Predicting Customer Approval of Motorcycles

W203 Lab 2 Report

Amina Alavi, Gary Kong, Vernon Robinson, Elizabeth Willard

Contents

1 Introduction

Motorcycle manufacturers face several key challenges in the current global motorcycle market. The market is highly competitive, with 368 distinct brands and 7677 models in bikez.com's 2013 - 22 database of motorcycles. Meanwhile, easy access to information makes customers more discriminating in purchasing decisions. Therefore, there is a need to design products in a data-driven way to satisfy customer preferences.

Many different aspects of product design could impact how well a motorcycle is received by customers. These include, but are not limited to, motorcycle category, engine responsiveness, transmission, handling, suspension, and styling. Nonetheless, engine responsiveness is one of the most important features customers look for. This analysis provides a starting point toward understanding how engine responsiveness impacts customer acceptance by asking the following research question:

> *How does motorcycle engine responsiveness impact customer approval?*

The answer to this question could help quantify how engine responsiveness impacts customer approval, especially compared to other features. Insights from our analysis are particularly relevant to product design teams, who can use this information to focus on designing products with features more likely to satisfy customer preferences. Investors may also use the outcomes of this research to better predict a given motorcycle's success.

2 Data and Methodology

The data in this study came from bikez.com, which provides high-quality motorcycle specification data from 1894 to 2022. A custom scraper extracted the data on 30 April 2022 to enrich an existing used motorcycle dataset for a Kaggle hackathon competition. It was compiled and made publicly available by Emmanuel F. Werr. The observational data includes 28 unique features that describe a given motorcycle. The data source is a .CSV file containing information of 38,472 motorcycles, each row representing a motorcycle (based on brand, model, and year combination).

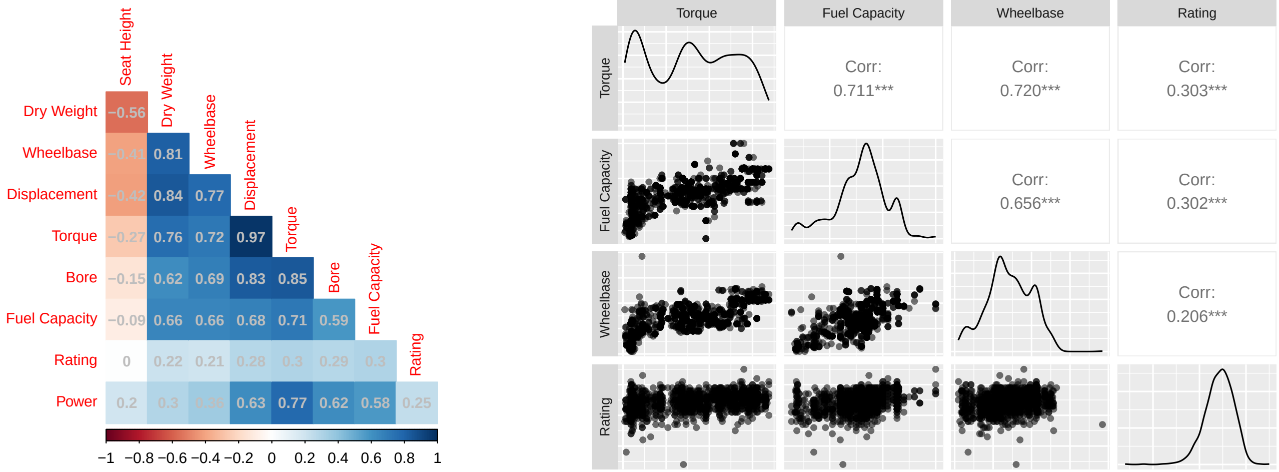
To operationalize our Y concept of customer approval, we used the *rating* variable, a mean of customer review scores (1-5 scale). Rating is a strong proxy for customer approval as it directly quantifies how well-received a motorcycle is by customers. Customers' individual ratings are ordinal but can be treated as interval due to equal differences between categories. Also, the rating variable in the dataset is metric as it is an aggregated mean of individual ratings.

To operationalize our X concept of engine responsiveness, we used the *torque* variable, which captures maximum torque in Newton-meters (Nm). The higher the torque, the faster the bike accelerates. Torque is one of the most important parameters that determine the responsiveness of a motorbike. We included *category*, *wheelbase*, *cooling system*, *front brakes*, *transmission type*, and *year* as secondary X variables.

The raw dataset includes 38472 motorcycles. We limited the dataset to motorcycles with a *year* between 2013 - 22, which leaves 14422 observations. We removed observations with missing values for any of our operationalization variables. First, we removed motorcycles with missing *rating* values, which leaves 1327 observations. No motorcycles had missing values for *torque*, *fuel capacity* and *wheelbase.* Second, we removed motorcycles with missing *cooling system* values, which leaves 1325 observations. Third, we removed motorcycles with missing *front brake* values, which leaves 1324 observations. Fourth, we removed motorcycles with missing *transmission type* values, which leaves 1296 observations.

We also manipulated four categorical variables (*cooling system*, *front brakes*, *transmission type*, and *category*), reducing the number of categories in each to simplify the model and facilitate model interpretation. For *cooling system*, we grouped similar categories (i.e., "Air" and "Oil & Air") together. For *front brakes*, *transmission type*, and *category*, we removed observations belonging to categories with a small number of Ns, which leaves 1135 observations.

To choose X variables to include in our model, we initially conducted a correlation analysis of all continuous variables (see Figure 1a). As the sample size after wrangling (n=1135) is not very large, we decided to not divide our data into exploratory and confirmation sets. We chose *torque* over *power* as the two are highly correlated (see Figure 1a), since power is a function of torque and speed (RPM). Some variables (*bore, stroke*, *displacement*) had been excluded from our research at the outset due to concerns of correlation to *power* and *torque,* which is confirmed here. Excluding these highly correlated variables minimizes the risk of having outcome variables on the right-hand side. While we initially hypothesized that *seat height* could be an explanatory variable, we chose to omit it as there is no correlation between *seat height* and *rating.* Also, since *wheelbase* and *dry weight* are highly correlated, we dropped *dry weight* and kept *wheelbase*, which is itself less correlated to *torque.*



(a) Correlations

(b) Pairwise Plots

Figure 1: Pearson Correlation of Continuous X Variables and Pairwise Plots of Continuous X and Y Variables

The bottom three plots in Figure 1b show *rating* as a function of *torque*, *fuel capacity*, and *wheelbase*. Notably, *rating* appears to increase as *torque* and *fuel capacity* increase, with moderate correlations between *fuel capacity* and *rating*, and *torque* and *rating*. In contrast, the relationship between *wheelbase* and *rating* appears less strong but is still significant. We chose to still include *wheelbase* as an X variable / covariate to reduce omitted variable bias.

Several categorical X variables were intentionally excluded from our analysis to minimize complexity. *Fuel system*, *fuel control*, *front suspension*, *rear suspension*, *color options* had excessively large numbers of distinct categories, making them unsuitable for regression. *Rear brakes* was excluded as *front brakes* are considered more important. *Engine cylinder*, *engine stroke* and *gearbox* were excluded as categories were ordinal and correlated with *torque*, so inclusion would have led to outcome variables on right-hand side. Based on the above, we specified five models:

Model 1: We wanted our first model to be simple and thus only used our primary X variable, *torque*, to predict *rating*, as exploratory plots show a linear relationship between the two variables. In this model, β_1 represents the change in *rating* for each unit increase in *torque* (all else kept constant). The model is:

$$\widehat{Rating} = \beta_0 + \beta_1 \cdot Torque + \epsilon$$

Model 2: We included all shortlisted X variables to assess statistical significance and to minimize omitted variable bias. All βs (except β_0) represent change in *rating* for each unit change of the X variable (all else kept constant). The model is:

$$\begin{aligned} \widehat{Rating} = \beta_0 + \beta_1 \cdot Torque + \beta_2 \cdot FuelCapacity + \beta_3 \cdot Wheelbase + \\ \beta_4 \cdot CoolingSystem + \beta_5 \cdot FrontBreaks + \beta_6 \cdot TransmissionType + \epsilon \end{aligned} \tag{1}$$

Model 3: Initial data exploration suggested that *category* affects *rating*. Thus, we included *category* to improve model fit[1]. Also, to simplify the model, we grouped all X variables with little significance under $\mathbf{Z}\gamma$. $\mathbf{Z}$ is a row vector for the covariates and γ is a column vector of coefficients.

$$\widehat{Rating} = \beta_0 + \beta_1 \cdot Torque + \beta_2 \cdot FuelCapacity + \beta_i \cdot Category_i + \mathbf{Z}\gamma + \epsilon \tag{2}$$

Model 4: The mean of *rating* appears to vary by year. We thus added *year* to our fourth model to correct for this and to improve the model's accuracy. *Year* is included as part of $\mathbf{Z}\gamma$ given many categories.

Model 5: For our final model we added *brand*. Like *year*, the mean of *rating* from brand to brand varies significantly. We added *brand* to our fifth model based on the same rationale as adding *year,* including *brand* as part of $\mathbf{Z}\gamma$.

We considered modeling interactions between *category* and other variables. However, the large number of categories would lead to excessive complexity in the model for the purposes of this assignment.

[1] The categories are represented by i in the equation. Indicator variables are "Classic", "Custom cruiser", "Enduro / offroad", "Naked bike", "Scooter", "Sport", "Super motard", "Touring". The baseline category is "All round".

3 Results

Table 1: Estimated Regressions

	Output Variable: Mean Rating (1-5 scale)				
	Rating				
	(1)	(2)	(3)	(4)	(5)
Torque (Nm)	0.002***	0.002***	0.002***	0.002***	0.001*
	(0.0002)	(0.0003)	(0.0004)	(0.0004)	(0.0005)
Fuel Capacity (Litres)		0.013***	0.016***	0.016***	0.018***
		(0.003)	(0.003)	(0.003)	(0.004)
Category: Classic			0.068	0.063	0.016
			(0.050)	(0.050)	(0.053)
Category: Custom / cruiser			−0.183**	−0.174**	−0.161*
			(0.065)	(0.064)	(0.067)
Category: Enduro / offroad			−0.148***	−0.139**	−0.198***
			(0.044)	(0.045)	(0.048)
Category: Naked bike			−0.105**	−0.095**	−0.109**
			(0.035)	(0.035)	(0.038)
Category: Scooter			−0.028	−0.014	−0.094
			(0.062)	(0.061)	(0.081)
Category: Sport			−0.174***	−0.159***	−0.187***
			(0.039)	(0.040)	(0.043)
Category: Super motard			−0.080*	−0.069	−0.134**
			(0.040)	(0.041)	(0.044)
Category: Touring			−0.127*	−0.117	−0.087
			(0.062)	(0.062)	(0.063)
Constant	3.274***	3.420***	3.471***	3.449***	3.680***
	(0.019)	(0.196)	(0.229)	(0.231)	(0.277)
Wheelbase (mm)		✓	✓	✓	✓
Cooling System		✓	✓	✓	✓
Front Brakes		✓	✓	✓	✓
Transmission Type		✓	✓	✓	✓
Year				✓	✓
Brand					✓
Observations	1,135	1,135	1,135	1,135	1,135
R^2	0.092	0.114	0.161	0.167	0.260
Residual Std. Error	0.296 (df = 1133)	0.293 (df = 1128)	0.286 (df = 1120)	0.286 (df = 1111)	0.276 (df = 1059)
Significance levels					*p<0.05; **p<0.01; ***p<0.001

Table 1 shows the results of five representative regressions[2]. The R-squared values increase as we go across the models, which indicates that the models become increasingly predictive. Across all models, the estimated coefficient for our primary X variable, *torque*, was highly statistically significant. Point estimates were consistent for the first four models at 0.002, and was 0.001 for model 5. To provide a sense of scale, consider a manufacturer which increased the torque of a motorcycle by 300 Nm, all else being equal. Applying model 5, the motorcycle's rating is expected to increase by 0.38.

[2] In model 2, wheelbase, cooling system, brakes, transmission type were part of the main model. However, they were pushed down as covariates starting from model 3 as they were not statistically significant. Robust standard errors are shown in parentheses.

Across all models, the coefficient for *fuel capacity* was also highly statistically significant. Point estimates ranged from 0.01 to 0.018. Another example to consider is two motorcycles with fuel capacity differing by 12 liters, all else being the same. Applying model 5, the rating of the motorcycle with the higher fuel capacity is expected to exceed the motorcycle with the lower fuel capacity by 0.21.

Most coefficients for *category* indicators were statistically significant. An example to consider is the "Enduro / offroad" category. Applying model 5, a motorcycle in the "Enduro / offroad" category is expected to have -0.2 lower ratings than the baseline category ("All round").

As previously noted, other covariates were included to adjust for omitted variable bias. Notably, the coefficients for *wheelbase*, *cooling system*, *front brakes*, *transmission type* were not statistically significant, as such, estimated coefficients for these should not be strictly used to make design decisions if the goal is to increase customer approval. We also note that *year* and *brand*, while being statistically significant for particular categories, are not practically significant as these are not parameters that a manufacturer can affect.

From a motorcycle manufacturer's point of view, when deciding which specifications to focus on, *torque* and *fuel capacity* seem to affect customer approval the most. Since *torque* is highly correlated with *power*, *bore*, *stroke* and displacement, these should also be included in the list of specifications to be optimized for if the goal is higher customer ratings. Manufacturers should also look into focusing on "Classic" motorbikes as this category is most likely to lead to the highest customer ratings.

4 Limitations

Consistent regression estimates require an assumption of independent and identically distributed (iid) observations. Motorcycles in the dataset belong to various brands, and motorcycles from specific brands may share common characteristics, so observations may not be fully independent. We partially account for this possibility in the fifth model by including *brand* as a covariate. Furthermore, observations may not be strictly identically distributed as motorcycles in the dataset were drawn from different years. We partially account for this possibility by restricting the *year* to 2013 - 22, and by including *year* as a covariate in models four and five.

Consistent regression estimates also require that a unique best linear predictor describes the population distribution. As shown in Figure 1b, there is no visual evidence of heavy-tailed distributions. There is also no perfect collinearity as correlations between pairs of X variables are all less than one, and no variables were dropped when fitting the regressions.

Heteroskedastic error is apparent in Figure 1b. This is less of a concern for the Large Sample Models used but is nonetheless accounted for by using robust standard errors. Additionally, it is unclear whether bikez.com used random sampling in choosing which motorcycles to include in its database. As such, there may be limitations as to the generalizeability of our findings across all motorcycles. Finally, while our models identified linear relationships between *rating.* and *torque* and *fuel capacity*, these relationships may not necessarily hold for very high values of *torque* and *fuel capacity*, particularly as *rating* should have a maximal value of 5.

Regarding structural limitations, a potential source of omitted variable bias is fuel efficiency, which did not have a corresponding variable in our data source. We postulate that *torque* has a negative relationship with *fuel efficiency*, and that *fuel efficiency* has a positive relationship with *rating.* As such, the direction of bias would be negative and towards zero, implying that the true effect of *torque* on *rating* is larger than the estimated effect. We do not consider reverse causality present, as rating should not cause changes in our X variables. We also do not have outcome variables on the right-hand side as X variable selection accounted for this (e.g., choosing torque over power due to correlation).

5 Conclusion

This study estimated the relationship between key motorcycle features and customer approval of a motorcycle, proxied by mean ratings. We also described and measured how motorcycle category predicts ratings. Follow-up models could include interactions between motorcycle categories and design features such as torque and fuel capacity, as the impact of different motorcycle features on ratings could differ depending on the motorcycle category. Follow-up models could also use simplified versions of categorical variables which were excluded due to excessive distinct categories. Motorcycle manufacturers may want to know, for example, whether there are benefits to choosing specific types of tires or suspension systems. Future research could also incorporate other datasets (e.g., sales data) to provide a more holistic view of product success. The ultimate goal of this line of work is to provide accurate tools for motorcycle manufacturers to develop motorcycles that are more likely to achieve success and reduce uncertainty in the design process.