Car MSRP Prediction with Multiple Linear Regression

CIS 4321

Beshoy Sadek, Sherleen Lee, Xiao Tong Gan, and Somar Dakak

May 14, 2021

Introduction

The problem for the manufacturers is to find a balance between what consumers want and pricing the particular vehicle

- Important to Know
    - which features and specs make the Manufacturer Suggested Retail Price
    - which features largely determine the price of the car
    - produce and sell cars that meet market demands for its target consumer group

The purpose of our group project is to figure out what features have an influence on a vehicle's Manufacturer Suggested Retail Price (MSRP) in the Car Features and MSRP dataset

- will be able to
    - know what features are most important to consumers by predicting an appropriate MSRP given a specific set of features
    - understand how those features affect the MSRP the most

Data Collection and Variable Description

Collecting Data:

- found our dataset on Kaggle
    - could not find the real source of the set
    - all the records were collected from Twitter and Edmunds

Dataset has 16 variables which are divided to 8 numerical and 8 categorical variables

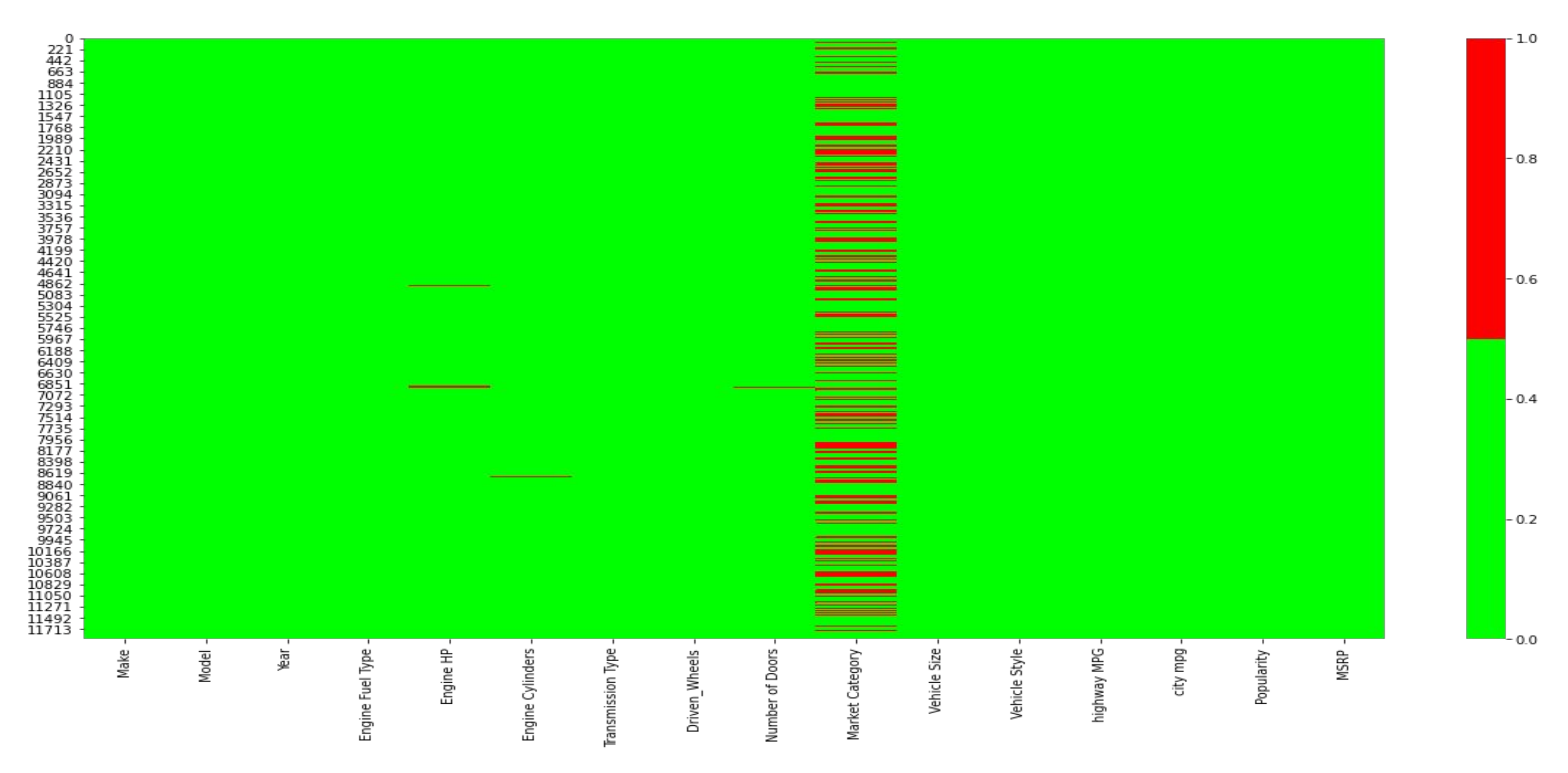
- numerical variables: Year, Engine HP, Engine Cylinders, Number of Doors, highway mpg, city mpg, Popularity, and MSRP
- categorical variables: Make, Model, Engine Fuel Type, Transmission type, Driven Wheels, Market Category, Vehicle Size, and Vehicle Style
    - categorical variables such as Transmission type, Vehicle Size, and Vehicle Style, are transformed to dummy variables

Data Collection and Variable Description

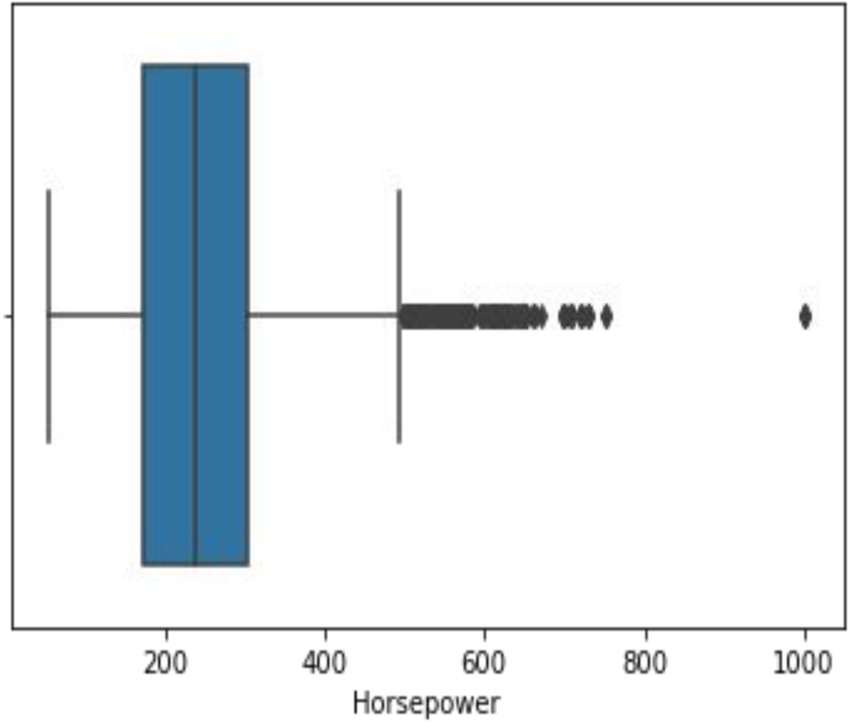
- Make : the brand of the vehicle
- Model : the model of the vehicle
- Year : the year of manufacture of the vehicle
- Engine fuel type : the type of you that the vehicle operates on
- Engine HP : the engine horsepower
- Engine cylinders :the number of cylinders in the combustion engine
- Transmission type: what kind of transmission the vehicle has
- Driven_Wheels : the amount of wheels that receive power directly from the engine
- Number of doors : how many doors the vehicle has

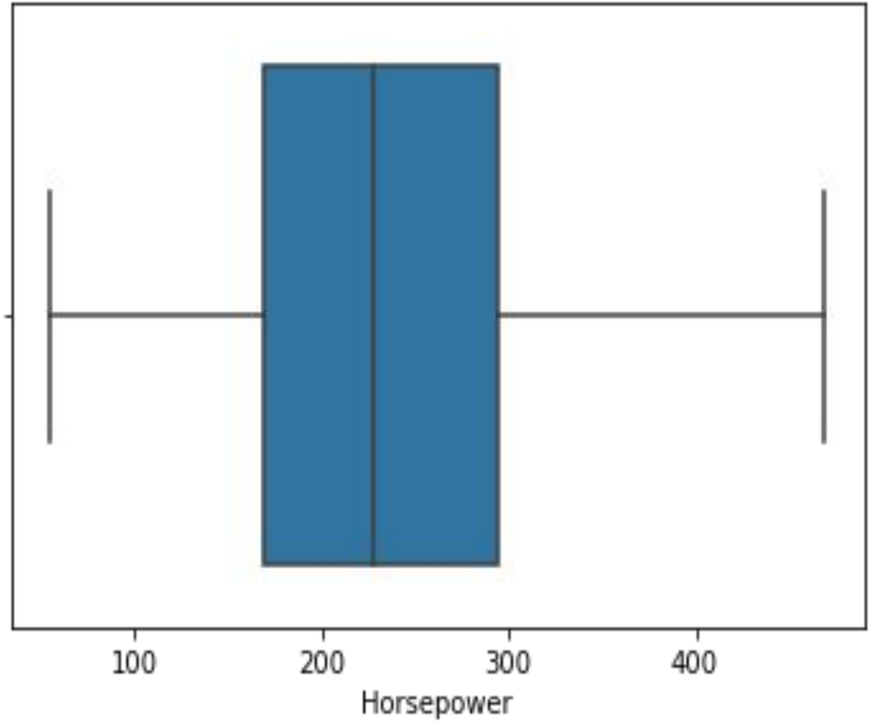
- Market category : the market category to which the vehicle belongs
- Vehicle size : the size of the vehicle based on its volume capacity
- Vehicle style : the style of the vehicle based on its physical shape
- Highway MPG : the number of miles the vehicle drives per gallon at highway speed
- City MPG : the number of miles the vehicle drives per gallon at a city speed
- Popularity : how popular the car is among people
- MSRP :the manufacturer’s suggested retail price of a vehicle

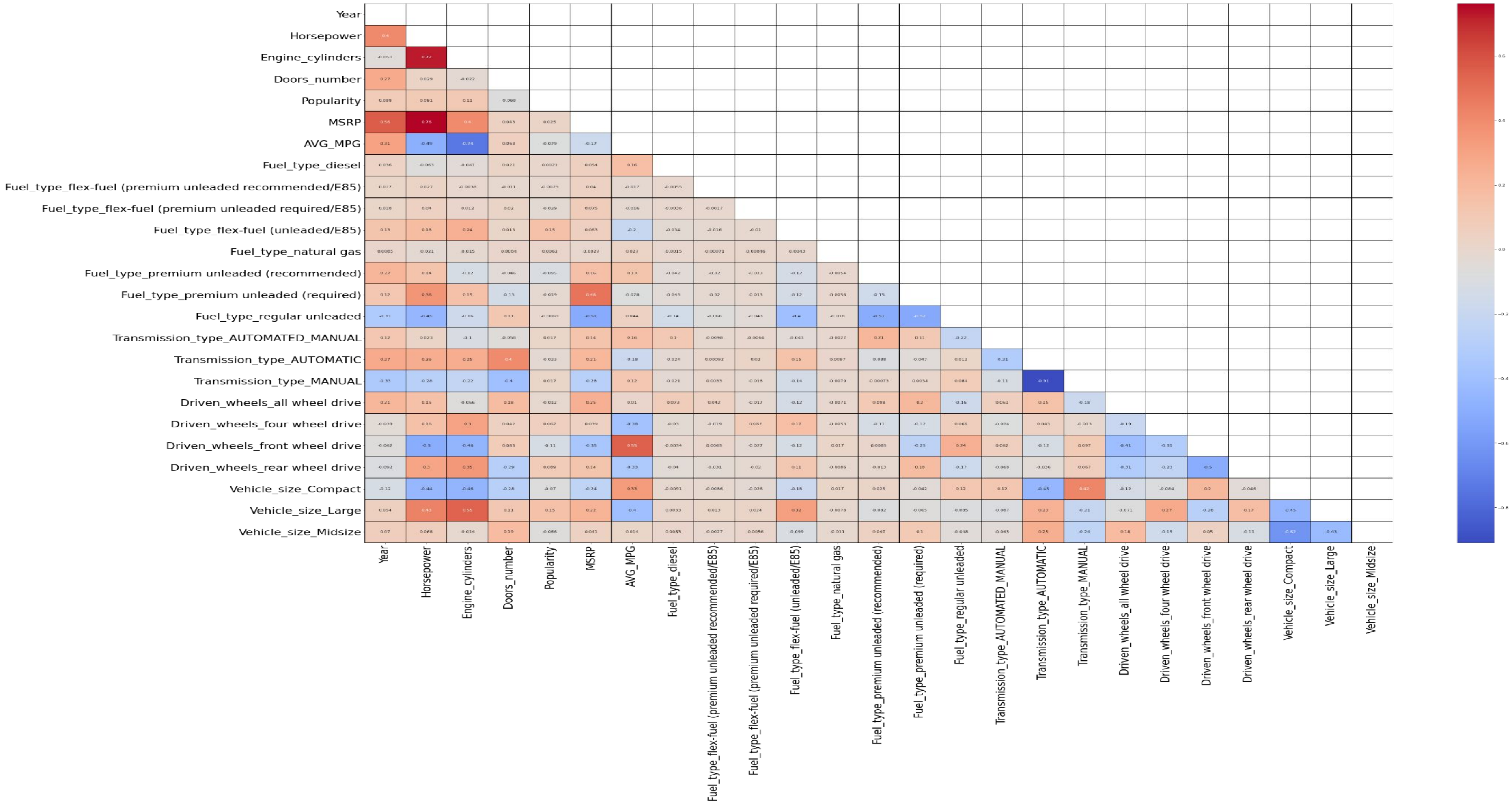
Descriptive Analysis

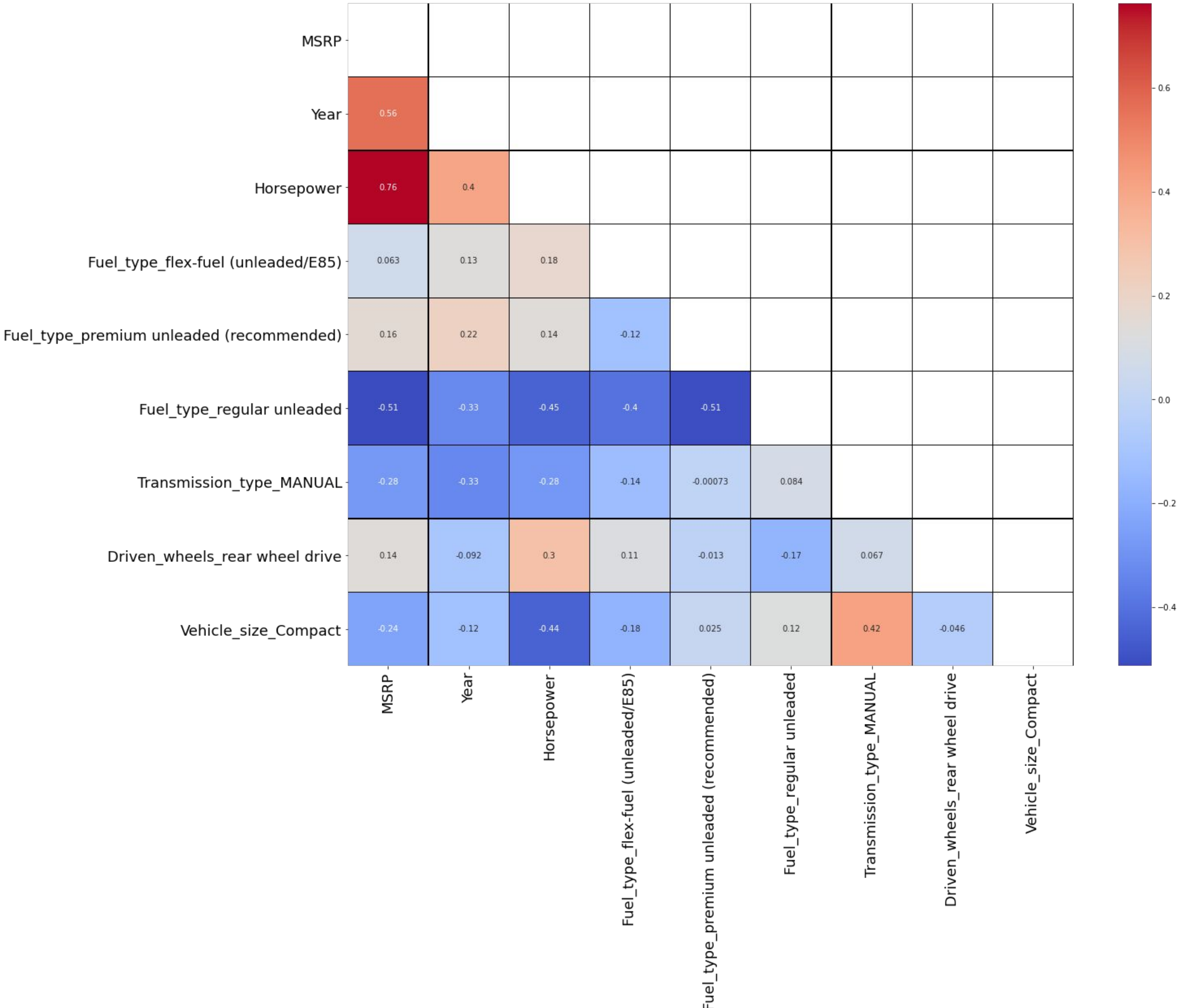


Descriptive Analysis






Year
Horsepower
0.4
Engine_cylinders
-0.051
0.72
Doors_number
0.27
0.029
-0.022
Popularity
0.088
0.091
0.11
-0.068
MSRP
0.56
0.76
0.4
0.043
0.025
AVG_MPG
0.31
-0.49
-0.74
0.063
-0.079
-0.17
Fuel_type_diesel
0.036
-0.063
-0.041
0.021
0.0021
0.054
0.16
Fuel_type_flex-fuel (premium unleaded recommended/E85)
0.017
0.027
-0.0038
-0.011
-0.0079
0.04
-0.017
-0.0055
Fuel_type_flex-fuel (premium unleaded required/E85)
0.018
0.04
0.012
0.02
-0.029
0.075
-0.016
-0.0036
-0.0017
Fuel_type_flex-fuel (unleaded/E85)
0.13
0.18
0.24
0.013
0.15
0.063
-0.2
-0.034
-0.016
-0.01
Fuel_type_natural gas
0.0085
-0.021
-0.015
0.0084
0.0062
-0.0027
0.027
-0.0015
-0.00071
-0.00046
-0.0043
Fuel_type_premium unleaded (recommended)
0.22
0.14
-0.12
-0.046
-0.095
0.16
0.13
-0.042
-0.02
-0.013
-0.12
-0.0054
Fuel_type_premium unleaded (required)
0.12
0.36
0.15
-0.13
-0.019
0.48
-0.078
-0.043
-0.02
-0.013
-0.12
-0.0056
-0.15
Fuel_type_regular unleaded
-0.33
-0.45
-0.16
0.11
-0.0069
-0.51
0.044
-0.14
-0.066
-0.043
-0.4
-0.018
-0.51
-0.52
Transmission_type_AUTOMATED_MANUAL
0.12
0.023
-0.1
-0.058
0.017
0.14
0.16
0.1
-0.0098
-0.0064
-0.043
-0.0027
0.21
0.11
-0.22
Transmission_type_AUTOMATIC
0.27
0.26
0.25
0.4
-0.023
0.21
-0.18
-0.024
0.00092
0.02
0.15
0.0087
-0.088
-0.047
0.012
-0.31
Transmission_type_MANUAL
-0.33
-0.28
-0.22
-0.4
0.017
-0.28
0.12
-0.021
0.0033
-0.018
-0.14
-0.0079
-0.00073
0.0034
0.084
-0.11
-0.91
Driven_wheels_all wheel drive
0.21
0.15
-0.066
0.18
-0.012
0.25
0.01
0.073
0.042
-0.017
-0.12
-0.0071
0.098
0.2
-0.16
0.061
0.15
-0.18
Driven_wheels_four wheel drive
-0.039
0.16
0.3
0.042
0.062
0.039
-0.38
-0.03
-0.019
0.087
0.17
-0.0053
-0.11
-0.12
0.066
-0.074
0.043
-0.013
-0.19
Driven_wheels_front wheel drive
-0.062
-0.5
-0.46
0.083
-0.11
-0.35
0.55
-0.0034
0.0065
-0.027
-0.12
0.017
0.0085
-0.25
0.24
0.062
-0.12
0.097
-0.41
-0.31
Driven_wheels_rear wheel drive
-0.092
0.3
0.35
-0.29
0.089
0.14
-0.33
-0.04
-0.031
-0.02
0.11
-0.0086
-0.013
0.18
-0.17
-0.068
-0.036
0.067
-0.31
-0.23
-0.5
Vehicle_size_Compact
-0.12
-0.44
-0.46
-0.28
-0.07
-0.24
0.33
-0.0091
-0.0086
-0.026
-0.18
0.017
0.025
-0.042
0.12
0.12
-0.45
0.42
-0.12
-0.084
0.2
-0.046
Vehicle_size_Large
0.054
0.43
0.55
0.11
0.15
0.22
-0.4
0.0033
0.013
0.024
0.32
-0.0078
-0.082
-0.065
-0.085
-0.087
0.23
-0.21
-0.071
0.27
-0.28
0.17
-0.45
Vehicle_size_Midsize
0.07
0.068
-0.014
0.19
-0.066
0.041
0.014
0.0063
-0.0027
0.0056
-0.099
-0.011
0.047
0.1
-0.048
-0.045
0.25
-0.24
0.18
-0.15
0.05
-0.11
-0.62
-0.43
Year
Horsepower
Engine_cylinders
Doors_number
Popularity
MSRP
AVG_MPG
Fuel_type_diesel
Fuel_type_flex-fuel (premium unleaded recommended/E85)
Fuel_type_flex-fuel (premium unleaded required/E85)
Fuel_type_flex-fuel (unleaded/E85)
Fuel_type_natural gas
Fuel_type_premium unleaded (recommended)
Fuel_type_premium unleaded (required)
Fuel_type_regular unleaded
Transmission_type_AUTOMATED_MANUAL
Transmission_type_AUTOMATIC
Transmission_type_MANUAL
Driven_wheels_all wheel drive
Driven_wheels_four wheel drive
Driven_wheels_front wheel drive
Driven_wheels_rear wheel drive
Vehicle_size_Compact
Vehicle_size_Large
Vehicle_size_Midsize
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8

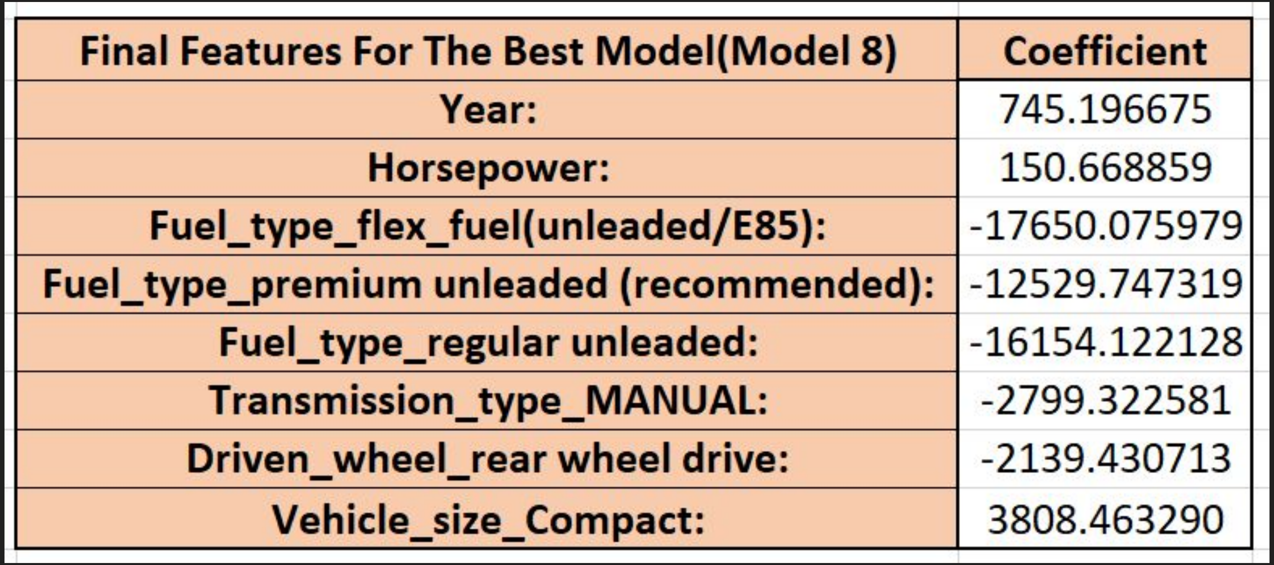

MSRP
Year
Horsepower
Fuel_type_flex-fuel (unleaded/E85)
Fuel_type_premium unleaded (recommended)
Fuel_type_regular unleaded
Transmission_type_MANUAL
Driven_wheels_rear wheel drive
Vehicle_size_Compact
0.56
0.76
0.4
0.063
0.13
0.18
0.16
0.22
0.14
-0.12
-0.51
-0.33
-0.45
-0.4
-0.51
-0.28
-0.33
-0.28
-0.14
-0.00073
0.084
0.14
-0.092
0.3
0.11
-0.013
-0.17
0.067
-0.24
-0.12
-0.44
-0.18
0.025
0.12
0.42
-0.046
MSRP
Year
Horsepower
Fuel_type_flex-fuel (unleaded/E85)
Fuel_type_premium unleaded (recommended)
Fuel_type_regular unleaded
Transmission_type_MANUAL
Driven_wheels_rear wheel drive
Vehicle_size_Compact
0.6
0.4
0.2
0.0
−0.2
−0.4

Data Analysis

- Use multiple linear regression
  - to predict one variable based on the known information about other variables, and
  - examine how multiple independent variables are related to one dependent variable
- Used recursive model elimination (RFE) to compare models with different sizes and selected Model 8* with highest adjusted R^2 of 0.75
- Compared the performance of the model using the training and test sets

Model 8* (Year, Horsepower, Fuel_type_flex-fuel (unleaded/E85), Fuel_type_premium unleaded (recommended), Fuel_type_regular unleaded, Transmission_type_MANUAL, Driven_wheels_rear wheel drive, Vehicle_size_Compact)

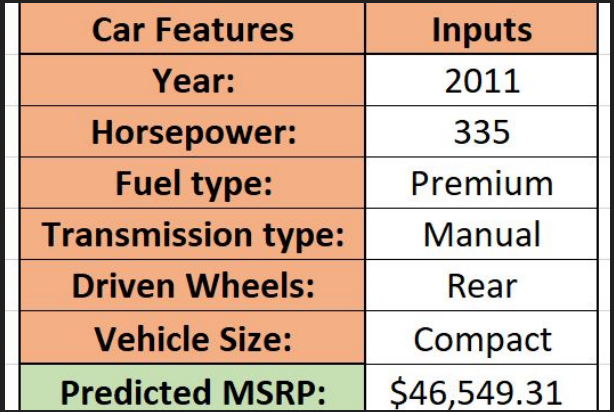
Model 8



Final Features For The Best Model(Model 8)	Coefficient
Year:	745.196675
Horsepower:	150.668859
Fuel_type_flex_fuel(unleaded/E85):	-17650.075979
Fuel_type_premium unleaded (recommended):	-12529.747319
Fuel_type_regular unleaded:	-16154.122128
Transmission_type_MANUAL:	-2799.322581
Driven_wheel_rear wheel drive:	-2139.430713
Vehicle_size_Compact:	3808.463290

Summary of Findings

- Result of the analysis: predict the MSRP of a car with a 75% accuracy
- We were able to predict a car’s MSRP with a given set of features
- We also concluded that the RFE is not the best model to predict a car’s MSRP based on the 75% accuracy that the RFE method yielded.
  - We were able to accurately predict a car's msrp after plugging in values for the features listed in Model 8



Car Features	Inputs
Year:	2011
Horsepower:	335
Fuel type:	Premium
Transmission type:	Manual
Driven Wheels:	Rear
Vehicle Size:	Compact
Predicted MSRP:	$46,549.31

Implications

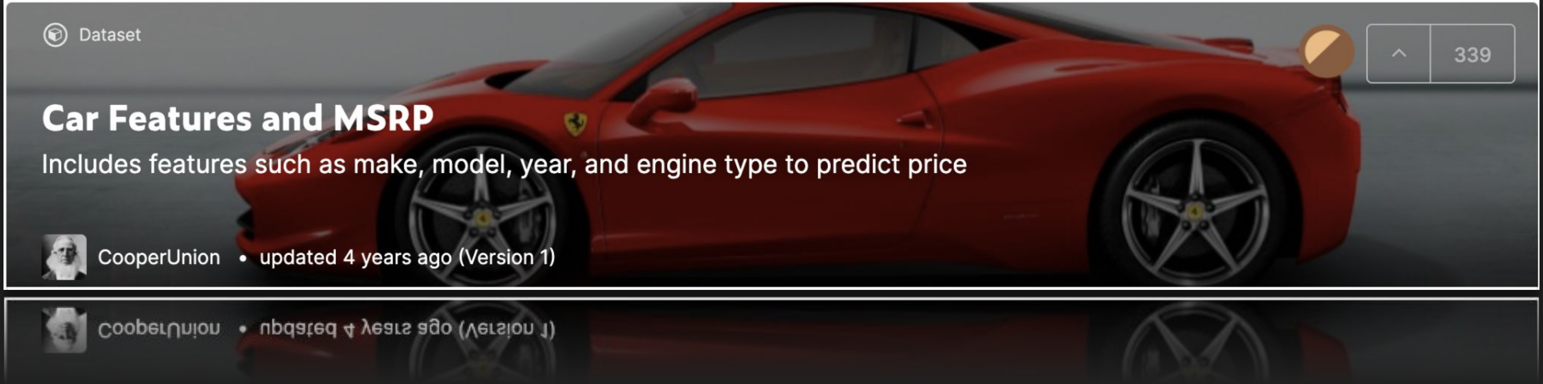
- At 75% accuracy in predicting MSRP, manufacturers can:
  - Adjust features and specs according to the preferred MSRP,
  - Budget early on,
  - Decide on how to market the car, and
  - Have a better understanding and control of the factors contributing to the MSRP

Limitations

- Lack of numerical variables
- Unknown values which needed to be dropped individually
- Irrelevant variables such as market category, vehicle style, and popularity.

Future steps

- Improve our results with a dataset that has more complete data and more relevant variables
- Use a different model for our prediction



Car Features and MSRP

Includes features such as make, model, year, and engine type to predict price

CooperUnion • updated 4 years ago (Version 1)

Classification

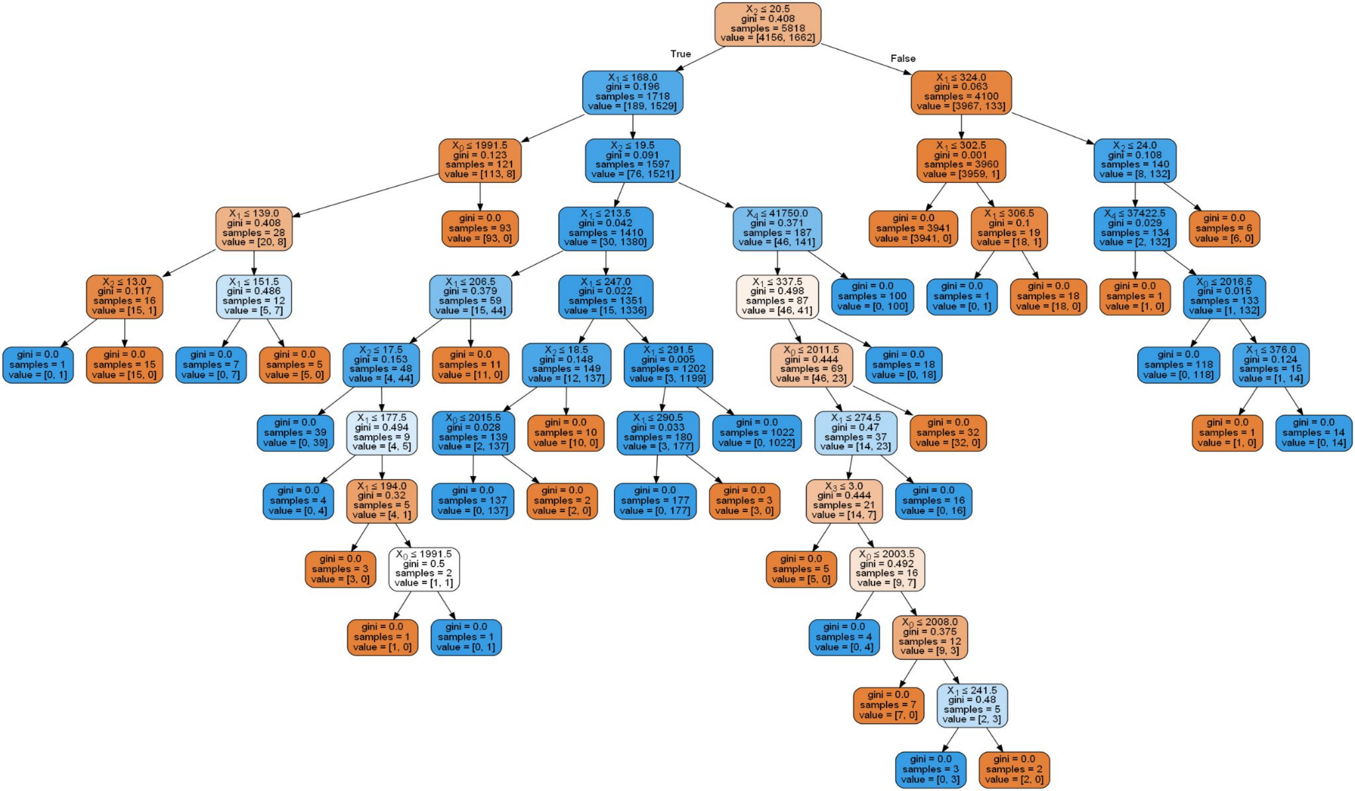
The problem: Filling a missing Engine Cylinder value on a car shopping website.

The approach: Using 3 different Binary Classification methods to solve this problem.

Used Methods: Decision Tree, Random Forest, and Naive Bayes

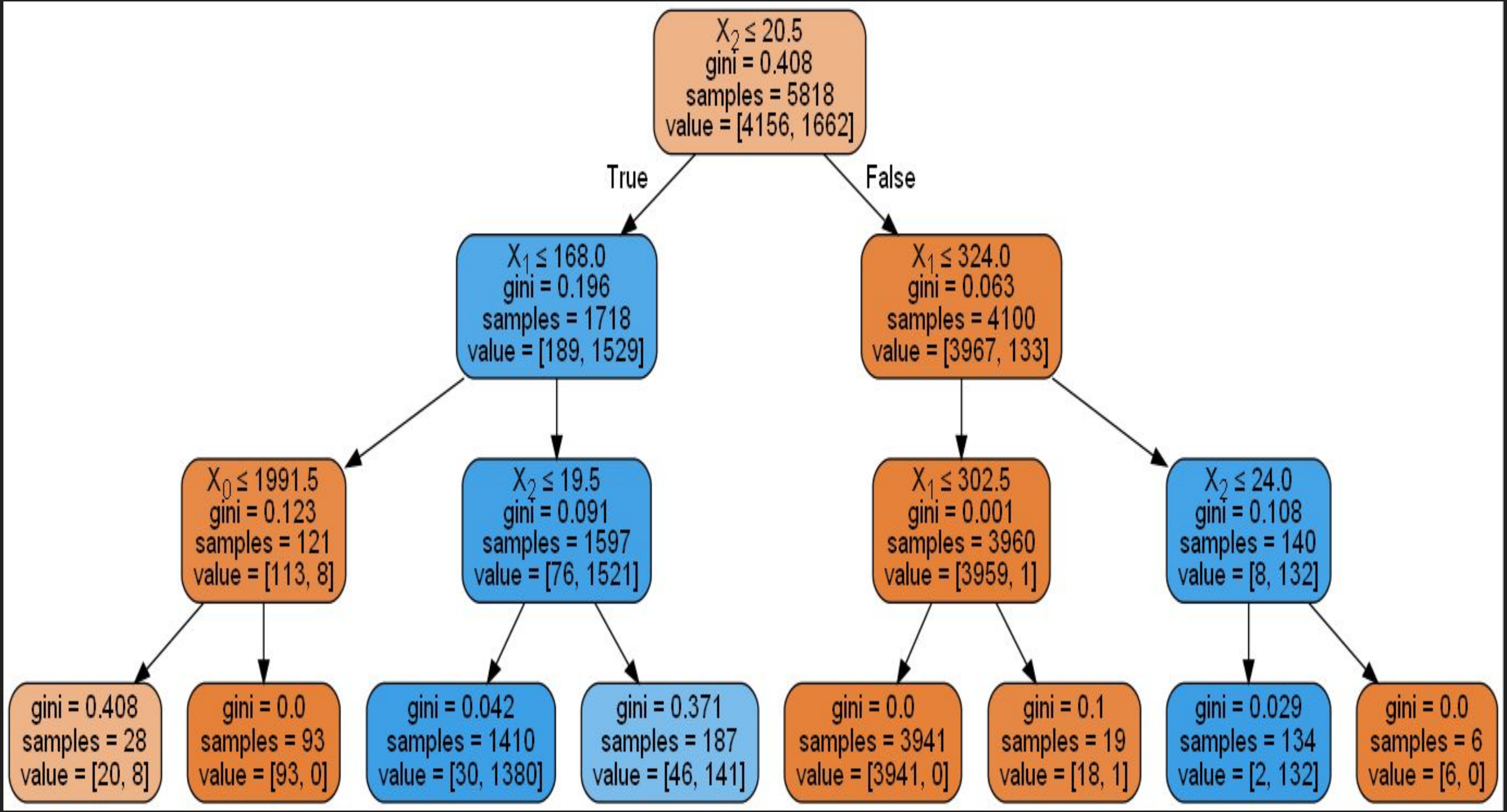
Results: All Methods yielded impressive Classification abilities of 97% and above accuracy scores.

Analysis and Results

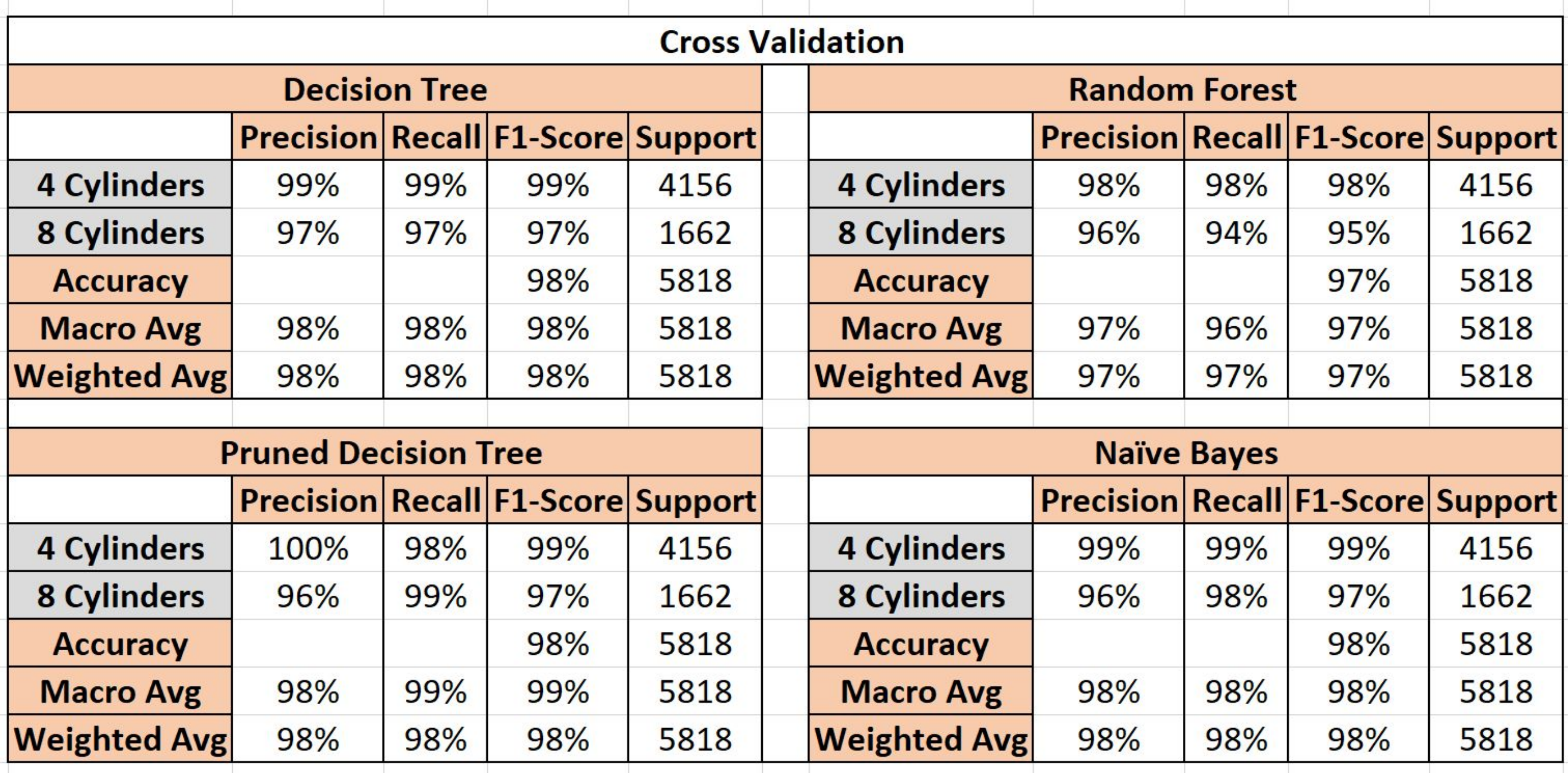


$X_2 \leq 20.5$ gini = 0.408 samples = 5818 value = [4156, 1662]
True
False
$X_1 \leq 168.0$ gini = 0.196 samples = 1718 value = [189, 1529]
$X_1 \leq 324.0$ gini = 0.063 samples = 4100 value = [3967, 133]
$X_0 \leq 1991.5$ gini = 0.123 samples = 121 value = [113, 8]
$X_2 \leq 19.5$ gini = 0.091 samples = 1597 value = [76, 1521]
$X_1 \leq 302.5$ gini = 0.001 samples = 3960 value = [3959, 1]
$X_2 \leq 24.0$ gini = 0.108 samples = 140 value = [8, 132]
$X_1 \leq 139.0$ gini = 0.408 samples = 28 value = [20, 8]
gini = 0.0 samples = 93 value = [93, 0]
$X_1 \leq 213.5$ gini = 0.042 samples = 1410 value = [30, 1380]
$X_4 \leq 41750.0$ gini = 0.371 samples = 187 value = [46, 141]
gini = 0.0 samples = 3941 value = [3941, 0]
$X_1 \leq 306.5$ gini = 0.1 samples = 19 value = [18, 1]
$X_4 \leq 37422.5$ gini = 0.029 samples = 134 value = [2, 132]
gini = 0.0 samples = 6 value = [6, 0]
$X_2 \leq 13.0$ gini = 0.117 samples = 16 value = [15, 1]
$X_1 \leq 151.5$ gini = 0.486 samples = 12 value = [5, 7]
$X_1 \leq 206.5$ gini = 0.379 samples = 59 value = [15, 44]
$X_1 \leq 247.0$ gini = 0.022 samples = 1351 value = [15, 1336]
$X_1 \leq 337.5$ gini = 0.498 samples = 87 value = [46, 41]
gini = 0.0 samples = 100 value = [0, 100]
gini = 0.0 samples = 1 value = [0, 1]
gini = 0.0 samples = 18 value = [18, 0]
gini = 0.0 samples = 1 value = [1, 0]
$X_0 \leq 2016.5$ gini = 0.015 samples = 133 value = [1, 132]
gini = 0.0 samples = 1 value = [0, 1]
gini = 0.0 samples = 15 value = [15, 0]
gini = 0.0 samples = 7 value = [0, 7]
gini = 0.0 samples = 5 value = [5, 0]
$X_2 \leq 17.5$ gini = 0.153 samples = 48 value = [4, 44]
gini = 0.0 samples = 11 value = [11, 0]
$X_2 \leq 18.5$ gini = 0.148 samples = 149 value = [12, 137]
$X_1 \leq 291.5$ gini = 0.005 samples = 1202 value = [3, 1199]
$X_0 \leq 2011.5$ gini = 0.444 samples = 69 value = [46, 23]
gini = 0.0 samples = 18 value = [0, 18]
gini = 0.0 samples = 118 value = [0, 118]
$X_1 \leq 376.0$ gini = 0.124 samples = 15 value = [1, 14]
gini = 0.0 samples = 39 value = [0, 39]
$X_1 \leq 177.5$ gini = 0.494 samples = 9 value = [4, 5]
$X_0 \leq 2015.5$ gini = 0.028 samples = 139 value = [2, 137]
gini = 0.0 samples = 10 value = [10, 0]
$X_1 \leq 290.5$ gini = 0.033 samples = 180 value = [3, 177]
gini = 0.0 samples = 1022 value = [0, 1022]
$X_1 \leq 274.5$ gini = 0.47 samples = 37 value = [14, 23]
gini = 0.0 samples = 32 value = [32, 0]
gini = 0.0 samples = 1 value = [1, 0]
gini = 0.0 samples = 14 value = [0, 14]
gini = 0.0 samples = 4 value = [0, 4]
$X_1 \leq 194.0$ gini = 0.32 samples = 5 value = [4, 1]
gini = 0.0 samples = 137 value = [0, 137]
gini = 0.0 samples = 2 value = [2, 0]
gini = 0.0 samples = 177 value = [0, 177]
gini = 0.0 samples = 3 value = [3, 0]
$X_3 \leq 3.0$ gini = 0.444 samples = 21 value = [14, 7]
gini = 0.0 samples = 16 value = [0, 16]
gini = 0.0 samples = 3 value = [3, 0]
$X_0 \leq 1991.5$ gini = 0.5 samples = 2 value = [1, 1]
gini = 0.0 samples = 5 value = [5, 0]
$X_0 \leq 2003.5$ gini = 0.492 samples = 16 value = [9, 7]
gini = 0.0 samples = 1 value = [1, 0]
gini = 0.0 samples = 1 value = [0, 1]
gini = 0.0 samples = 4 value = [0, 4]
$X_0 \leq 2008.0$ gini = 0.375 samples = 12 value = [9, 3]
gini = 0.0 samples = 7 value = [7, 0]
$X_1 \leq 241.5$ gini = 0.48 samples = 5 value = [2, 3]
gini = 0.0 samples = 3 value = [0, 3]
gini = 0.0 samples = 2 value = [2, 0]

Full Decision Tree



Pruned Decision Tree



Cross Validation

Decision Tree

	Precision	Recall	F1-Score	Support
4 Cylinders	99%	99%	99%	4156
8 Cylinders	97%	97%	97%	1662
Accuracy			98%	5818
Macro Avg	98%	98%	98%	5818
Weighted Avg	98%	98%	98%	5818

Random Forest

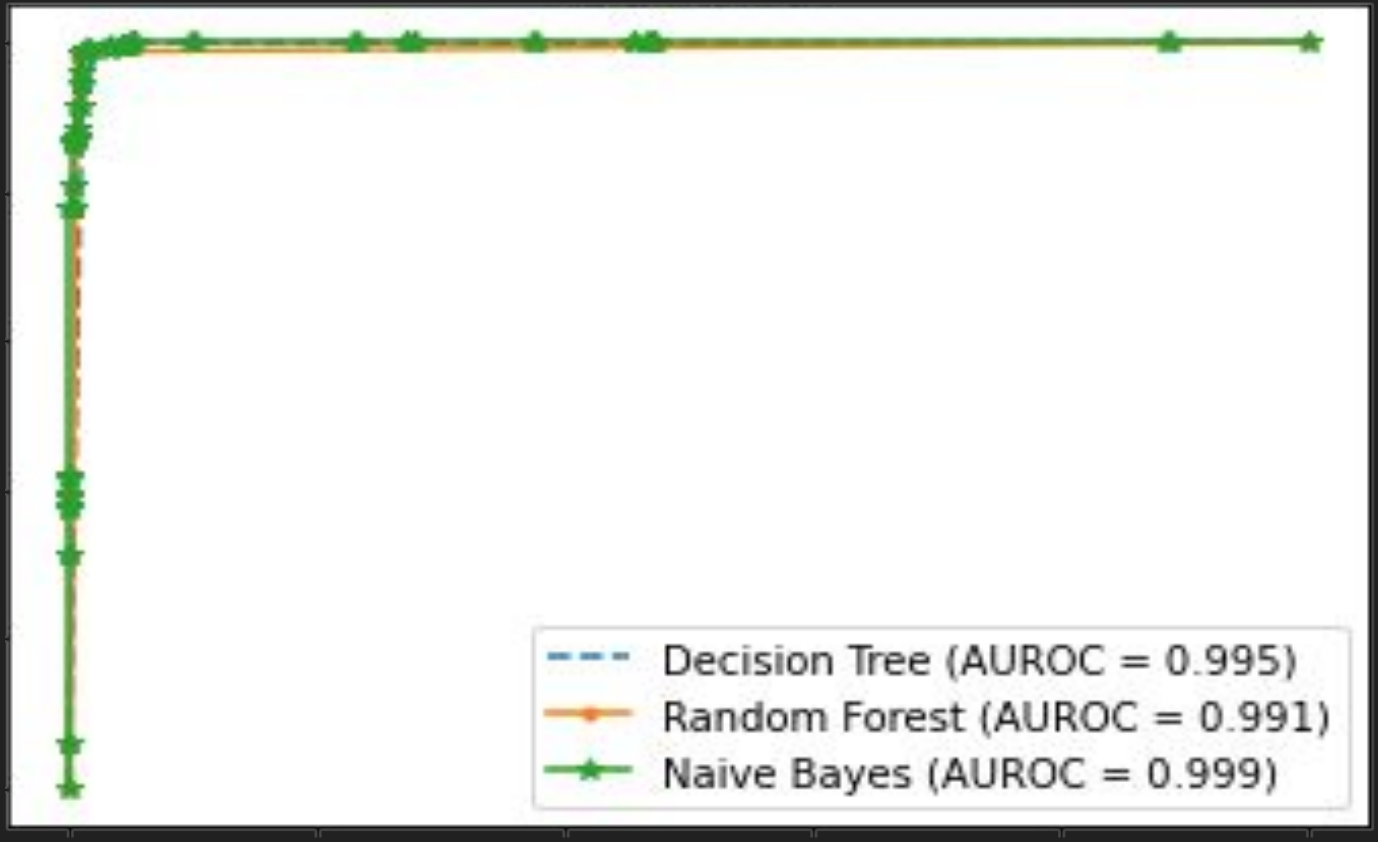
	Precision	Recall	F1-Score	Support
4 Cylinders	98%	98%	98%	4156
8 Cylinders	96%	94%	95%	1662
Accuracy			97%	5818
Macro Avg	97%	96%	97%	5818
Weighted Avg	97%	97%	97%	5818

Pruned Decision Tree

	Precision	Recall	F1-Score	Support
4 Cylinders	100%	98%	99%	4156
8 Cylinders	96%	99%	97%	1662
Accuracy			98%	5818
Macro Avg	98%	99%	99%	5818
Weighted Avg	98%	98%	98%	5818

Naïve Bayes

	Precision	Recall	F1-Score	Support
4 Cylinders	99%	99%	99%	4156
8 Cylinders	96%	98%	97%	1662
Accuracy			98%	5818
Macro Avg	98%	98%	98%	5818
Weighted Avg	98%	98%	98%	5818


Decision Tree (AUROC = 0.995)
Random Forest (AUROC = 0.991)
Naive Bayes (AUROC = 0.999)

ROC Curve