Vehicle Fuel Economy Prediction

Presented by
Artem Ramus

Introduction and Background

The goal of this project is to predict fuel consumption, aka 'combined MPG', from 80 cars' attributes.

Content

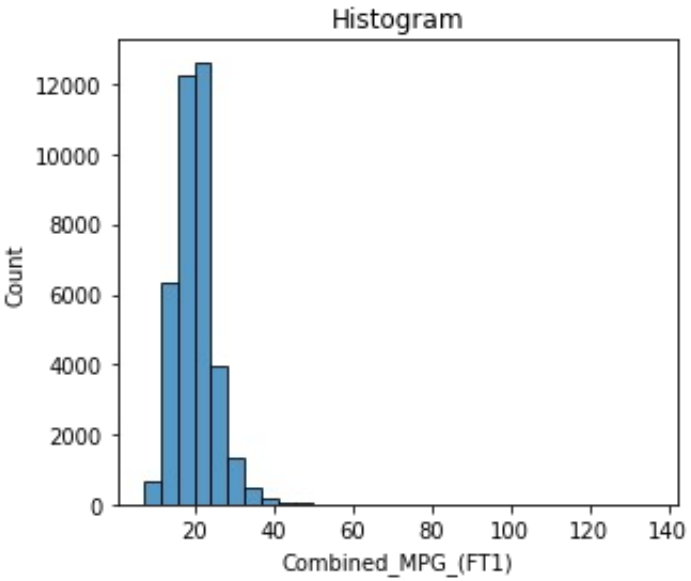
The purpose of EPA's fuel economy estimates is to provide a reliable basis for comparing vehicles. Most vehicles in the database (other than plug-in hybrids) have three fuel economy estimates: a "city" estimate that represents urban driving, in which a vehicle is started in the morning (after being parked all night) and driven in stop-and-go traffic; a "highway" estimate that represents a mixture of rural and interstate highway driving in a warmed-up vehicle, typical of longer trips in free-flowing traffic; and a "combined" estimate that represents a combination of city driving (55%) and highway driving (45%). Estimates for all vehicles are based on laboratory testing under standardized conditions to allow for fair comparisons.

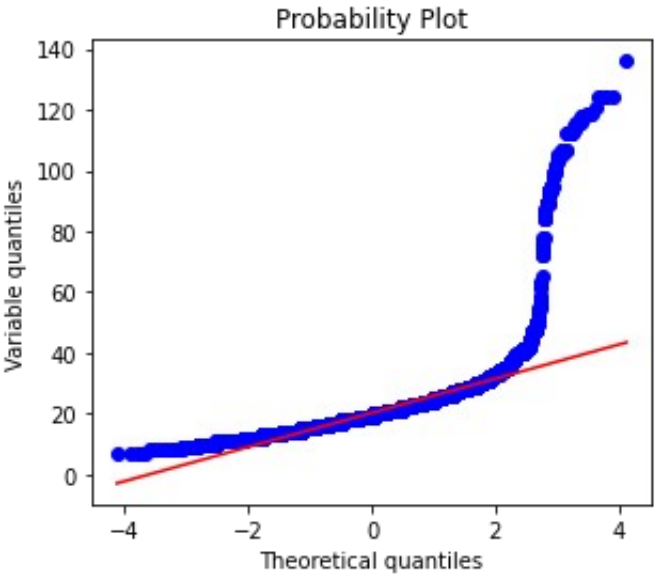
Introduction and Background

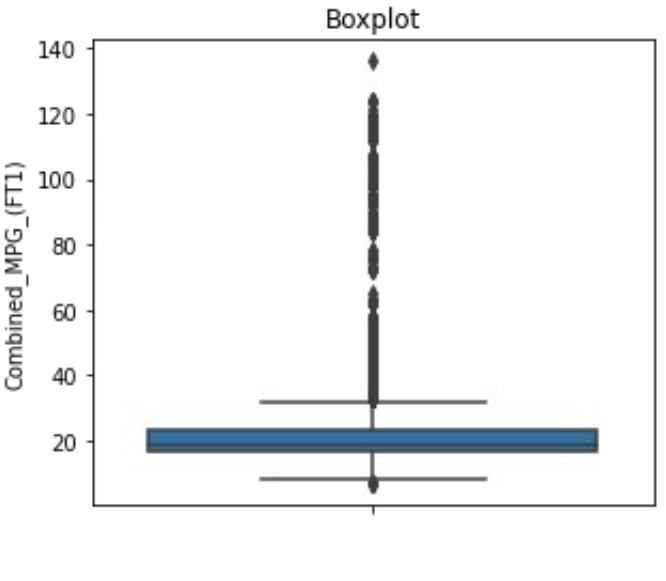
Acknowledgments
Fuel economy data are produced during vehicle testing at the Environmental Protection Agency's National Vehicle and Fuel Emissions Laboratory in Ann Arbor, Michigan, and by vehicle manufacturers with EPA oversight.

Link to the data set at Kaggle:
https://www.kaggle.com/prasy46/credit-score-prediction

Exploratory Data Analysis – target

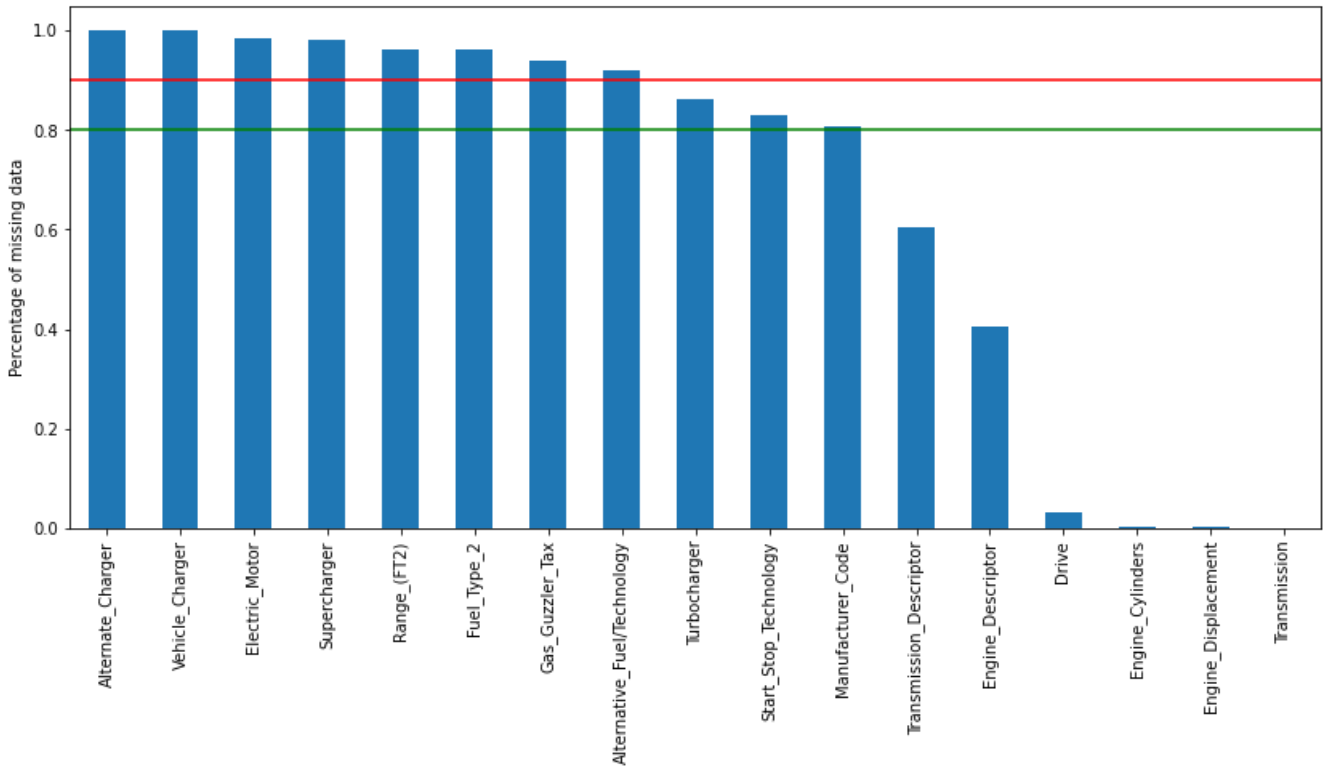






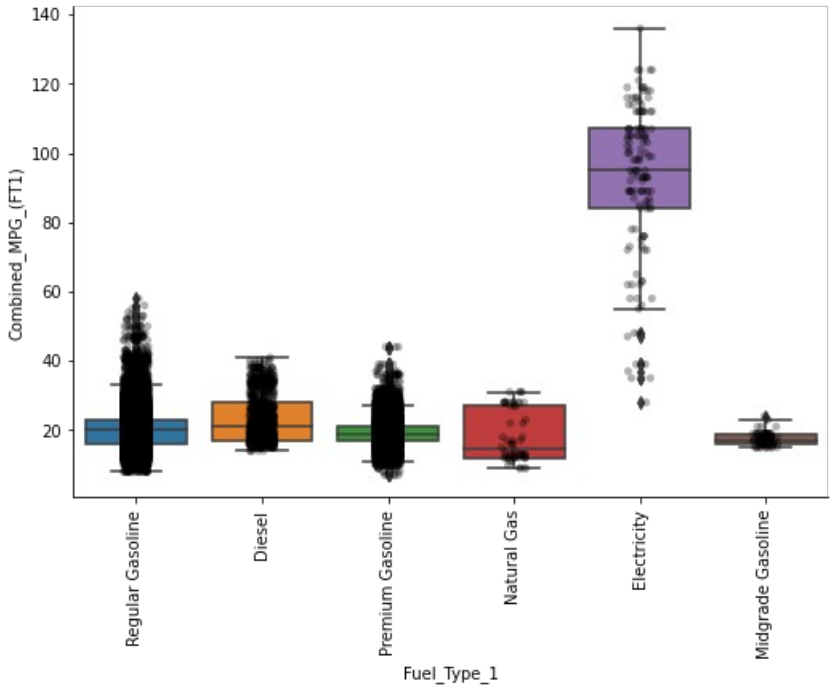
The target distribution is right skewed. Skewness of Combined_MPG_(FT1) is 5.73.

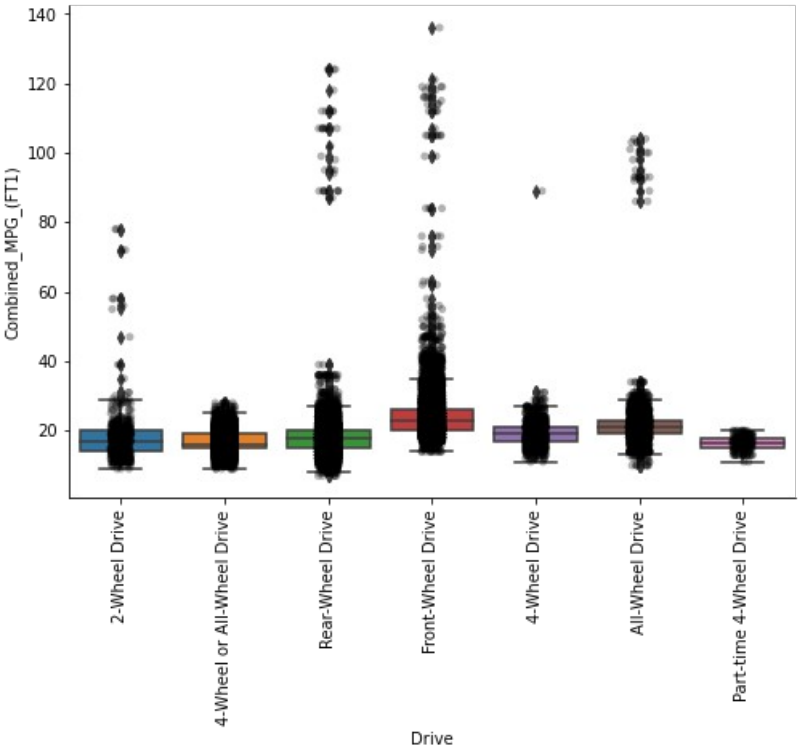
Exploratory Data Analysis – missing values



There are 17 variables with missing values. 9 of them has more than 90% missing values.

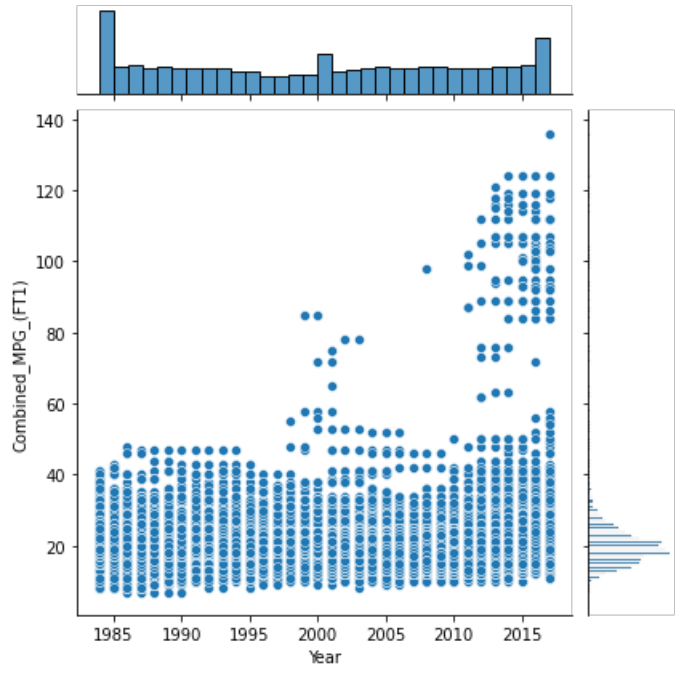
Exploratory Data Analysis – categorical variables

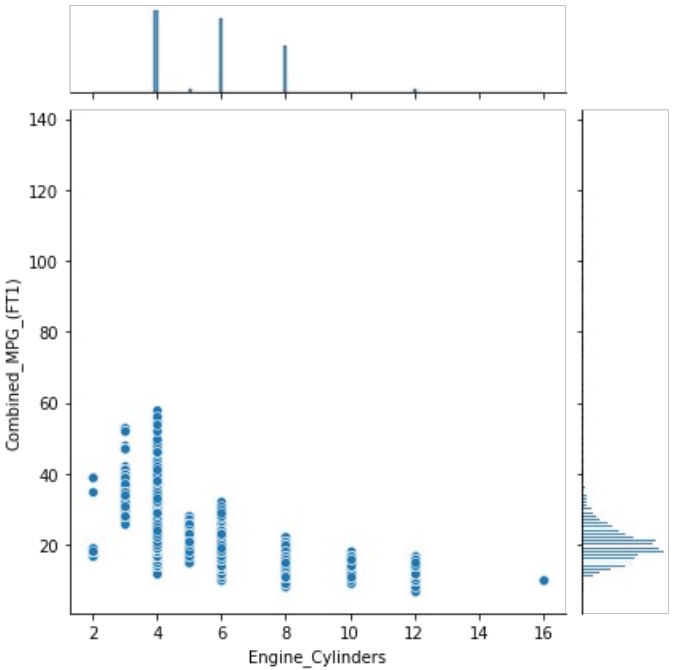


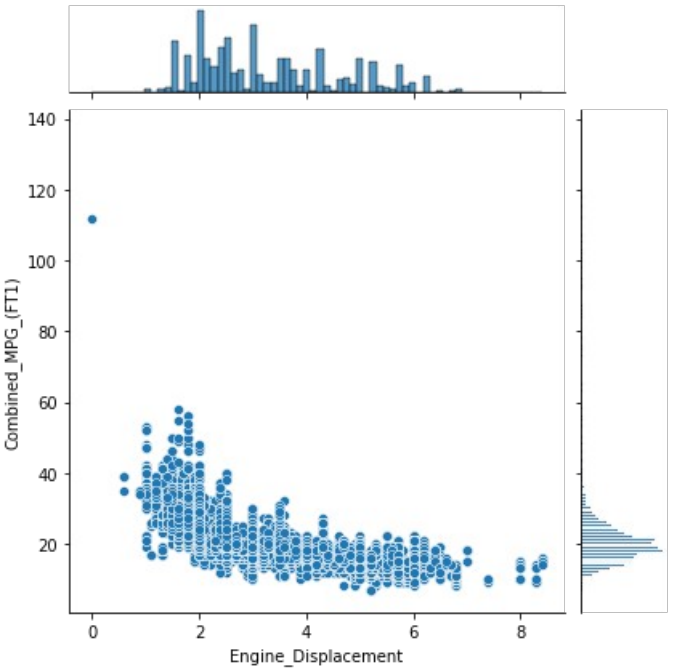


There are 18 variables. 9 of them vary with the target and may be used as predictors.

Exploratory Data Analysis – continuous variables







There are 10 variables . All of them are correlated with the target. Some variables, like engine cylinders and displacement are inter-correlated.

Feature Engineering – missing values imputation

Boolean
For boolean variables, N was imputed.

Categorical
For categorical variables, mode was imputed.

Numerical
engine displacement was imputed with the train set mean value.

Feature Engineering – categorical variables

Some categorical variables are fixed based on domain knowledge and their influence on the target.

Compact Cars	4980
Subcompact Cars	4372
Midsize Cars	3925
Standard Pickup Trucks	2115
Sport Utility Vehicle - 4WD	1889
Large Cars	1706
Two Seaters	1671
Sport Utility Vehicle - 2WD	1467
Small Station Wagons	1359
Special Purpose Vehicles	1332
Minicompact Cars	1127
Standard Pickup Trucks 2WD	1067
Vans	1018
Standard Pickup Trucks 4WD	898
Midsize-Large Station Wagons	575
Special Purpose Vehicle 2WD	548
Small Pickup Trucks	496
Midsize Station Wagons	471
Small Sport Utility Vehicle 4WD	464
Standard Sport Utility Vehicle 4WD	402
Small Pickup Trucks 2WD	394
Vans, Cargo Type	386
Small Sport Utility Vehicle 2WD	358
Minivan - 2WD	308
Vans, Passenger Type	285
Special Purpose Vehicle 4WD	278
Small Pickup Trucks 4WD	190
Standard Sport Utility Vehicle 2WD	167
Minivan - 4WD	42
Standard Pickup Trucks/2wd	4
Vans Passenger	2
Special Purpose Vehicles/4wd	2
Special Purpose Vehicles/2wd	2
Special Purpose Vehicle	1

Class variable fixed

Subcompact Car	5499
Compact Cars	4980
Midsize Car	4971
Sport Utility Vehicle	4747
Standard Pickup Truck	4084
Special Purpose Vehicle	2163
Large Cars	1706
Van	1691
Two Seaters	1671
Small Station Wagons	1359
Small Pickup Truck	1080
Minivan	350

Feature Selection

- Based on the correlation analysis, 4 inter-correlated numerical features removed.
- Based on the correlation analysis, 2 low correlated with target features, quasi-constant categorical features removed
- Based on feature shuffling, 3 features removed.
- 8 features selected for modelling

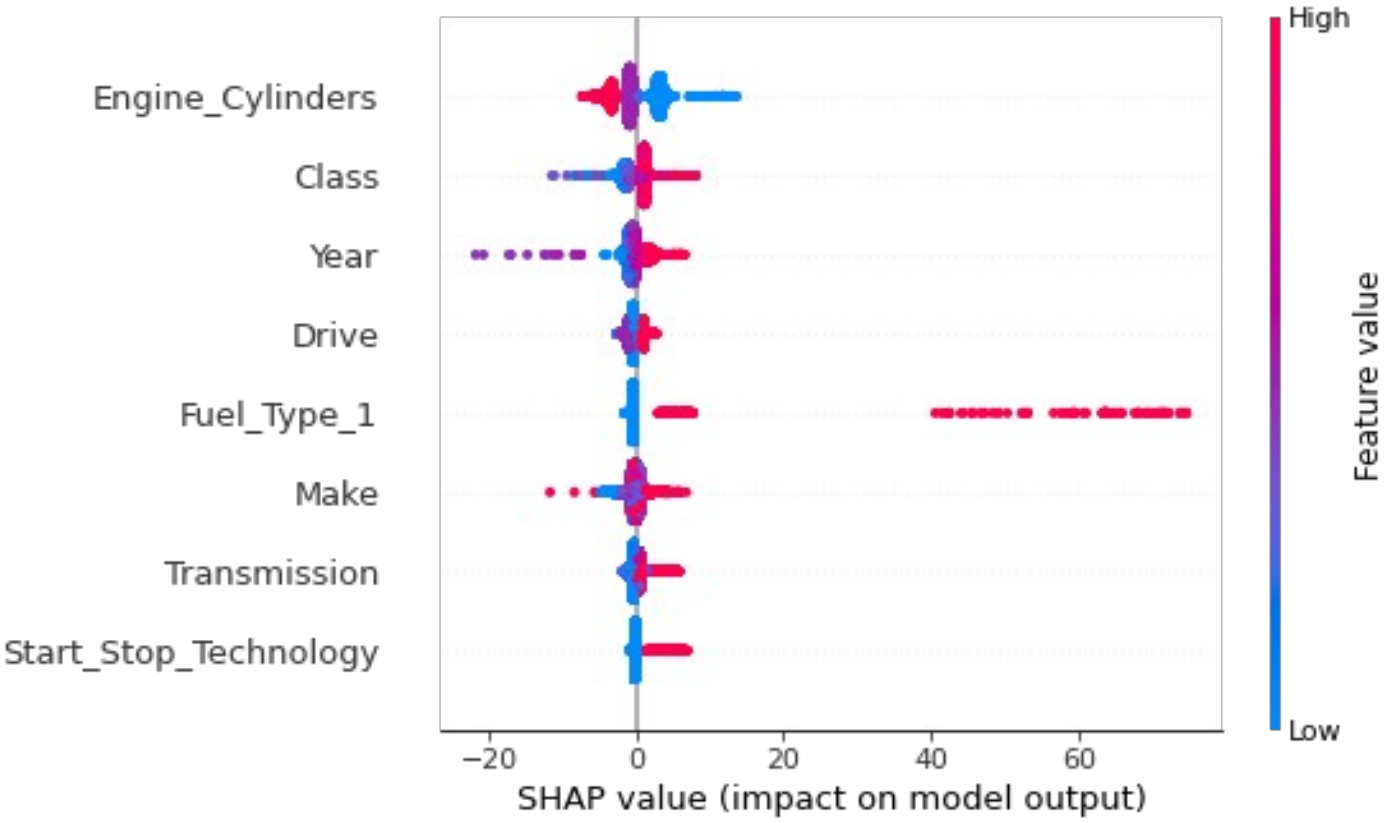
Performance

Model
XGBoost regressor

Cross validation
RMSE: 92.75 %, Standard Deviation: 1.09 %

Evaluation
RMSE train: 2.92 MPG
RMSE test : 3.01 MPG
r2 train : 0.94
r2 test   : 0.94

Feature Importance - SHAP



Summary

- 8 of 81 features are selected for modeling.
- With the selected features credit scores were predicted with RMSE_train of 2.92 and RMSE_test of 3.01 MPG. R2_train 0.94 and R2_test 0.94.

The end

Thank you for your attention!