Report – Automotive Performance Metric Design

Yidan Nie

Summary of report

In this analytics project, I first stated my understanding of a good metric and a good work flow of an analytics project in the company.

Then I summarized three challenges in the automotive industry and stated my assumptions about this task, the design of the metric should reflect efforts to overcome these challenges. Next, I cleaned and explored the dataset with visualizations and statistics. The rationales of metric design are also depicted with formulas and results of top 10 cars ranking based a score metric:

Score = a * rank_mpg + b * rank_acce + c * rank_weight

The alternative method I recommended is conjoint analysis and I also talked about the improvement of this task with additional data sources.

Metrics

First of all, establishing a metric should consider about the following four targets:

1. **Closely associated with objectives**

A good metric should be derived from a concrete product objective. The scope and granularity of this metric should be defined explicitly depending on the resources and time constraints.

2. **Measurable and actionable**

A good metric should be trackable and easy to analyze. An actionable metric means that the measurement should emphasize activity that does add value to the product, and we can deliver solid rationales for decision-making based on the trend of this metric.

3. **Easy to understand**

Metrics should act a role as the medium to help technical and non-technical people communicate effectively. A good metric should be simple to explain and make less confusion among people from different backgrounds because our ultimate goal is utilizing this metric to help people understand and improve their core product or performance.

4. **Updated**

Since metrics are dynamic and they will evolve over time as our product evolves, we always need to check which metric is no longer valuable and update our metrics to better monitor the trends.

The work flow of an analytics project

Because this task intentionally leaves out the specific needs and details of the project, I pretend that I am a data scientist in an internal consulting team, and I am assigned this project with limited guidance.

Before the first meeting with stakeholders, I need to explore the data and get some initial solution ideation on my own.

In practice, I think the work flow for an analytics project would be:

1. Define the objective and KPIs with stakeholders
2. Research on specific topic and deeper data exploration
3. Model development/ experiment implementation/ dashboard construction
4. Product launch/ model deployment
5. Maintenance

For each above step, it will include many feedback iterations and validations.

Challenges for automotive manufacturer Industry

1. Find the best balance of car features under the constraints (cost, profit, customer needs, government regulations)
2. Fierce competitions in the market
3. How to differentiate their product with low risk

Assumptions

For this task, I develop my analysis based on some assumptions

1. Customers of vehicles are always greedy about the car performance, typically they want good power performance with good fuel economy, superior interior space and infotainment, and an affordable price.
2. The acceleration could be a proxy for the engine performance.
3. Although the dataset is created in 1983, it still can help us make informed decisions for current car models.

Signals

Based on the feature 'car name', It is known that the majority of cars in this dataset are sedans and coupes, only 1 pickup.

The detailed steps for data cleaning can be seen in the Appendix.

The signals conclude 6 features and the key information is shown in Table 1.

Table 1. Automotive features

Name	Type	Physical Definition	category
mpg	numerical	Miles per Gallen	Fuel economy
cylinder	categorial	Power unit of an engine	Power
displacement	numerical	The total volume of all the cylinders in an engine	Power

horsepower	numerical	The amount of power an engine develops	Power
weight	numerical	The weight of a car	Safety
acceleration	numerical	The rate of the change of velocity	Power

Feature 'mpg' is often used to measure fuel economy. 'cylinder', 'displacement', 'horsepower', and 'acceleration' are measurements of the engine performance. 'weight' has significant impact on the power requirement [1] and safety for drivers.

The complete descriptive statistic table and scatterplots can be seen be seen in the Appendix.

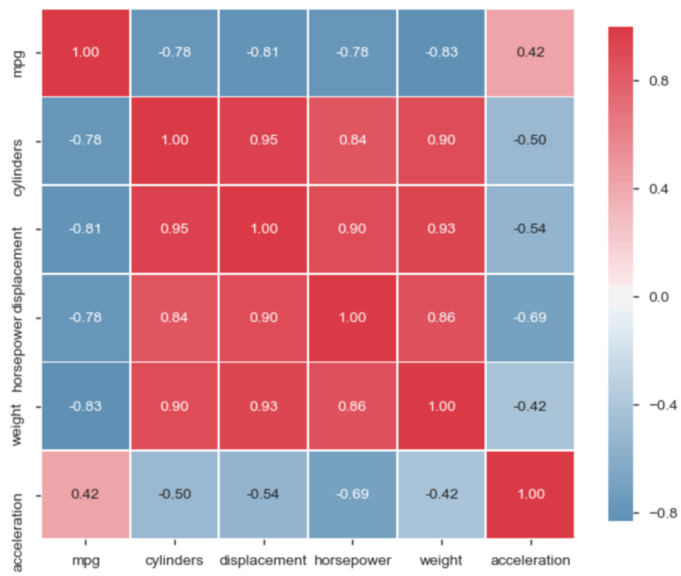


Figure 1. Correlation matrix for six features

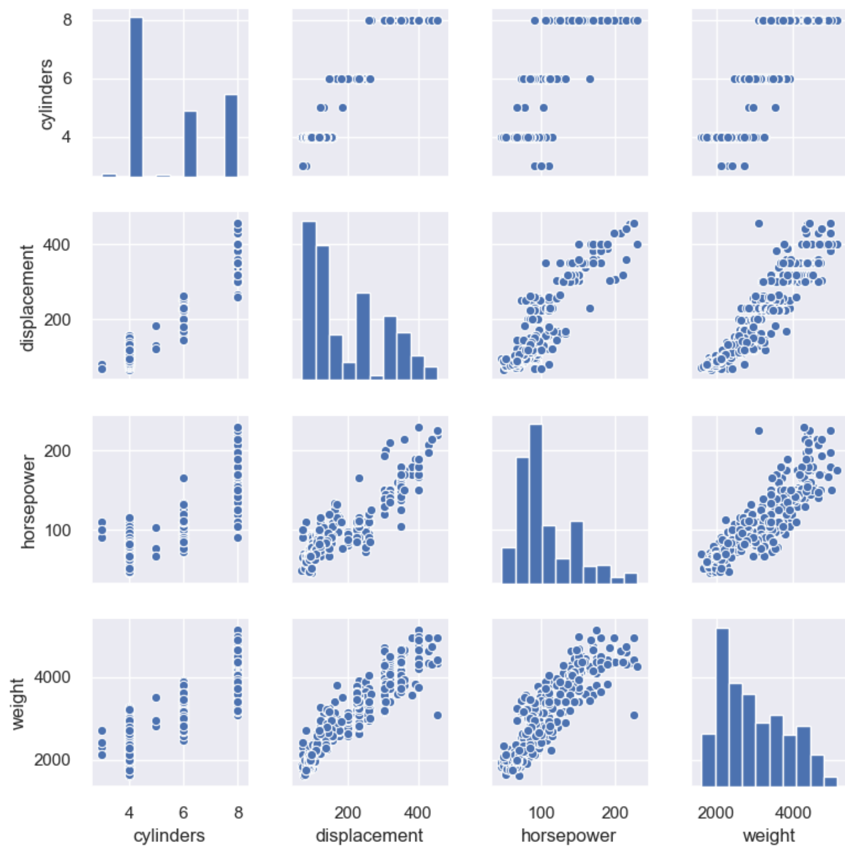


Figure 2. Paired scatter plots for six features

From the plot of correlation matrix, it can be seen clearly that the 'cylinder', 'displacement', 'horsepower' and 'weight' are highly correlated, and their correlations are greater than 0.9. With such high correlations, I have to drop three of them. Because feature 'weight' is the only feature representing safety issue, so I decide to remove 'cylinder', 'displacement' and 'horsepower' for the analysis and only leave feature 'acceleration' as the proxy for the power performance.

Table 2. Selected features

Name	**Type**	**Physical Definition**	**Category**	**Value**
mpg	numerical	Miles per Gallen	Fuel economy	9, …, 46.6
weight	numerical	The weight of a car	Safety	1613, …, 5140
acceleration	numerical	The rate of the change of velocity (m/s)	Power	8, …, 24.8

Table 3. Descriptive Statistics for selected features

stat	**mpg**	**acceleration**	**weight**
count	392	392	392
mean	23.45	15.54	2977.58
std	7.81	2.76	849.40
min	9.00	8.00	1613.00

25%	17.00	13.78	2225.25
50%	22.75	15.50	2803.50
75%	29.00	17.025	3614.75
max	46.60	24.80	5140.00

After the data exploration, I extracted three features respectively representing three main categories of the car performance. Compared to deleted features describing characteristics of the car engine, these three features are much easier to understand.

The detailed scatter plots and descriptive statistics are shown below.

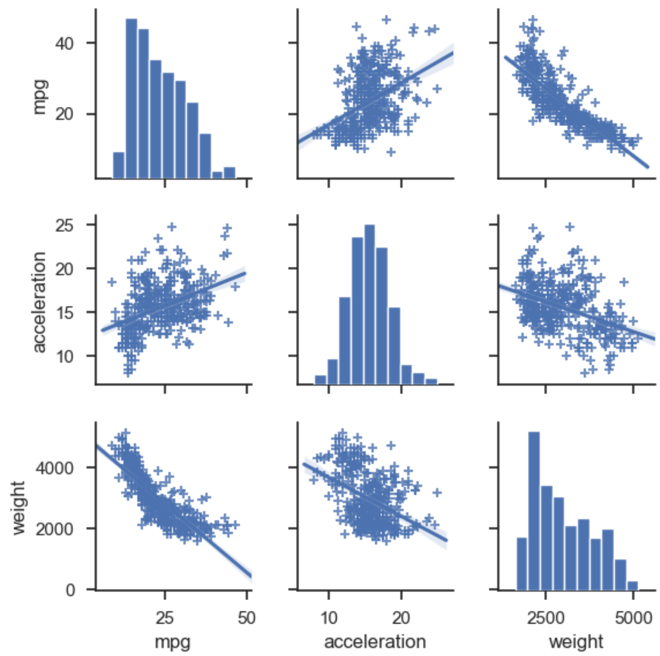


Figure 3. Scatter plots for selected plots

First, based on diagonal histograms, it can be seen that the distribution of acceleration is approximately symmetric. The distribution of mpg and weight is moderately skewed with skewness of 0.46 and 0.51. Based on scatter plots of feature pairs, it can be seen there is some negative association between weight and other two features. What's more, with the increase of mpg, the acceleration also increases. These relationships bring insights about the underlying relationships among the different categories of car performance.

Based on the dataset, the performance of cars in different categories has the following trends:

Fuel economy increase → **Weight** decrease *(relatively strong relationship)*
Engine Power increase *(relatively weak relationship)*

Engine Power increase → **Weight** decrease *(relatively weak relationship)*

Metrics Creation

To explore the optimal method to create a metric, I first tried to use ratio like 'mpg/weight' and 'mpg/acceleration' to evaluate the performance. The value of features is normalized using MinMax method. However, the result of this approach is not good and it's hard for further exploration of a ratio metric.

Finally, I decided to utilize the ranking index to build the metric. I first rank each feature from 1 to 391 as 'rank_mpg', 'rank_acce' and 'rank_weight' in ascending order. The larger rank number the car has, the larger value in that specific performance category. Then I build a score based on different business objective.

For this task, the stakeholder wants to create a new metric that can directly tell how good a car's performance is. In the automotive industry, car manufacturers break down the huge market into smaller segmentations that share similar needs. Because the definitions of 'good' vary hugely across the segmentations, my analysis will only focus on three main customer segmentations.

1. Business objective: create a metric for evaluating **an all-round car**
Good → Good balance of three categories

This dataset was collected in 1983. During that period, the domestic automotive industry expanded unprecedentedly. Customers preferred to buy trucks and larger, well-equipped cars [2]. Almost forty years have passed, the mainstream of customers still have similar preference. Based on the top 10 best-selling vehicles in the united states in 2018, it is obvious that cars with excellent balance of fuel economy, power performance and weight are the most popular choices [3].

Score = 0.33*rank_mpg + 0.33*rank_acce + 0.33*rank_weight

Table 4. Top 10 all-around cars

	mpg	acceleration	weight	rank_mpg	rank_acce	rank_weight	score
audi 5000s (diesel)	36.4	19.9	2950	369.0	369.0	217.0	315.15
mercedes-benz 240d	30.0	21.8	3250	307.0	383.0	252.0	310.86
volvo diesel	30.7	19.6	3160	311.0	367.0	241.0	303.27
oldsmobile cutlass ls	26.6	19.0	3725	260.0	354.0	304.0	302.94
peugeot 505s turbo diesel	28.1	20.4	3230	286.0	372.0	249.0	299.31
peugeot 504	27.2	24.8	3190	271.0	391.0	243.0	298.65
oldsmobile cutlass ciera (diesel)	38.0	17.0	3015	379.0	293.0	227.0	296.67
vw dasher (diesel)	43.4	23.7	2335	387.0	389.0	120.0	295.68
mercedes benz 300d	25.4	20.1	3530	238.0	370.0	286.0	295.02
oldsmobile cutlass salon brougham	23.9	22.2	3420	211.0	387.0	271.0	286.77

2. Business objective: create a metric for evaluating the performance of a **fuel-efficient heavier car**
Good → Good balance of fuel economy and weight

Within the 2018 best-selling car ranking, it can be seen the top 6 best-selling cars are pickups and efficient SUVs. Larger and heavier cars make people feel safe and power of control, but they are less environmental. So, an efficient heavier car should attract many customers to buy because of its superior engine design and aerodynamics.

Score = 0.6*rank_mpg + 0.4*rank_weight

Table 5. Top 10 fuel-efficient heavier cars

	mpg	acceleration	weight	rank_mpg	rank_acce	rank_weight	score
oldsmobile cutlass ciera (diesel)	38.0	17.0	3015	379.0	293.0	227.0	318.2
audi 5000s (diesel)	36.4	19.9	2950	369.0	369.0	217.0	308.2
datsun 280-zx	32.7	11.4	2910	337.0	23.0	209.0	285.8
mercedes-benz 240d	30.0	21.8	3250	307.0	383.0	252.0	285.0
volvo diesel	30.7	19.6	3160	311.0	367.0	241.0	283.0
vw dasher (diesel)	43.4	23.7	2335	387.0	389.0	120.0	280.2
oldsmobile cutlass ls	26.6	19.0	3725	260.0	354.0	304.0	277.6
datsun 510 hatchback	37.0	15.0	2434	370.0	178.0	137.0	276.8
triumph tr7 coupe	35.0	15.1	2500	358.0	181.0	143.0	272.0
peugeot 505s turbo diesel	28.1	20.4	3230	286.0	372.0	249.0	271.2

3. Business objective: create a metric for evaluating the performance of a **fuel-efficient performance car**
Good → Good balance of fuel economy and acceleration

For those customers who prefer a performance car with high rate of fuel utilization, the good balance of fuel economy and power performance would be a better performance metric.

Score = 0.6*rank_mpg + 0.4*rank_acce

Table 6. Top 10 fuel-efficient performance cars

	mpg	acceleration	weight	rank_mpg	rank_acce	rank_weight	score
vw pickup	44.0	24.6	2130	388.0	390.0	72.0	388.8
vw dasher (diesel)	43.4	23.7	2335	387.0	389.0	120.0	387.8
vw rabbit c (diesel)	44.3	21.7	2085	389.0	382.0	58.0	386.2
volkswagen rabbit custom diesel	43.1	21.5	1985	386.0	381.0	37.0	384.0
datsun 210	40.8	19.2	2110	384.0	357.0	62.0	373.2
audi 5000s (diesel)	36.4	19.9	2950	369.0	369.0	217.0	369.0
datsun 210 mpg	37.0	19.4	1975	371.0	359.0	34.0	366.2

toyota corolla tercel	38.1	18.8	1968	380.0	343.0	31.0	365.2
datsun b210 gx	39.4	18.6	2070	383.0	338.0	55.0	365.0
mazda glc	46.6	17.9	2110	391.0	316.0	61.0	361.0

It should be noted that the coefficients of each score metric set arbitrarily, in practice it should be designed relying on research on historical data and customer surveys.

Alternatives

I would recommend using conjoint analysis to collect more data about customer's evaluation on car performance. Conjoint analysis is a statistical technique used in market research to determine how people value different features that make up an individual product or service.

The objective of conjoint analysis is to determine what combination of a limited number of attributes is most influential on respondent choice or decision making, which is highly relevant to the needs of this task. A controlled set of potential products is shown to respondents and by analyzing how they make preferences between these products, the implicit valuation of the individual elements making up the product or service can be determined. Once the preference ranking is obtained, we can find the utilities of different values of each attribute that would result in the respondent's order of preference.

This method is efficient because the survey does not need to be conducted using every possible combination of attributes. The utilities can be determined using a subset of possible attribute combinations. From these results one can predict the desirability of the combinations that were not tested. The result of conjoint analysis can be used to create market prediction models that estimate market share, revenue and even profitability of new designs.

Improvement for this dataset

The modern car creation process is a much complex system: after engineers build a concept for car models, finance and strategy departments will join to research on risk and price point of this car under some strict budget constraints.

For this task, if the stakeholder has to build a performance metric based on a limited dataset, it would be better if he can collect new data sources on comfort, safety and electronics categories into this dataset. Because usually customers rate cars based on Performance, Comfort and Quality, Safety, Features, and Energy Green[4], more signals or features would definitely make our metrics more robust.

Reference

[1] https://www.sciencedirect.com/topics/engineering/fuel-economy
[2] Automotive Fuel Economy, National Research Council, Chapter 5 IMPACTS ON THE AUTOMOTIVE INDUSTRY
[3] https://www.foxnews.com/auto/the-10-best-selling-vehicles-in-the-united-states-in-2018-were-mostly-trucks-and-suvs
[4] https://cars.usnews.com/cars-trucks/honda/civic/performance

Appendix

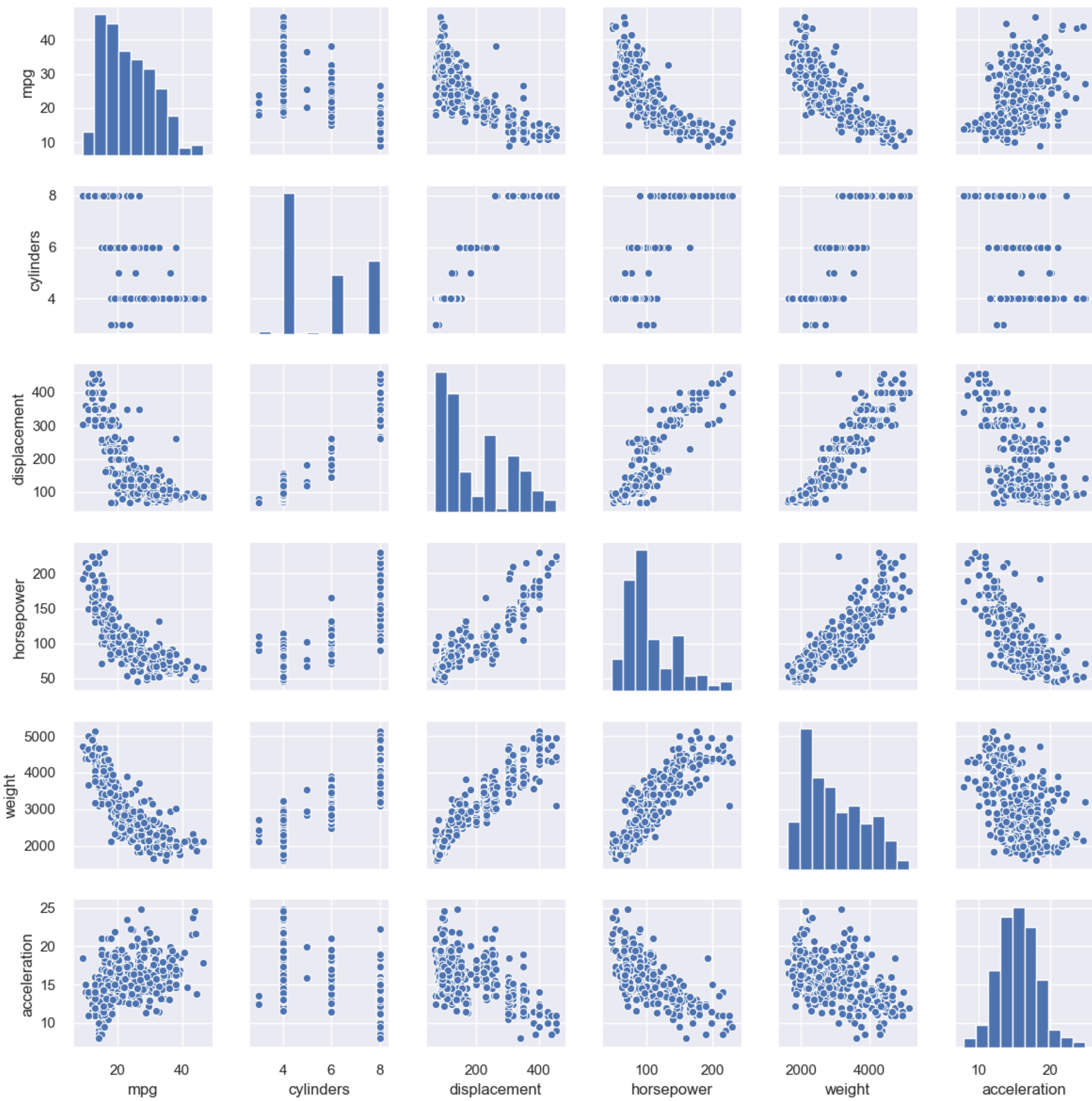
1.Data cleaning and engineering

1. drop columns 'model year' and 'origin'
2. drop the rows with missing values for horsepower
3. drop one duplicate row (row 18)
4. transform data in 'horsepower' column from categorial value into numerical value

2.Descriptive statistics for all signals

	mpg	cylinders	displacement	horsepower	weight	acceleration
count	392.000000	392.000000	392.000000	392.000000	392.000000	392.000000
mean	23.445918	5.471939	194.411990	104.469388	2977.584184	15.541327
std	7.805007	1.705783	104.644004	38.491160	849.402560	2.758864
min	9.000000	3.000000	68.000000	46.000000	1613.000000	8.000000
25%	17.000000	4.000000	105.000000	75.000000	2225.250000	13.775000
50%	22.750000	4.000000	151.000000	93.500000	2803.500000	15.500000
75%	29.000000	8.000000	275.750000	126.000000	3614.750000	17.025000
max	46.600000	8.000000	455.000000	230.000000	5140.000000	24.800000

3.Scatter plots for all signals


mpg
cylinders
displacement
horsepower
weight
acceleration