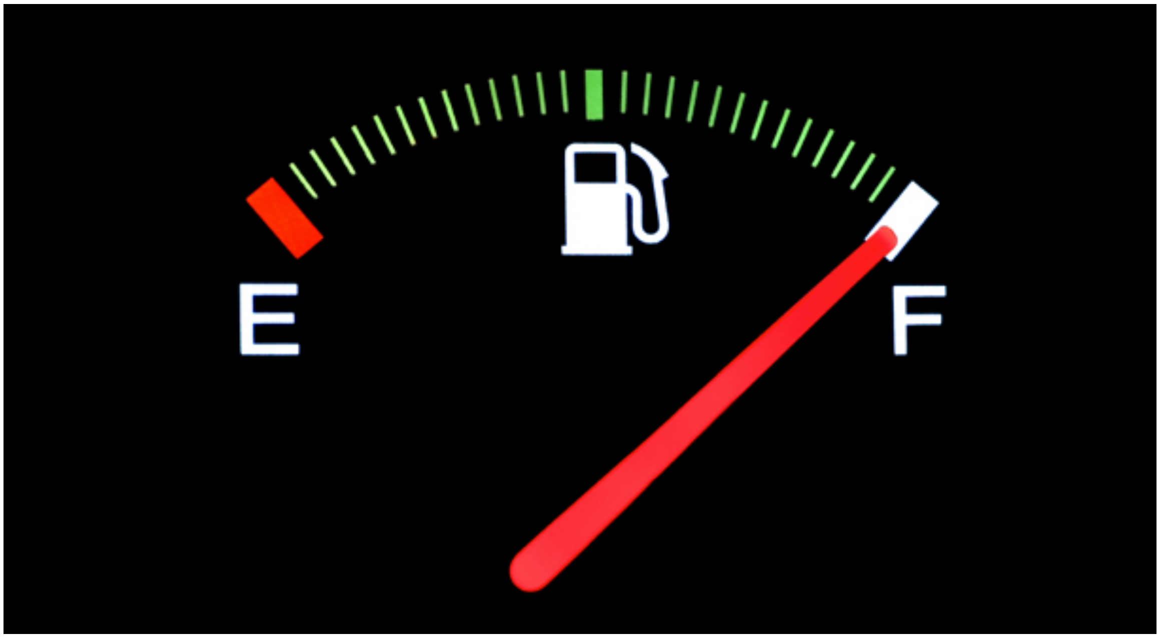
MILES-PER-GALLON, THE MILESTONE IS HERE



Nathan Nguyen

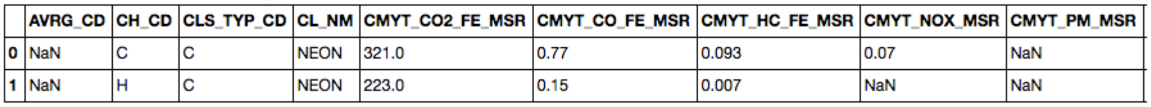
Objective

For years, EPA (Environmental Protection Agency) has been testing market accessible vehicles manufactured by various makers. These tests are used to verify the each of the components specified by the makers and compile into a large dataset for each year. This report contains the discussion and analysis on the fuel consumption reported by the EPA and multiple attempt to predict the **MPG (miles-per-gallon)**. The predictions will be based on the given information by the manufacturers to the EPA for each model at that given time.

Discussion

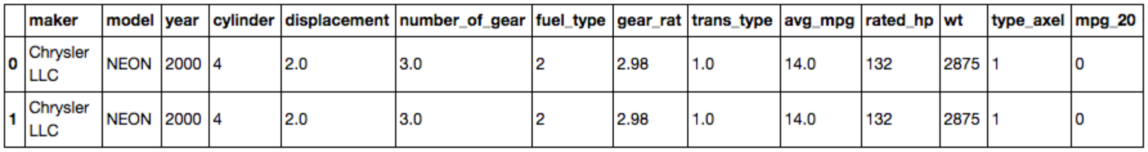
The datasets were obtained from EPA online databank in multiple files specified for each year (ranged from 2000-2009). The datasets contains labels such as: maker, model, engine size, engine displacement, manufactured rated horsepower, curb weight, etc. These features are encoded in alpha-numerical texts and contains either text, numerical or a mixture of both.

Before data cleaning:



	AVRG_CD	CH_CD	CLS_TYP_CD	CL_NM	CMYT_CO2_FE_MSR	CMYT_CO_FE_MSR	CMYT_HC_FE_MSR	CMYT_NOX_MSR	CMYT_PM_MSR
0	NaN	C	C	NEON	321.0	0.77	0.093	0.07	NaN
1	NaN	H	C	NEON	223.0	0.15	0.007	NaN	NaN

After data cleaning:



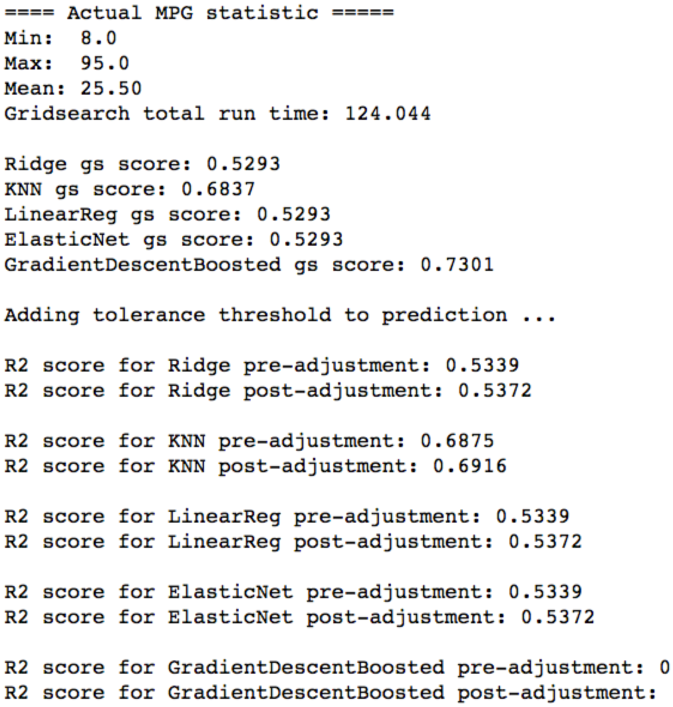
	maker	model	year	cylinder	displacement	number_of_gear	fuel_type	gear_rat	trans_type	avg_mpg	rated_hp	wt	type_axel	mpg_20
0	Chrysler LLC	NEON	2000	4	2.0	3.0	2	2.98	1.0	14.0	132	2875	1	0
1	Chrysler LLC	NEON	2000	4	2.0	3.0	2	2.98	1.0	14.0	132	2875	1	0

For a preliminary research, a new column was created (see above) labeled "mpg_20" to mark whether a model of a particular vehicle get at least 20 miles to the gallon based on average consumption.

Selected features:

- Engine cylinders
- Engine displacement
- Year of manufacture
- Rated Horsepower
- Curb weight

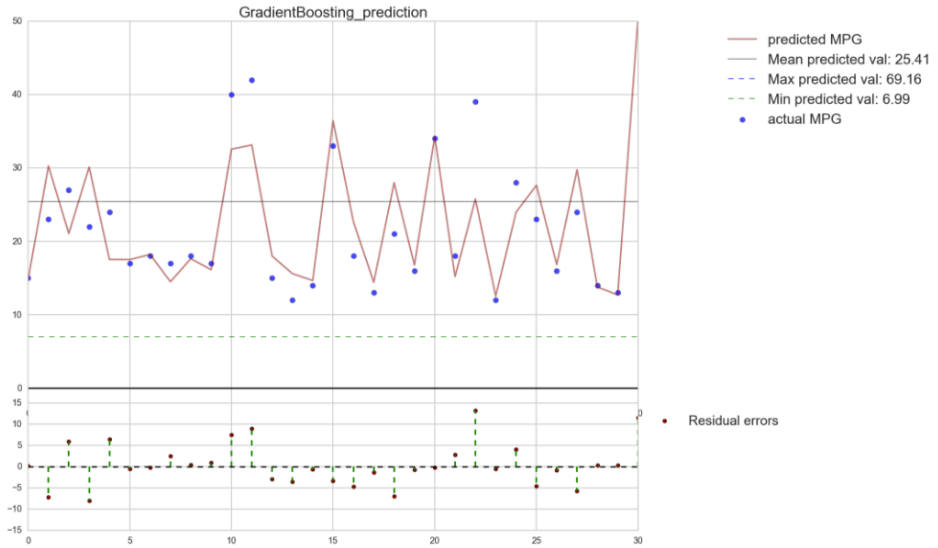
The reason behind which of the features to keep rely on the information given to consumers when they go to any dealership. Even those the information seem limited, the prediction was gracious enough to yield acceptable numbers.



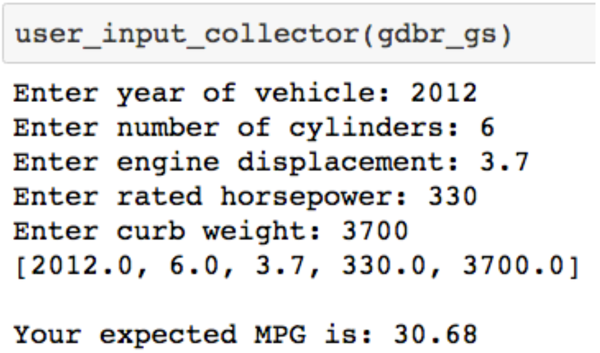
```
==== Actual MPG statistic =====
Min:  8.0
Max:  95.0
Mean: 25.50
Gridsearch total run time: 124.044

Ridge gs score: 0.5293
KNN gs score: 0.6837
LinearReg gs score: 0.5293
ElasticNet gs score: 0.5293
GradientDescentBoosted gs score: 0.7301

Adding tolerance threshold to prediction ...

R2 score for Ridge pre-adjustment: 0.5339
R2 score for Ridge post-adjustment: 0.5372

R2 score for KNN pre-adjustment: 0.6875
R2 score for KNN post-adjustment: 0.6916

R2 score for LinearReg pre-adjustment: 0.5339
R2 score for LinearReg post-adjustment: 0.5372

R2 score for ElasticNet pre-adjustment: 0.5339
R2 score for ElasticNet post-adjustment: 0.5372

R2 score for GradientDescentBoosted pre-adjustment: 0
R2 score for GradientDescentBoosted post-adjustment:
```

Note:

- GS scores are the accuracy scores yielded after cross-validation (the higher the score, the more accurate the prediction agreed with actual values)
- R2 scores are the squared errors, which indicate the percent of error that can be explained (the measure of slope fitment)



The information provided by Gradient Boosting regressor was the likely match to what we extracted from the actual dataset, thus it was selected. A function which takes in user input showcase the prediction on MPG when a user would like to check on their prior belief of a car's fuel consumption.



```
user_input_collector(gdbr_gs)
```

```
Enter year of vehicle: 2012
Enter number of cylinders: 6
Enter engine displacement: 3.7
Enter rated horsepower: 330
Enter curb weight: 3700
[2012.0, 6.0, 3.7, 330.0, 3700.0]

Your expected MPG is: 30.68
```

Conclusion

The model was successfully built with relatively high score of accuracy, which may seem to predict well. But as we run prediction on a single input, the number did not seem to agree with realistic scenario. More data will be required to allow this model reduce less error.