Modeling Rides and Riders

Complex statistical models can accurately model intricate processes. But they also run the risk of overfitting to the data. To avoid this, we build up our models from simple to complex, comparing the models with cross validation to make sure the complexities introduced add real value.

In this chapter we focus on building models that incorporate information about rider, weather conditions, time of day, and ride length. In brief, our models start with a logistic regression model considering only ride-level variables, and formulate more complex models by adding various terms. Table 1 describes each model briefly along with the models label.

Table 1: Brief descriptions of Models 1–6

Model	**Description**
Model 1	(Baseline) logistic regression
Model 2	Add rider intercepts
Model 3	Add trigonometric terms for time of day
Model 4	Additive model with cubic cyclic spline for time of day
Model 5	Additive model with spline for ride length
Model 6	Remove random rider intercepts from Model 4

Six Models for Probability of a Negative Ride Rating

Model 1, which we will use as the baseline for comparing further models, is a logistic regression model:

$$\mathbb{P}(Y_i = 1) = \text{logit}^{-1}(\alpha + X_i\beta),$$

where $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$ are parameters to be estimated. (X is the matrix of ride-level predictors specified at the end of **??**.)

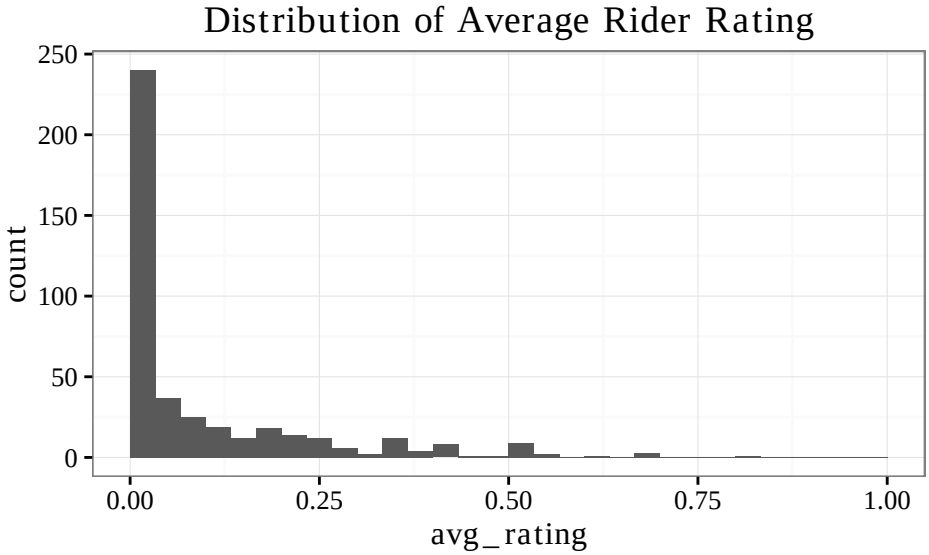

Figure 1: The overall rates at which each rider gives a negative rating for a ride varies greatly. This is our primary motivation for including rider intercepts and predictors.

Riders appear to have different tendencies to rate rides negatively more often, as we note in Figure 1. In fact, many riders give zero or nearly zero negative ratings. For **Model 2**, we account for this variability by adding intercepts that vary by rider:

$$y_i \sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\alpha + \alpha_{j[i]} + X_i\beta\right)\right), \tag{1}$$

for $i = 1, \ldots, n$.

Rider intercepts themselves aren't as interesting as how they deviate from the mean, so we keep a fixed intercept α and constrain the rider intercepts, α_j, by specifying

$$\alpha_j \sim N(0, \sigma_\alpha^2).$$

Starting with **Model 3**, we address time of day, $t \in [0, 24)$ as a predictor. (We measure time of day in hours since midnight.) We use time of day to account for all the daily trends that may affect ratings, including as a simple way to model the overall traffic level, which is difficult to model on its own. These patterns are cyclic and very non-linear, so we can't model time as a linear term. One approach is to add sinusoidal terms with a period of one day. We would be interested in fitting a term,

$$\beta \sin(Tx^{\text{time}} + \phi).$$

Estimating β wouldn't be hard: we can easily estimate coefficents of transformed terms; it's more difficult to estimate T and ϕ. But, we know that we want to restrict our terms to fitting trends that happen over the course of one day, so we can set $T = 2\pi/d$, where d is 24 hours or some fraction of that.

As for ϕ, a trigonometric transformation reframes the estimation of a phase shift parameter into the estimation of two coefficients for trigonometric functions with no phase shift:

$$\begin{aligned}
\beta \sin(Tx + \phi) &= \beta\left(\sin(Tx)\cos(\phi) + \cos(Tx)\sin(\phi)\right) \\
&= \beta\cos(\phi))\sin(Tx) + \sin(\phi)\cos(Tx) \\
&= \beta_1 \sin(Tx) + \beta_2 \cos(Tx),
\end{aligned}$$

where $\beta_1 = \beta\cos(\phi)$ and $\beta_2 = \sin(\phi)$. At this point, we are now just estimating the coefficients of a couple of transformed variables, which can easily be done in any package that does generalized linear regressions.

We also want to take into account that weekday hourly patterns may be different than weekend patterns. We use a variable X^{weekend} that serves as a weekend indicator. For Model 3, we add two sets of sinusoidal terms: one set for weekdays and one for weekends. More explicitly, we define the model,

$$\begin{aligned}
\mathbb{P}(Y_i = 1) = \text{logit}^{-1}(&\alpha + \alpha_{j[i]} + X_i\beta \\
&+ X^{\text{weekend}} \cdot [\beta^{t1}\sin(T \cdot t) + \beta^{t2}\cos(T \cdot t) \\
&+ \beta^{t3}\sin(T/2 \cdot t) + \beta^{t4}\cos(T/2 \cdot t)] \\
&+ (1 - X^{\text{weekend}}) \cdot [\beta^{t1}\sin(T \cdot t) + \beta^{t2}\cos(T \cdot t) \\
&+ \beta^{t3}\sin(T/2 \cdot t) + \beta^{t4}\cos(T/2 \cdot t))].
\end{aligned} \tag{2}$$

For **Model 4** we abandon parametric methods and use a cyclic non-parametric smoother to model time of day, making our model,

$$y_i \sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\alpha + \alpha_{j[i]} + X_i\beta + X^{\text{weekend}} \cdot f^{\text{time.w}}(t_i) + (1 - X^{\text{weekend}}) \cdot f^{\text{time}}(t_i)\right)\right), \tag{3}$$

for $i = 1, \ldots, n$, where α_j is specified like Model 2 and $f^{\text{time.w}}$ and f^{time} are cyclic cubic spline terms for weekend and weekday rides, respectively, with knots at 0, 3, 6, 9, 12, 15, 18, 21, and 24 (0, again) hours.

Model 5 extends Model 4 by adding a cubic spline for ride_length:

$$y_i \sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\begin{array}{l}\alpha + \alpha_{j[i]} + X_i\beta + f^{\text{length}}(x_i^{\text{log.length}}) + \\ X^{\text{weekend}} \cdot f^{\text{time.w}}(t_i) + (1 - X^{\text{weekend}}) \cdot f^{\text{time}}(t_i)\end{array}\right)\right), \tag{4}$$

$$y_i \sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\alpha + \alpha_{j[i]} + X_i\beta + X^{\text{weekend}} \cdot f^{\text{time.w}}(t_i) + (1 - X^{\text{weekend}}) \cdot f^{\text{time}}(t_i) + f^{\text{length}}(x_i^{\text{log.length}})\right)\right), \tag{5}$$

for $i = 1, \ldots, n$, where f^{length} is a cubic spline smoother.

Finally, **Model 6** is identical to Model 5, but without the rider intercepts:

$$y_i \sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\alpha + X_i\beta + f^{\text{time}}(t_i)\right)\right), \tag{6}$$

for $i = 1, \ldots, n$, where f^{time} is a cyclic cubic spline term, with the same knots as in Model 4.

Model Evaluation

Table 2: Fit summaries for Models 1–6.

Model	**Separation Plot**	$\log(\mathcal{L})$	**AIC**	**AUC**$_{\text{CV}}$[1]
Model 1		-4,786	9,586	0.552
Model 2		-3,971	7,957	0.797
Model 3		-3,923	7,877	0.802
Model 4		-3,930	7,870	0.802
Model 5		-3,928	7,878	0.803
Model 6		-4,713	9,455	0.601

To fit the data, we got all of the rides in Portland, OR, from December 3, 2014 to February 8, 2016 for riders that had over 20 rated rides. (We only look at riders with a certain number of rides to make sure we get can get estimates for rider-level parameters, like rider random intercepts, that aren't too uncertain.) There were 25,397 rides, 14,032 of which were rated. Overall, 10.88 percent of these rides were given a negative rating. There were 138 riders in the data set.

The separation plots in Table 2 give a clear initial picture of how these model fits compare. Model 1 performs very poorly compared to those that include rider intercepts, assigning the same probability to most observations. Models that include the rider intercept perform similarly to each other. The log likelihoods and AIC,[2] shown in Table 2, corroborate this. Adding time dependency doesn't seem to impact predictive ability. We will see later, however, that it gives a fascinating result to interpret.

The gains from the rider intercepts are great, but we are compelled to ask: how much of that gain could have been achieved with randomly chosen groups? In other words, if riders were randomly assigned to rides, would

[1] Area under ROC curve estimated with 10-fold cross-validation.

[2] Akaike Information Criterion (AIC) is a metric that penalizes the $\log(\mathcal{L})$ with the number of parameters estimated k, with lower values being better. It is defined as $\text{AIC} = 2k - 2\log(\mathcal{L})$.

the flexibility in the model created by allowing intercepts to vary increase predictive performance to the same degree? To test this, we ran a Model 4 after we randomly assign the rides a rider, by randomly permuting the rider column. This quick test nullified this skepticism, as you can see in the resulting separation plots in Figure 2.

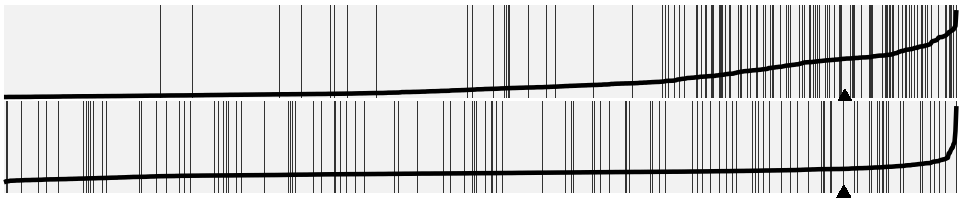

Figure 2: Separation plots for models 2 compared to a similar model where riders are randomly assigned to rides. This test demonstrates that the improvement in predictive performance provided by the rider intercepts was not coincidence.

Model Results

Table 3: Regression coefficients for Model 1, Model 2, Model 4, and Model 6. 95% confidence intervals are given in parentheses.

Regression Term	Model 1	Model 2	Model 4	Model 6
Log(length)	-0.122	-0.100	-0.092	-0.114
	(-0.180, -0.063)	(-0.168, -0.032)	(-0.162, -0.022)	(-0.174, -0.054)
Mean Temp.	0.053	0.076	0.075	0.069
	(-0.0004, 0.110)	(0.005, 0.147)	(0.003, 0.147)	(0.012, 0.127)
Mean Wind speed	0.028	0.014	0.012	0.027
	(0.004, 0.052)	(-0.013, 0.041)	(-0.014, 0.039)	(0.002, 0.051)
Gust speed	-0.003	0.001	0.001	-0.003
	(-0.015, 0.008)	(-0.012, 0.013)	(-0.012, 0.013)	(-0.014, 0.009)
Rainfall	0.008	0.012	0.008	0.005
	(-0.015, 0.031)	(-0.015, 0.038)	(-0.019, 0.035)	(-0.019, 0.027)
Rainfall 4-Hour	0.013	0.016	0.017	0.014
	(0.005, 0.021)	(0.007, 0.025)	(0.008, 0.027)	(0.006, 0.022)
Intercept	-2.2868	-3.075	-3.127	-2.313
	(-2.428, -2.108)	(-3.386, -2.764)	(-3.436, -2.818)	(-2.475, -2.151)

Table 3 presents the fixed effect estimates for our models. Length has a robust strong negative effect. This makes sense if we think of length as the only information we have about route in these models: it seems routes that are longer tend to be less likely to be rated negatively. Perhaps longer rides tend to be for leisure or sport rather than commuting, so and so are less likely to go along routes with high traffic and other dangers. Temperature also seems to have a small effect. It could be the case that the temperature itself is important, or perhaps season is a confounding factor. It could be the case that the type of rides taken during the warmer months are more likely to be rated negatively. Wind and gust speed don't seem to have robust effects. These models suggest that four hour cumulative rainfall was more important than rainfall during the hour of the ride. This supports our suspicion that weather effects that are more relevant to safety than comfort— like wet road and puddles rather than rain at the time of a ride—are important in determining a cyclist's rating of their ride. These coefficients, however, aren't nearly as enlightening as the time of day fits.

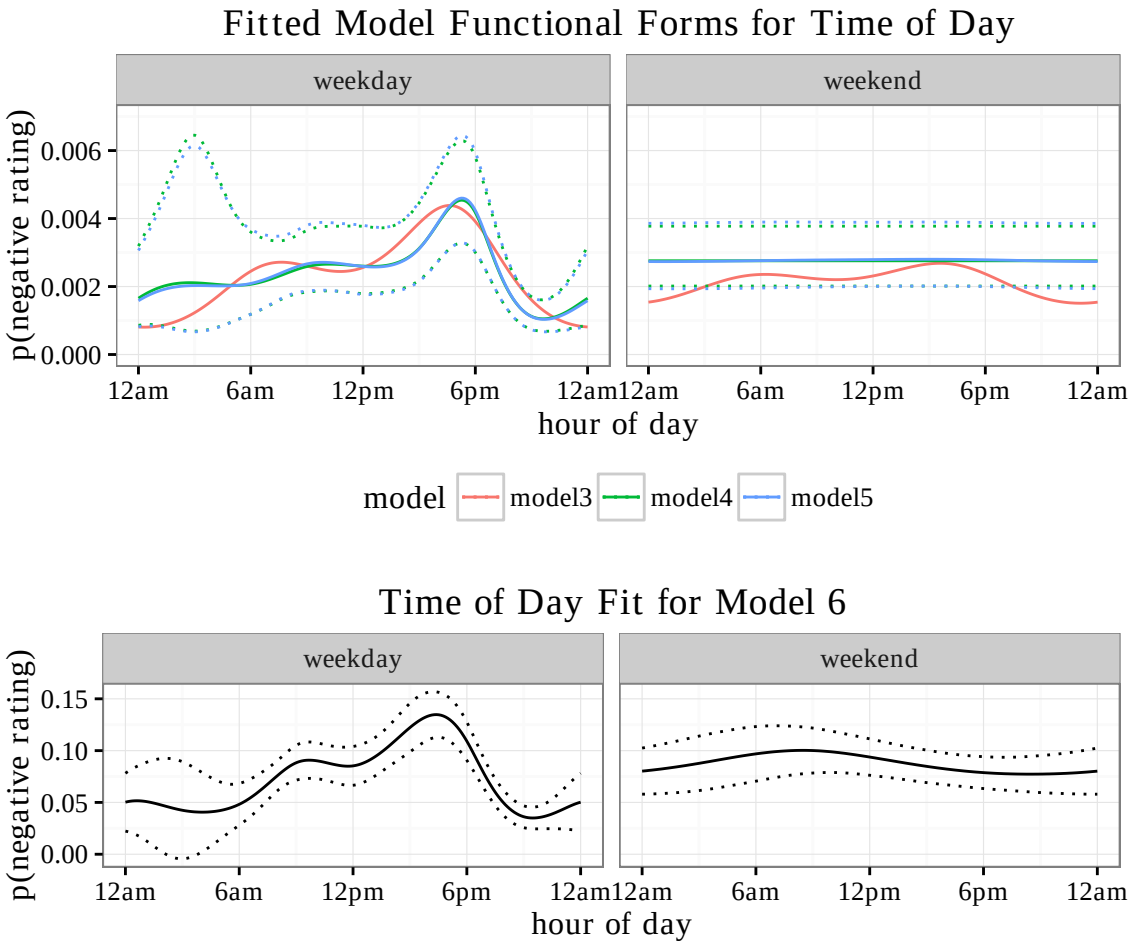

Figure 3: Predicted probabilities of a negative rating by time for a typical ride. The rider was chosen so the intercept was closest to the mean intercept for model 5. The median length and average mean temperature were used, and all other predictors were set to zero. The dotted lines show ±2 standard errors from the predictions.

The marginal fits for time of day, shown in Figure 3, are predictable. On weekdays, the probability of a negative rating peaks in the afternoon from 4–6 p.m., around when we expect rush hour traffic, and on weekends it stays steady throughout the day. While Model 4 and Model 5 give similar fits for time of day, Model 3's predictions peak at different times on weekdays and exhibit much more variability on weekends. There are two probable reasons for these differences: first, the sinusoidal terms are less flexible than the splines; second, the splines, because they are non-parametric functions, penalize complexity of the fit while the parametric sinusoidal form does not, making the splines more conservative in their "curviness." The former explains the discrepancies in the weekday fits while the latter explains the discrepancies in the weekend fits. Given these differences, fitting time of day with splines is preferable; there is no motivation to constrain the functional form to any strict parametric form.

But these marginal time-of-day fits don't just tell a story about our time terms; they also reveal part of why the random rider intercepts are such powerful predictors. Notice that in comparing Model 6 to the other models in Figure 3, the scale at which the Model 6 time fitted probabilities vary is much larger than the scale at which the other models' predictions vary. Without allowing for varying rider intercepts, the time terms take on a significant role. Yet, interestingly, the time term has nowhere near the amount of information that the rider intercepts seem to encode, according to the separation plots in Table 2.

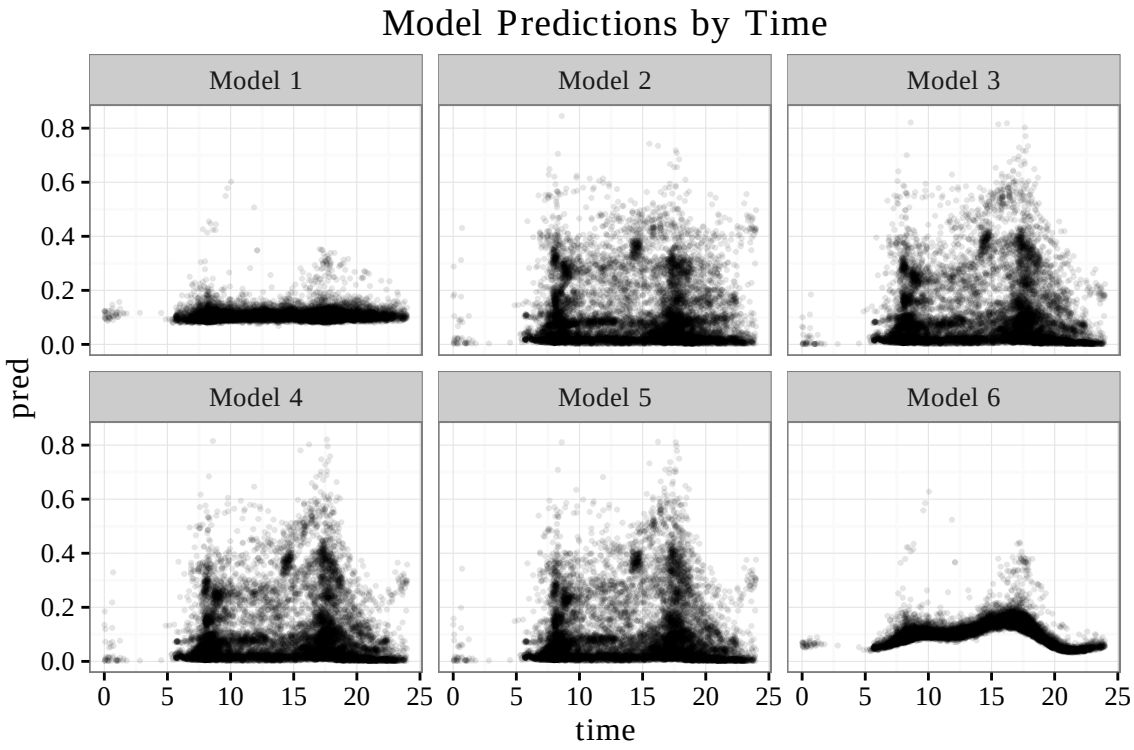

Figure 4: Predicted probabilities of negative rating by time of day for Models 3–6. Notice how starting with Model 2, daily trends start to emerge. This indicates that the rider intercepts are picking up on time of day trends, which must be reflected in a rider's typical ride.

Figure 4 paints a clearer picture of what is going on. These models show two different ways to look at the time of day pattern in ride rating: Model 2 suggests *who is riding* determines these patterns while Model 6 suggests there is something inherent in that time of day, such as traffic, that determines these patterns. The models between fit a combination of these, but as we saw in Figure 3, the time dependence is more than an order of magnitude weaker after accounting for the rider intercepts. The two black pillars of rides in the predictions of Models 2–5 line up with when we expect commuters to be riding, suggesting that that the riders with high rider intercepts are commuters. The converse, however, is not true: a great number of rides during commuting times are predicted to have almost zero probability of receiving a negative rating.

What explains this relationship between the temporal patterns and the rider intercepts? We suspect ride route is a confouding factor here. Figure 4 confirms our suspicion that riders have particular schedules they stick to; so it's also likely, given that these are mostly commuting cyclists, that most of their rides follow the same route as well. These models ignore ride route, so we suspect the typical ride route is encoded in the rider intercepts; *i.e.* a rider whose commuting route goes through many of the most stressful intersections and streets in Portland will likely have a higher intercept than most riders. This hypothesis can only be tested, however, when future research develops models with random rider intercepts and a model for how routes effect ratings.