number of Doors, City MPG, and Highway MPG
  - Target: Car Price (MSRP)
- Various models performance compared with numerical features
  - Linear Regression
  - Random Forest Regression
  - K Nearest Neighbors Regression
  - Support Vector Machine Regression
- Overall best R^2 value for numerical features combined will determine the model to select

Comparing Model Performance

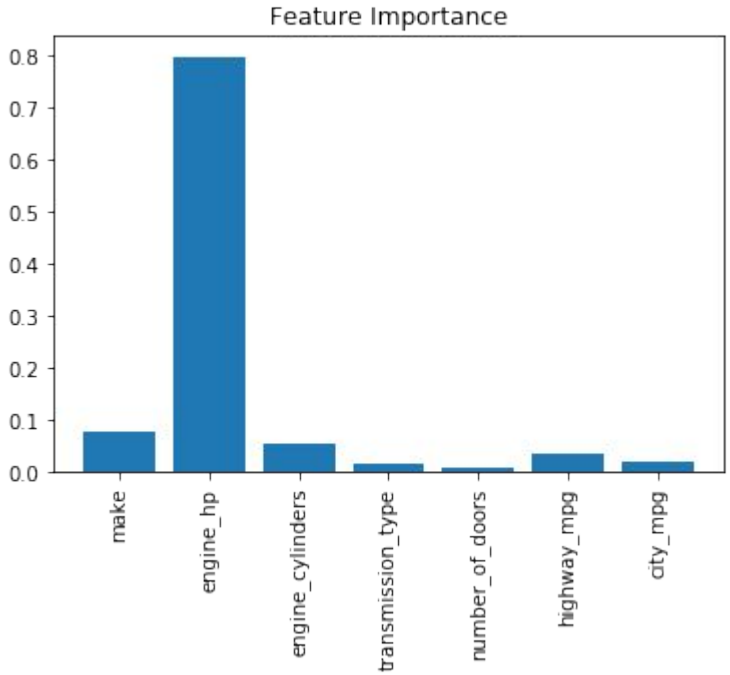






In the bar graphs above, I compared the R^2 scores of the models for highway MPG, city MPG, and engine HP. Overall, the Random Forest Regression model has the best overall goodness of fit.

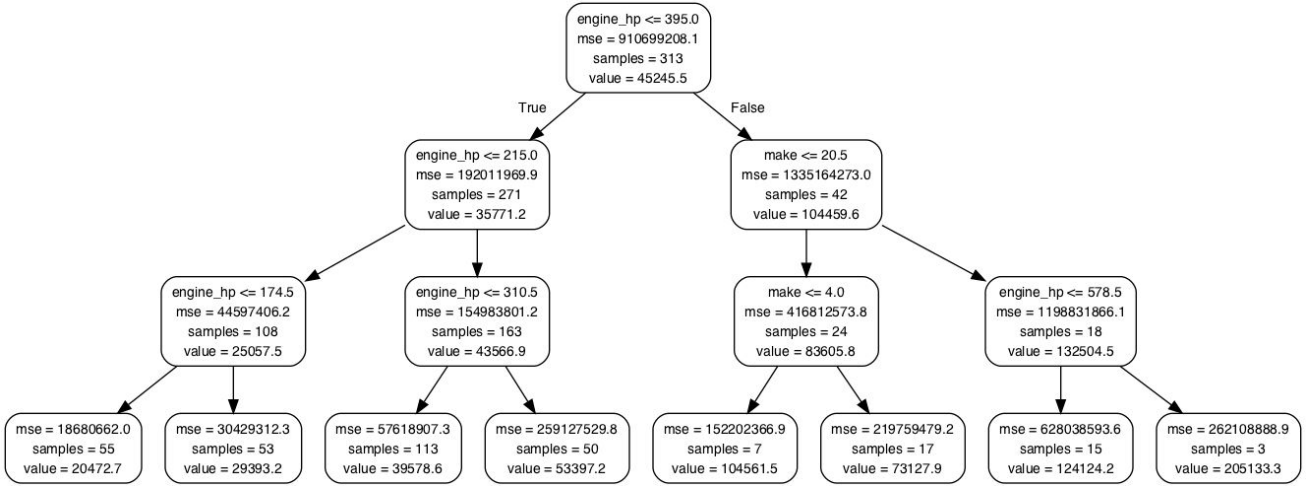
Feature Importance in Random Forest



Horsepower plays a crucial role in predicting a vehicle's price, followed by its brand and number of cylinders, respectively.

For buyers looking to make an informed decision, these features are the most critical to consider. These factors also are important for car manufacturers looking to make vehicle prices competitive in target markets.

Random Forest



The graphic here shows a demonstration of the random forest model.

It involves decision trees which will determine a target value base on the condition of the features supplied by the test set.

Applying the Model

- We can look at examples of applying our model to answer the questions introduced earlier in this presentation:
  A. I want a car with these features. What is my price point?
  B. Between Kia and Toyota, which brand has the competitive price edge?
  C. Is Mercedes more expensive than its competitor, BMW?

Applying the Model (cont.)

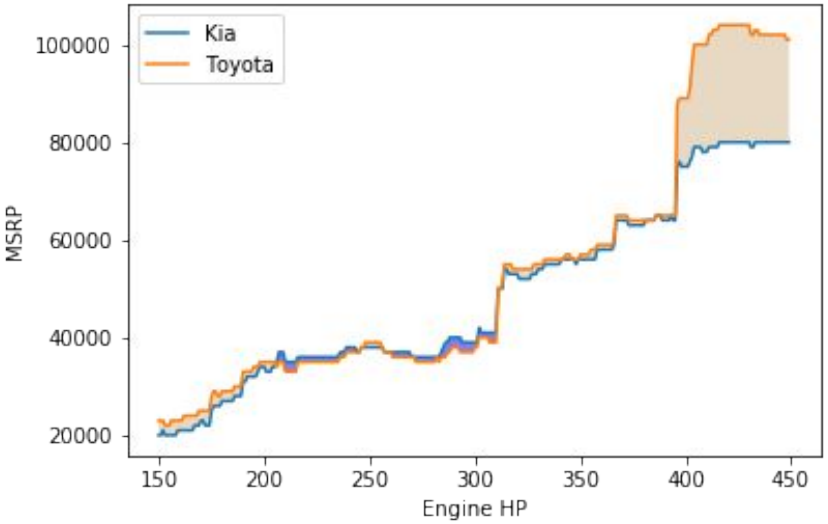
- To address question A, we simply call the Random Forest Regressor model with the following inputs:
  - make = Kia
  - engine_hp = 150
  - engine_cylinders = 4
  - transmission_type = AUTOMATIC
  - number_of_doors = 4
  - highway_mpg = 40
  - city_mpg = 35

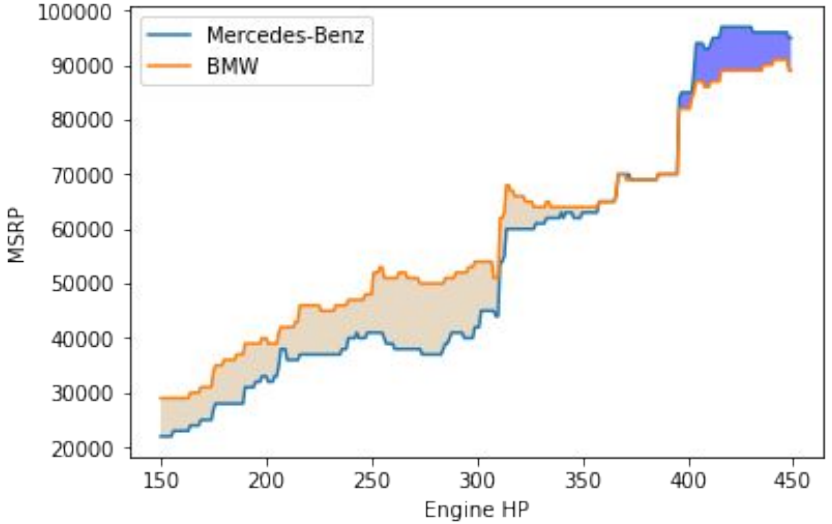
- The resulting output is ***$20,000***

Applying the Model (cont.)

- To answer questions B and C, a single prediction output will not suffice
- we need to have a more comprehensive approach to providing insight.
- Since engine HP plays the most significant role in predicting the MSRP of a car, we can let this feature vary while keeping the other features at fixed value.

Applying the Model (cont.)





Applying the Model (cont.)

- As shown in the two plots above the answers to questions B and C would be that it depends.
- In the case of Kia vs. Toyota, Kia has an overall competitive edge against Toyota when it comes to pricing vs HP.
    - There are, however, small intervals where Toyota is priced lower, however it's not sufficiently significant to compete against Kia.
- Whereas in the case of Mercedes-Benz vs. BMW, Mercedes-Benz is the more overall competitive brand until you compare cars that have an engine HP of over 400. In the 400-450 HP interval, BMW has more competitive options.

Recommendations and Conclusion

- Features crucial to car pricing are: make, engine HP, engine cylinders, transmission type, and number of doors.
- Makes or brands with large price spreads cater to diverse customers, ranging from economy to luxury models, while makes with small spreads cater to niche markets (exclusively compact/economy to exclusively luxury)
- For buyers and car manufacturers, it is recommended to minimize features, such as engine HP and engine cylinders, and focus on brands that mainly sell compact/economy cars. If a manual transmission option is available, that also helps.
- Most importantly, customer and business satisfaction!

Thank You

image courtesy of: https://techflourish.com