Car Manufacturer Suggested Retail Price Prediction

Lanxi Liu

URL: https://github.com/LanxyL/Capstone2

Why?

- Demand for cars is high in US:
  - 89% of all American adults have a driver’s license (>25 years old)
  - The U.S. auto industry sold over 3.4 million cars in 2020
- Buying a car is stressful:
  - There is no much transparency about the price
  - Too many brands, models and features to choose from

Who might be the potential user?

Car Buyers

- What kind of car I can afford?
- Which feature affect the most on price?
- What will be a reasonable price for a set of features I chose?
- Is this car model overpriced?

Car Manufacturers

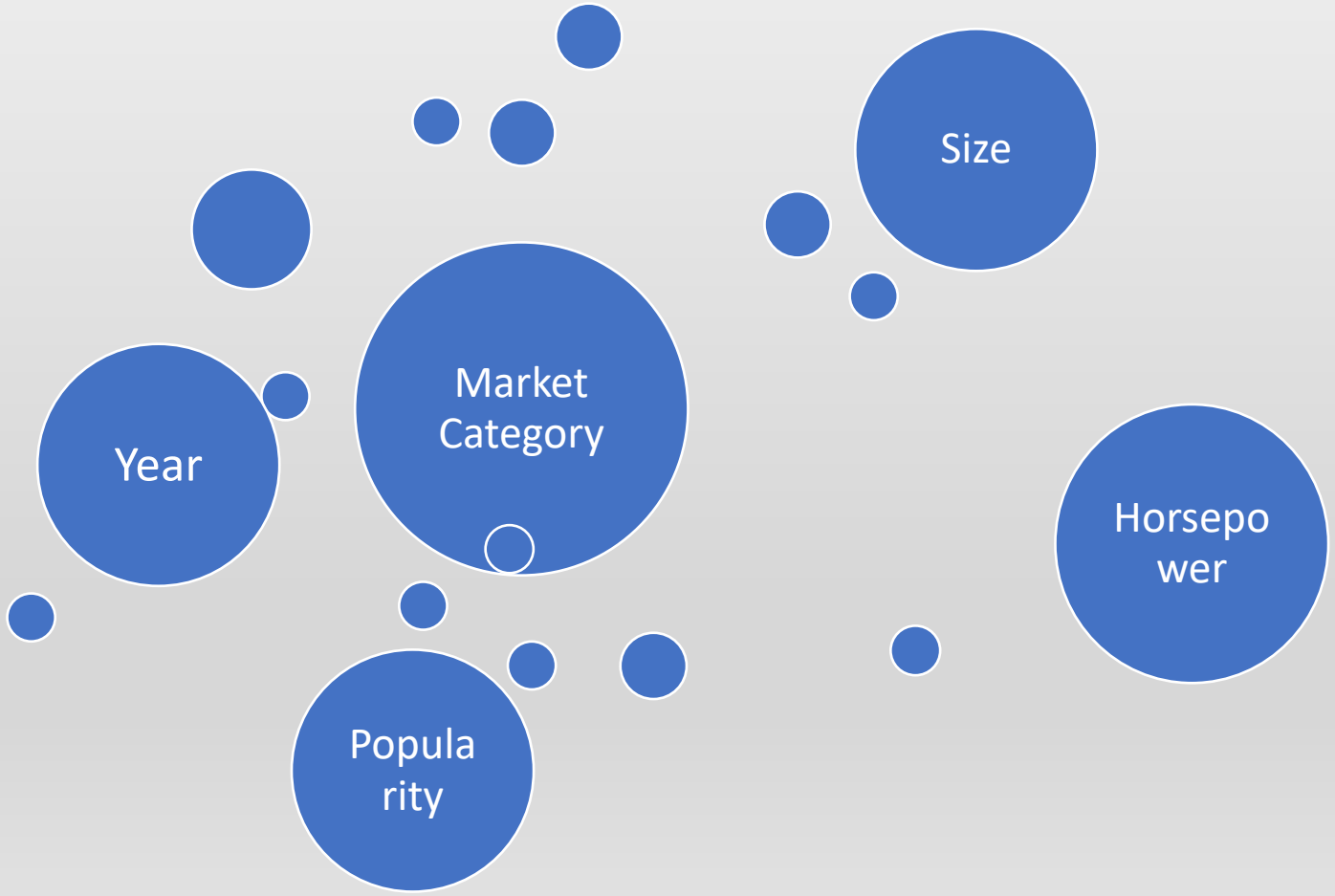
- What price should we set for the new launched car?

Car Dealers

- What would be a reasonable price for a particular car?

What features might affect the retail price?

- Data source: Kaggle dataset scraped from Edmunds and Twitter



About the Data

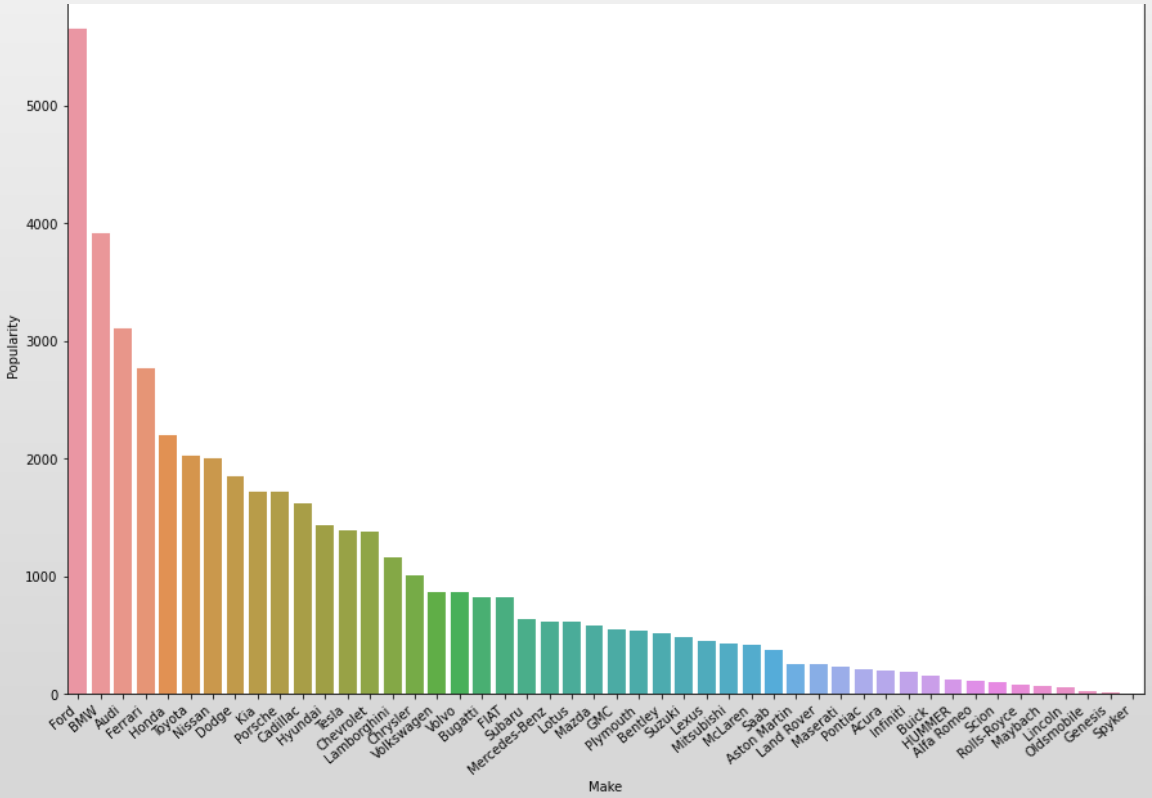
- Year range: 1990-2017
- It covers 48 brands with 11914 entries
- Price range: $2,000-$2,065,902
- 16 Features included:

  (Make; Model; Year; Engine Fuel Type; Driven Wheels; Number of Doors; Market Category; Vehicle Size; Vehicle Style; Highway/City MPG; Popularity; MSRP)

Data Wrangling

- There are 5 columns with missing values.
- Most missing values on “Market Category”
  - No clear pattern about the reason
  - Data might be lost at the parsing process or the original data has no category assigned
- Engine horsepower has second-most missing data
  - Most information were filled up with horsepower found on the internet
  - The rest are for electric car which has no horsepower information
- Dropping all duplicate rows: 11914 entries left
- Correct a wrong entries which found as an outlier

Exploratory Data Analysis

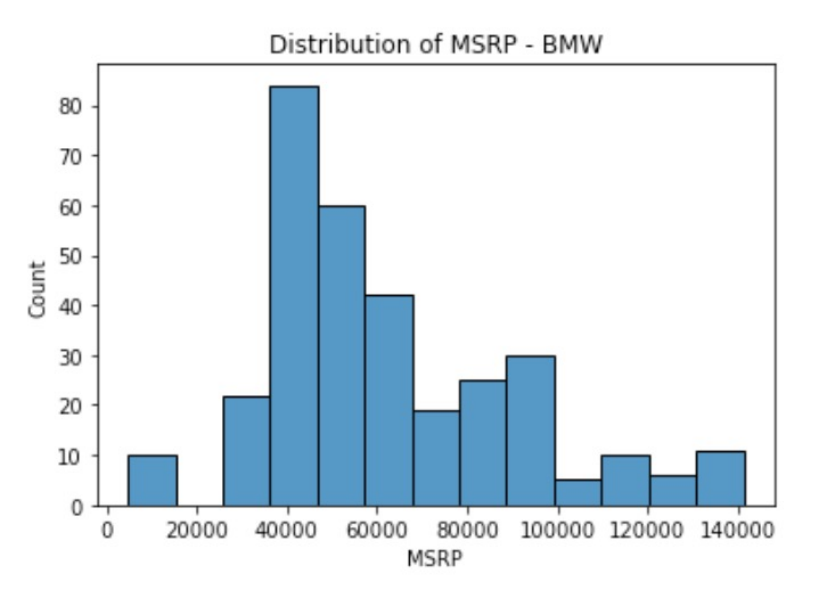


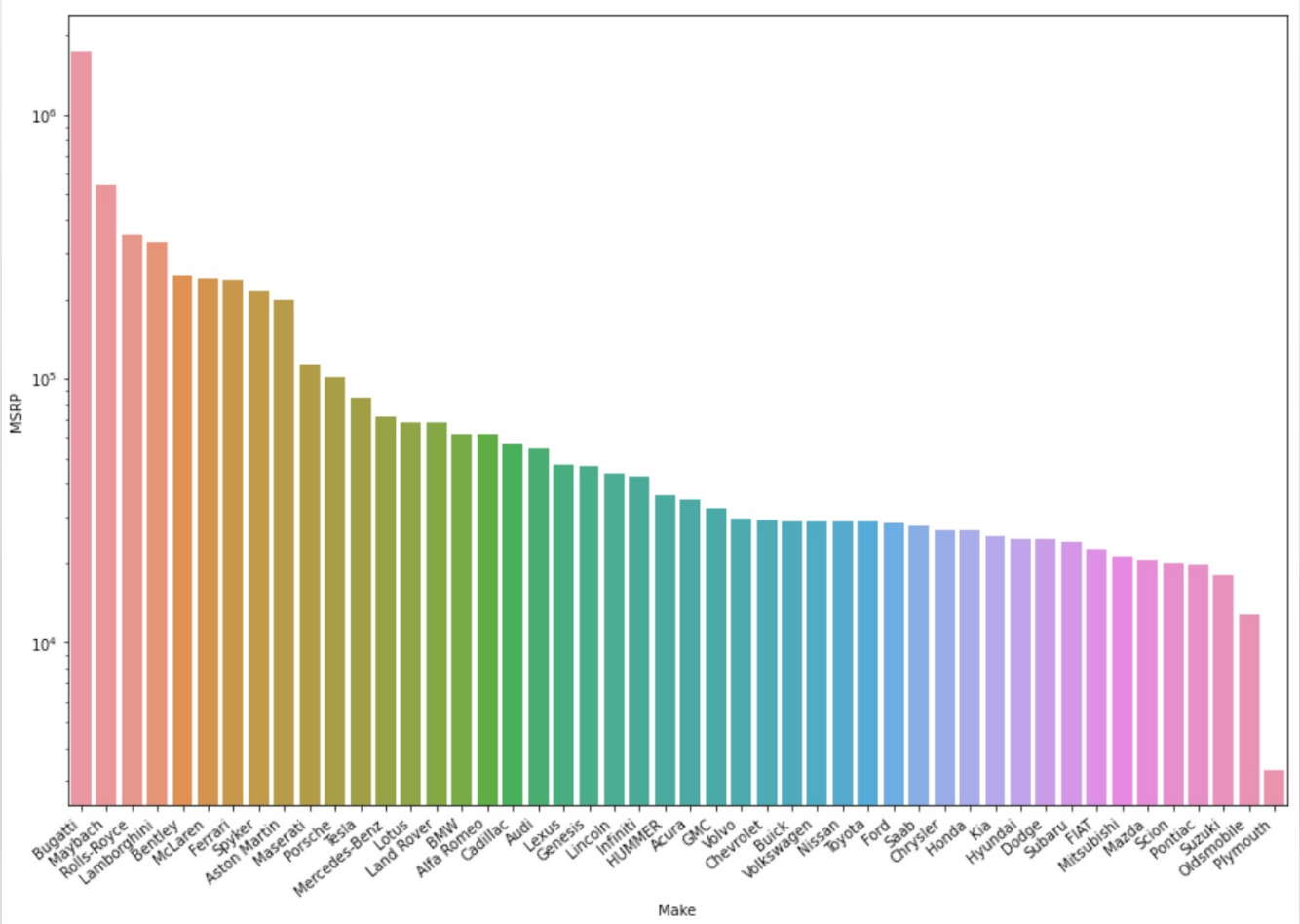
- The popularity variable is directly associated with brand
- The top 3 brands are:
  - Ford
  - BMW
  - Audi

Brands vs. MSRP

Some brands like Bugatti, Maybach, Rolls-Royce and Lamborghini focus more on the high performance cars which typically sold for more than $200,000

However, each brand typically has several models that covers a range of prices.



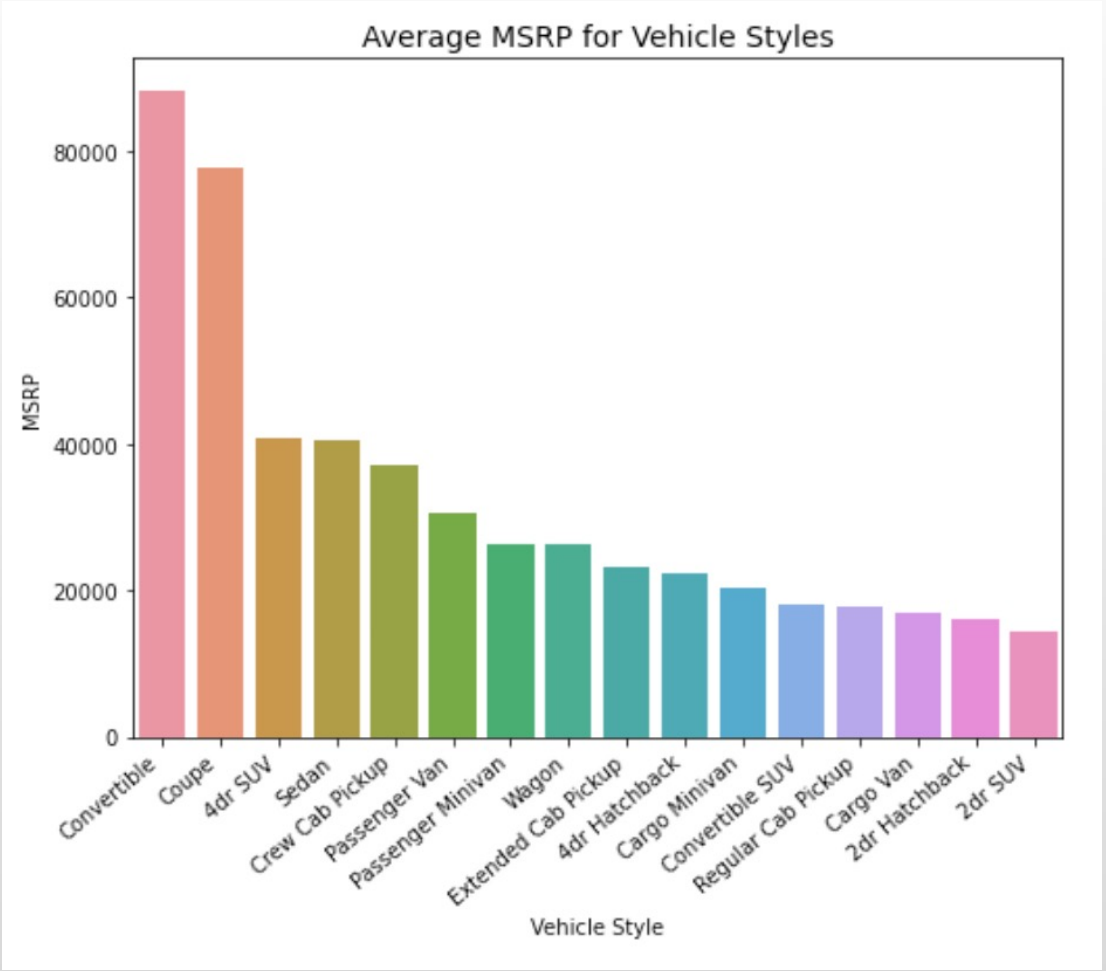


Styles vs. MSRP

Top two styles:

- Convertible
- Coupe

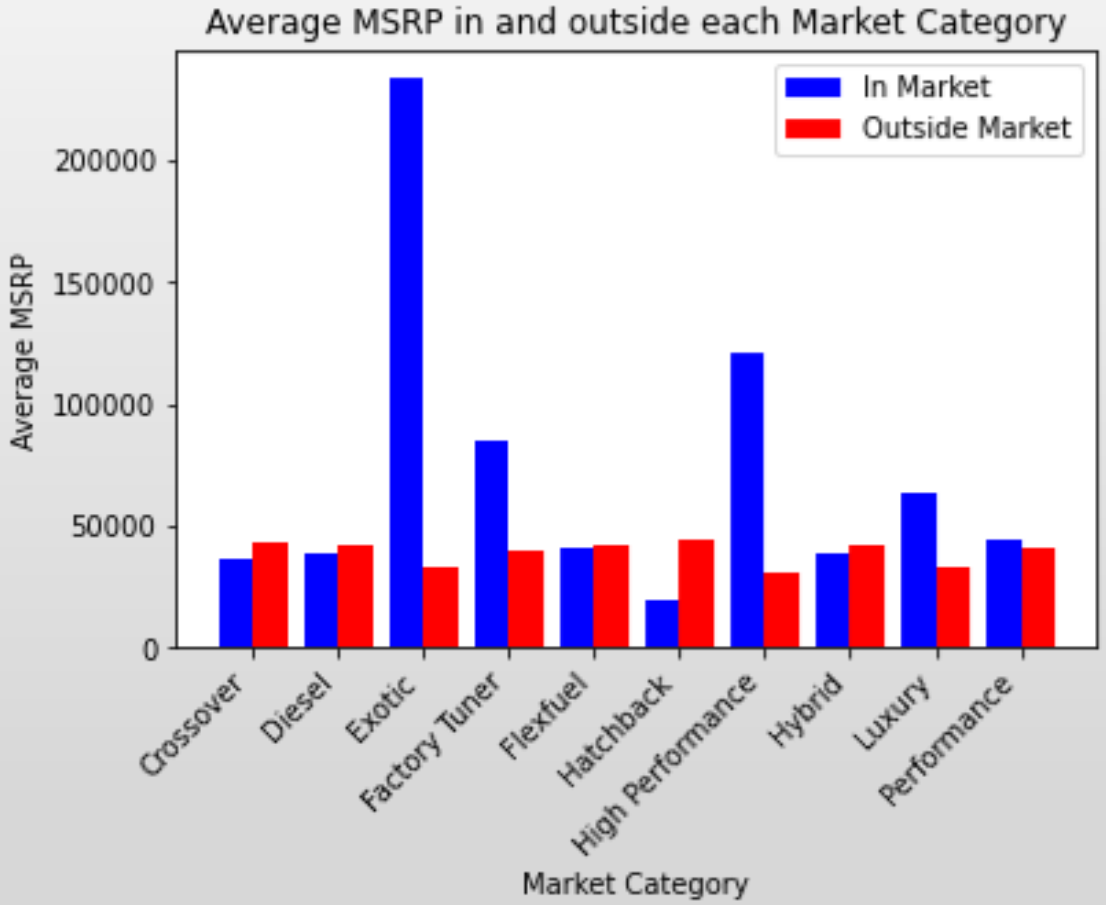
These two styles tend to show up on luxury cars.



Market Categories vs. MSRP

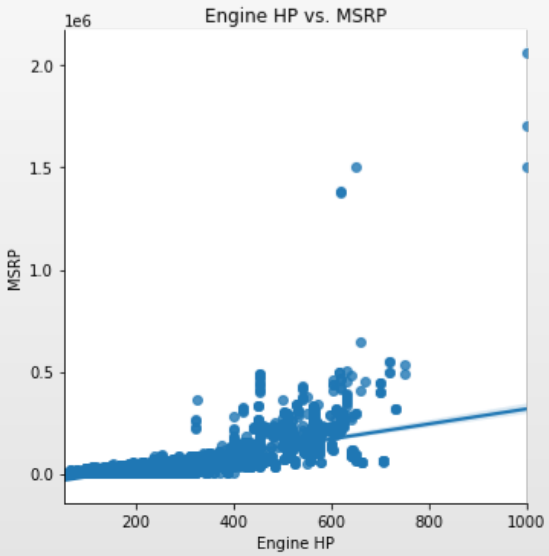
- Exotic
- High Performance
- Factory Tuner
- Luxury

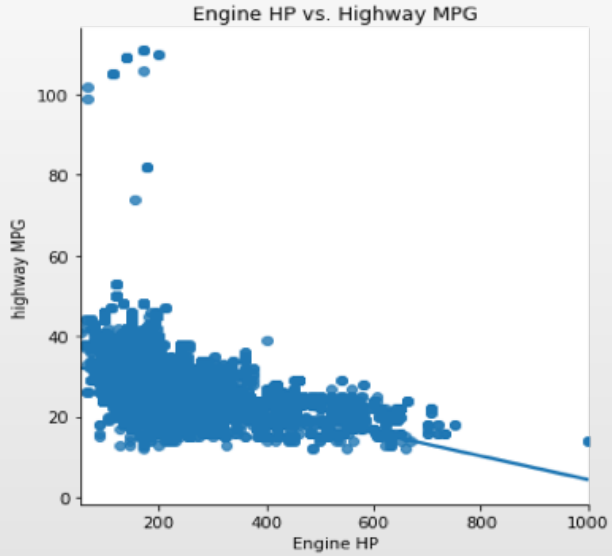
These 4 market categories obviously affect the price.

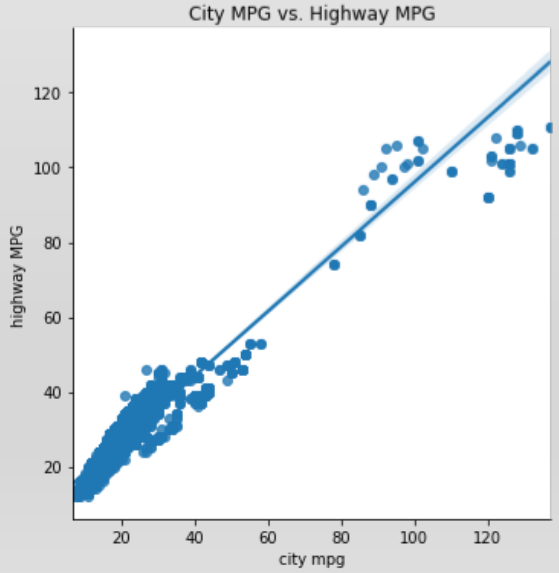


Numerical Features

- Engine Horse Power ↑
  MSRP ↑
- Engine Horse Power ↑
  Highway MPG ↓
- City MPG ↑
  Highway MPG ↑







Machine Learning Modeling

- Type: Supervised Learning
- Data Separation: 70% - Train Set; 30% - Test Set
- Two models were tested:
  - Linear regression model
  - Random forest regression model

Linear Regression Model

Pipeline Steps:

- Missing data imputation
- Scaler
- Feature Selection
  - With “SelectKBest” function
  - f_regression as soring method
- Cross validation (CV) for hyperparameter tuning:
  - 5 fold cv
  - Using Scikit-learn Grid Search method
- Pick K with highest accuracy

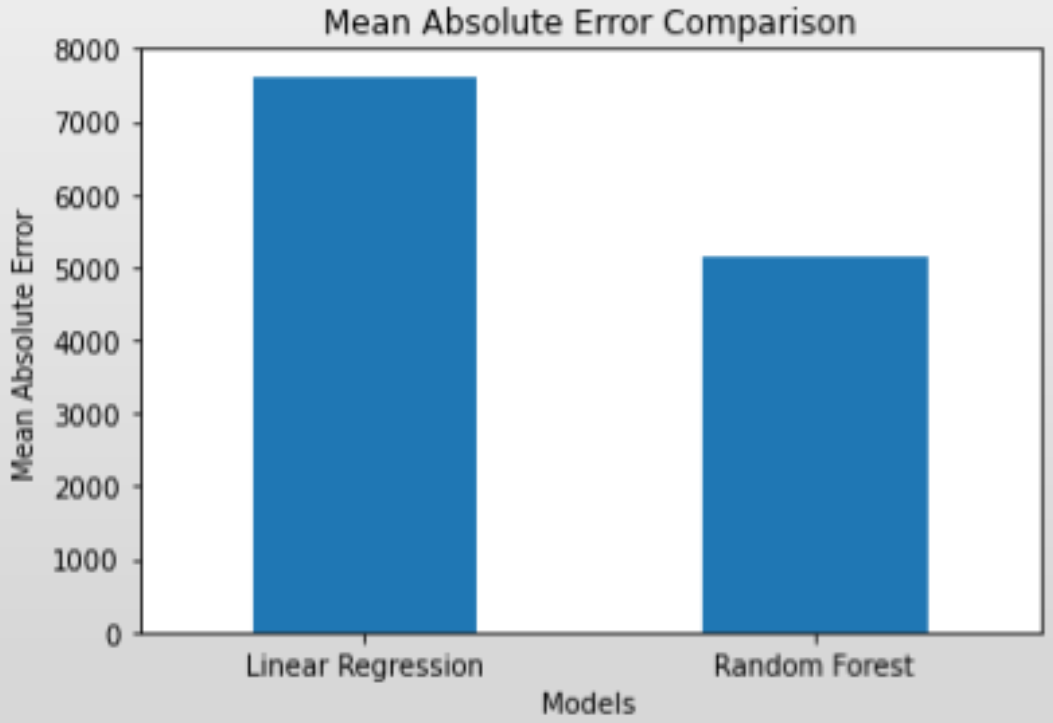
Random Forest Model

Pipeline Steps:

- Missing data imputation
- Scaler
- Cross validation (CV) for hyperparameter tuning:
  - 5 fold cv
  - Using Scikit-learn Grid Search method
- Choose best parameters that give highest accuracy:
  - Depth of tree
  - Number of estimator
  - Imputer method
  - Scale or not

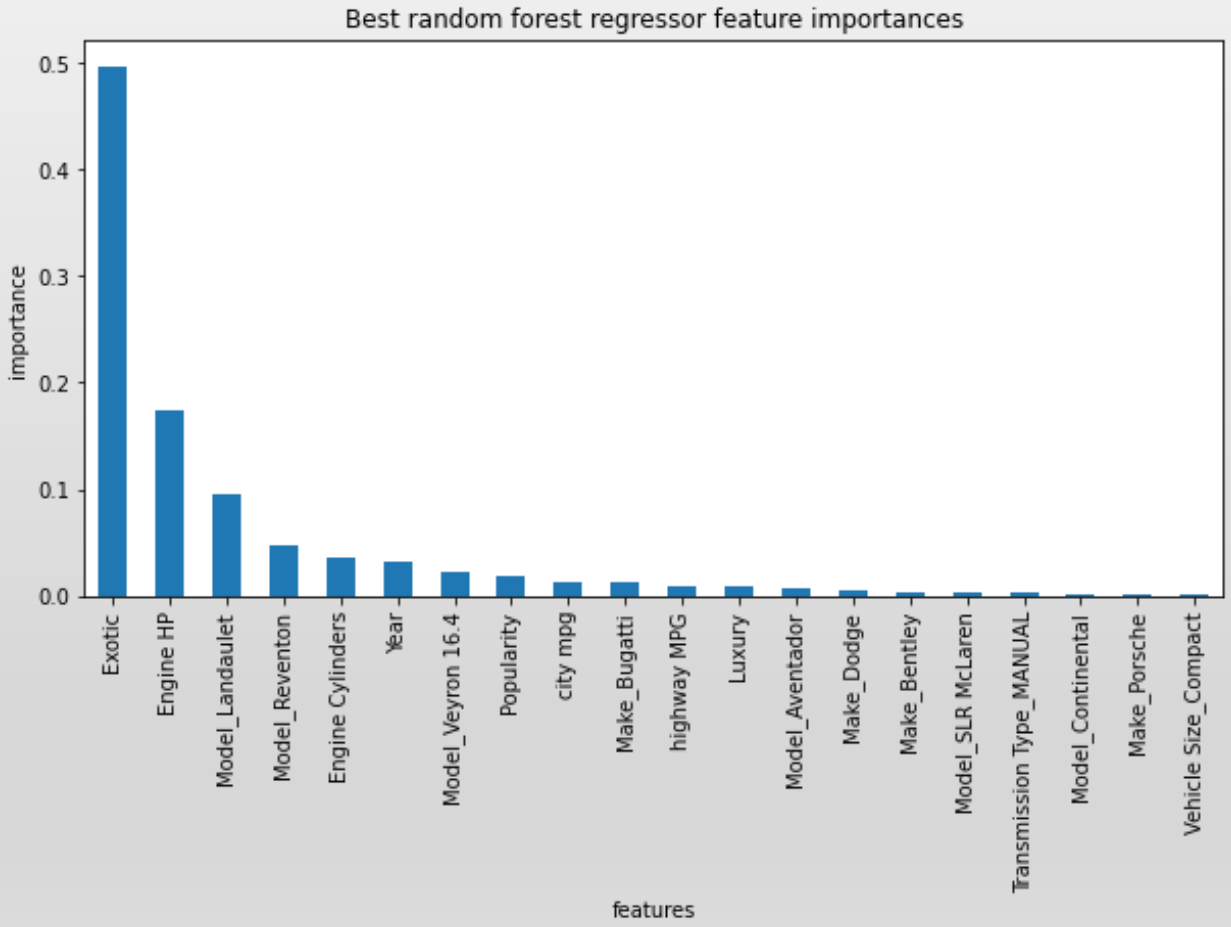
Model Performance

- Mean absolute error was used to exam the model performance
  - Linear regression: $7622
  - Random forest: $5133
- Random forest model not only has lower MAE, also higher R^2 score and higher accuracy



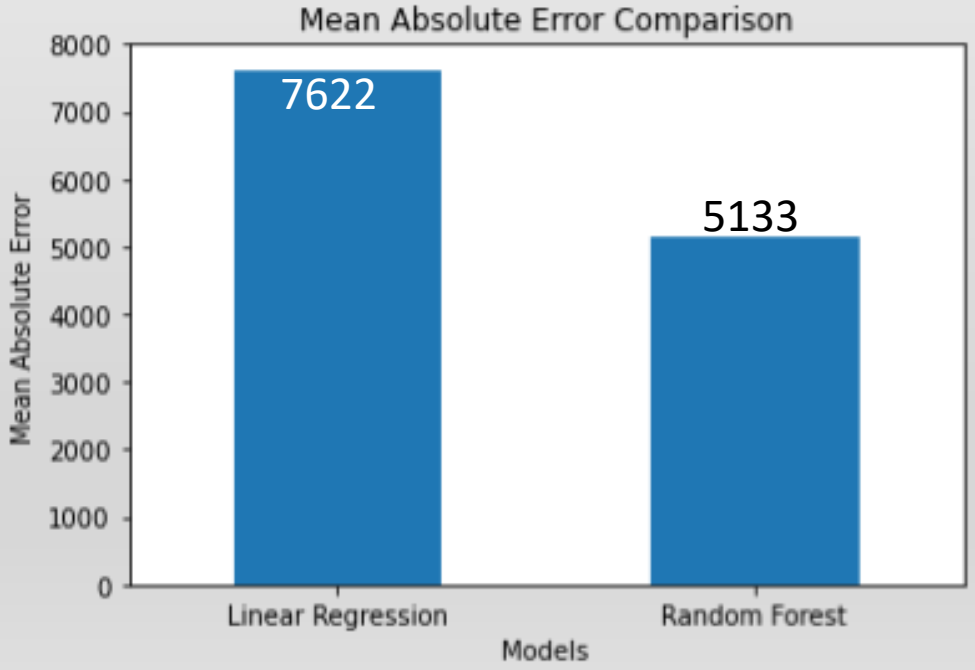
Important Features

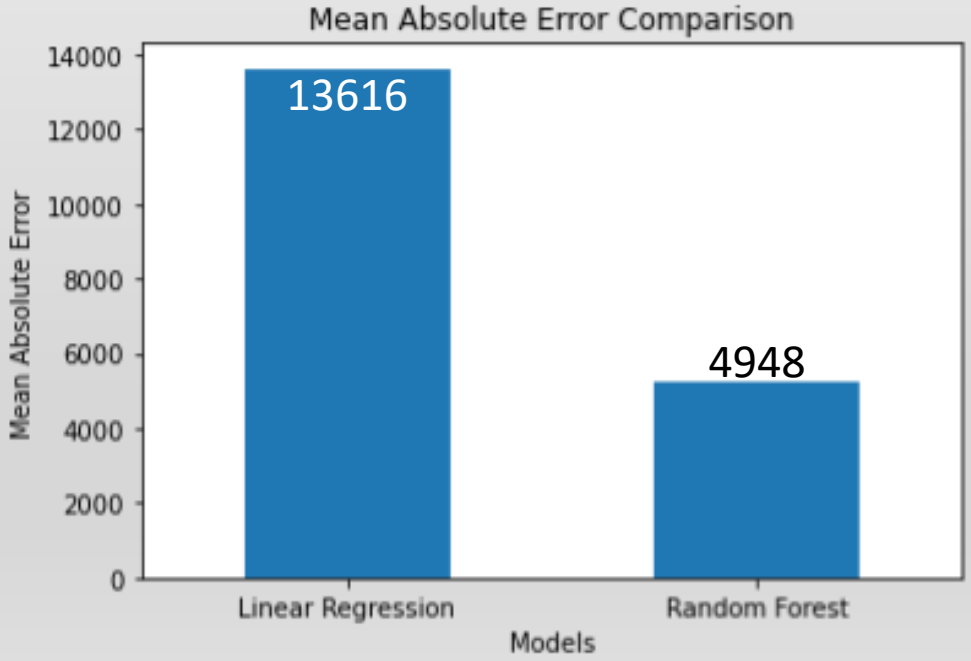
- Top 3 features which have biggest influence on price:
  - Market Category – Exotic
  - Engine Horsepower
  - Model – Landaulet
- Brands and Models are important features to the prediction model, but **what if we hide them all?**



With or Without Info about Brands and Models

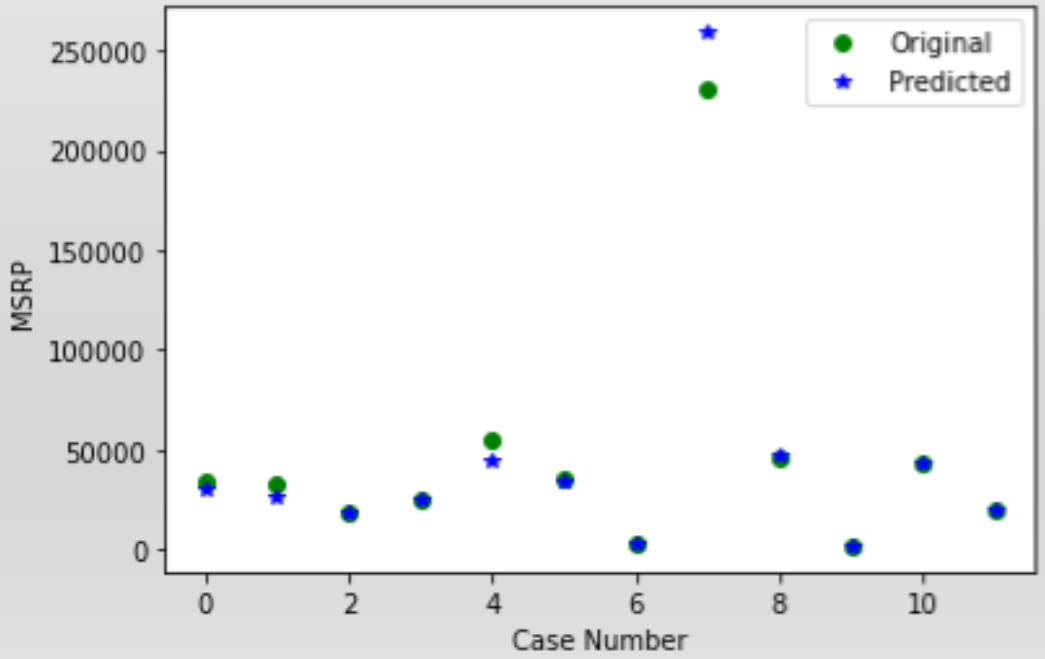
- The performance for random forest model is slight better without brands and models, which the performance for linear regression model got worst with twice the error without brands and models.

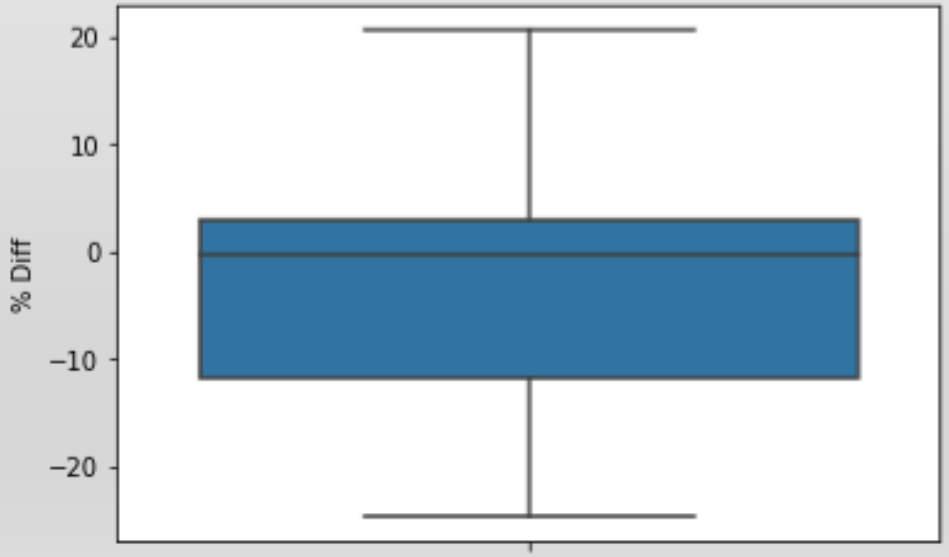




Prediction

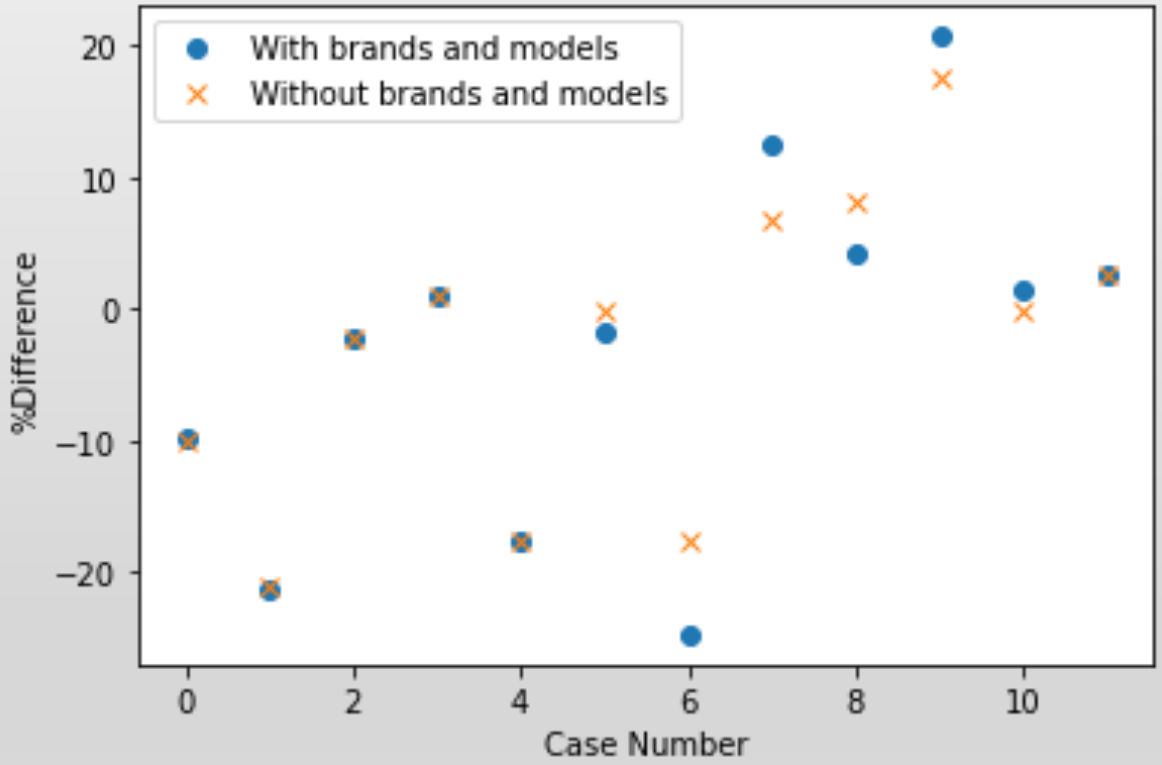
- % of difference between Predicted price and the actual MSRP ranges from -24.7% to 21.2% while 75% of the cases are within 12%.
- The model meets the initial target of 80% accuracy.





Prediction with/without brands and models

- The prediction was generally more accurate without information about brands and models.



Conclusions

- Random Forest Model provided the best results
- The average error of prediction is about $5000
- Future improvements:
  - Use more recent data

    (there are no data between 2017 to 2021)
  - More features

    (Material, Quality, Brand reputation, etc.)
  - Model flexibility

    (ability to fit the model with missing inputs, output a price range instead of a exact price)