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: List of models we evaluate in this chapter.

Model	**Description**
Model 1	(Baseline) Classical 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

The Models

Model 1, which we will use as the baseline for comparing further models, is a multiple logistic regression model. Then, our first model will be,

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

where $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$ are parameters to be estimated.

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:

$$\mathbb{P}(Y_i = 1) = \text{logit}^{-1}(\alpha + \alpha_{j[i]} + X_i\beta).$$

Rider intercepts themselves aren't as interesting as how they deviate from the mean, so we actually 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 the various daily trends that may affect ratings, including as a simple way to model the overall traffic level, which is difficult to model on it's own. These patterns are cyclic and very non-linear, which means we have to be a little more creative in how we incorporate them into our model. One approach is to add sinusoidal terms with a period of one day. We would be interested in fitting a term,

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

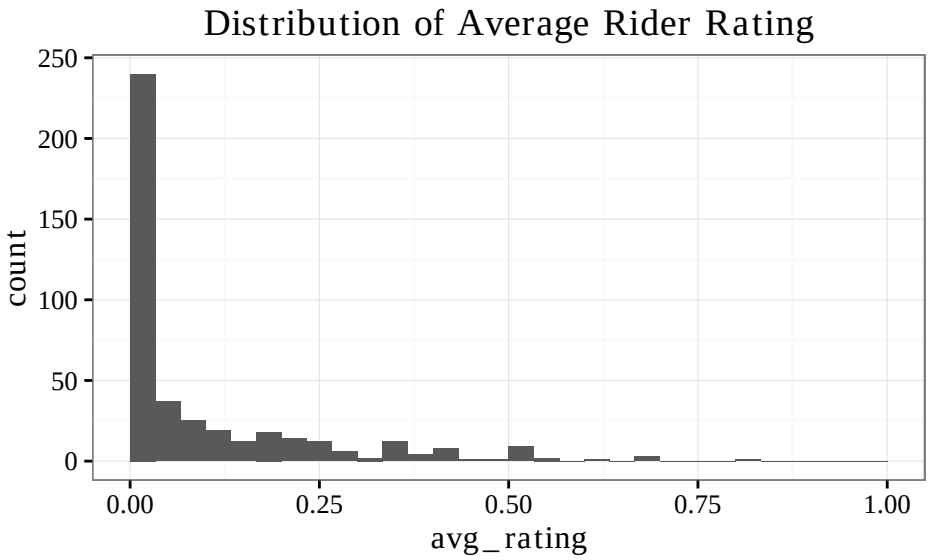


Figure 1: The overall rates at which each rider gives a negative rating for a ride varies greatly. This is our primary motivation for including rider intercepts and predictors.

Estimating β wouldn't be hard; we can easily estimate coefficents of tranformed terms. But it's more difficult to estimate T and ϕ. Except 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.

ϕ isn't actually that much of a problem, because we can rewrite this function as a sum of two trignometric terms 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 estimates generalized linear regressions.

We also want to take into account that weekday patterns may be different than weekend hourly patterns. We also have 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,

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

For **Model 4** and **Model 5**, we abandon parametric methods and use a cyclic non-parametric smoother to model time of day. The only problem is that we need a way to combine our parametric and multilevel parts of the model with a new non-parametric part. This is where additive models come in.

Additive Models and Smoothing Splines

We want to explore using non-parametric methods to model the relationship with time and length, but we wish to keep the other parts of our model. We can do this with an additive model. Additive models assume that the response is the sum of functions of each of the predictors:

$$\text{logit}(\mathbb{P}(y_i = 1)) = \alpha + \sum_{j=1}^{p} f_j(x_{ij}).$$

These functions can be linear, so generalized linear regression is a subset of additive models. But more interestingly, these functions can be non-parametric. [1] One of the most common types of functions fit are smoothing splines.

Smoothing splines are essentially cubic functions stitched together at points called "knots" such that the full piece-wise function is continuous and has continuous first and second derivative. One can further define cyclic cubic splines, which simply have the constraint that the last knot and first knot are treated as the same, thus allowing a continuous cyclic function to be fit. [2]

Computation of multilevel additive models with splines is available in the `gamm4` package, which we use to fit the two following models.

Model 4 will introduce a cyclic cubic splines for time of day.

Model Evaluation

To fit the data, we got all of the rides in Portland, OR from [insert data here] to [insert date here]for riders that had over 20 rides. There were 35,370 rides, 14,032 of which were rated. Overall, 10.88 percent of these rides were given a negative rating. There were 518 riders in the data set.

All six models were fit twice: first to the full data set to get good estimates of the parameters, and second to a sample of 80 percent of the rides. We used the latter fit to make predictions for the remaining 20 percent of rides, to determine the predictive accuracy for out-of-sample observations. (The sample had to be stratified by rider, to guarantee that every rider had at least on ride in the training set; otherwise models with random intercepts would be unable to make predictions for some rides.)

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 scores, 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? *i.e.* if riders were randomly assigned to rides, would 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. This quick test nullified this skepticism, as you can see in the resulting separation plots in Figure 2.

Table 2: Model fit summaries for full-data fittings.

Model	**Separation Plot**	$\log(\mathcal{L})$	**AIC**	**AUC$_{\text{CV}}$**[3]
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

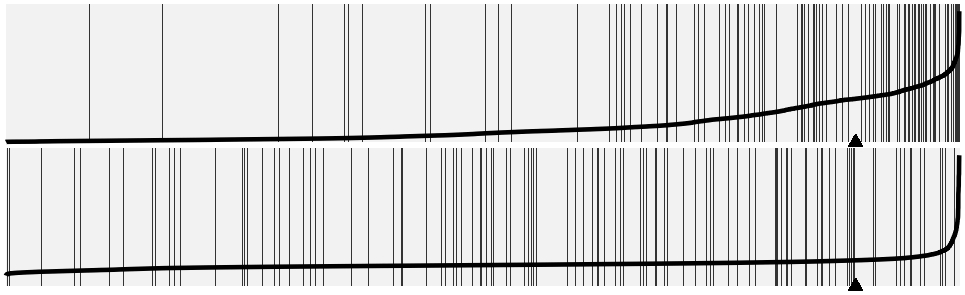


Figure 2: Separation plots for models 2 compared to a similar model where riders are randomly assigned to rides.

Model Results

Table 3 presents the fixed effect estimates for our models.

The marginal fits for time of day, shown in **??**, 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 flexibile 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 **??** to **??**, the scale at which the Model 6 time fitted probabilities vary is much larger than the scale at which the other models' predictions vary. (The differences is so large we had to show them in separate plots!) Without allowing

[1]How are these models fit? Using what's known as the Backfitting Algorithm. We define the kth partial residuals $Y^{(k)} = Y - \left(\alpha + \sum_{j \neq k} f_j(x_j)\right)$. (That is, define the portion of Y leftover for $f_k(x_k)$ to fit to after the other f_j's have had their share.) Then, iteratively fit each of the functions f_j on the partial residuals $Y^{(j)}$ until each of the functions converge. For a further quick look at additive models, check out Cosma Shalizi's lecture notes (@cosmaadditive)

[2]For a brief and entertaining introduction to smoothing splines, see @cosmasplines. For a more in-depth look at splines, check out @wood2006

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

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

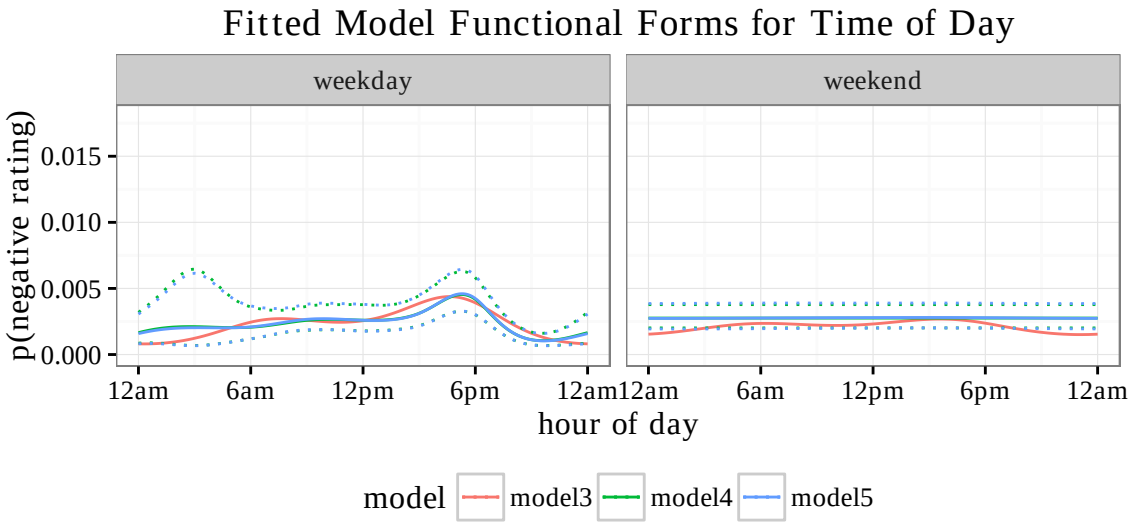
Regression Term	Model 1	Model 2	Model 4	Model 6
Log(length)	-0.122	-0.100	-0.092	-0.114
	(-0.180, -0.063)	(-0.168, -0.032)	(-0.162, -0.022)	(-0.174, -0.054)
Mean Temp.	0.053	0.076	0.075	0.069
	(-0.0004, 0.110)	(0.005, 0.147)	(0.003, 0.147)	(0.012, 0.127)
Mean Wind speed	0.028	0.014	0.012	0.027
	(0.004, 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)

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.

A clearer picture of what is going on is painted in Figure 4. The predictions of Model 1, which has a fixed intercept and no time dependence, don't vary by time of day. The models with random intercepts (2–5) show strong temporal patterns, with concentrated spikes in the morning and evening. Model 6, which had the time of day spline but a fixed intercept, has a temporal pattern, but does not represent the same degree time dependence that the random intercept models do.

One should be cautious about interpreting the intercepts. It's tempting to say they represent a riders general tendency to rate rides negatively, but this is ignoring their typical route. Just like how figure Figure 4 demonstrates that riders have unique patterns of what times of days they take rides, riders also likely have very particular routes that make up the majority of their rides. Because our models do not take into account the route, the intercepts are likely encoding the information about a riders typical route.

The fact that riders have apparent patterns of time of ride (and very likely route of ride), indicates there will be some value in trying to differentiate the types of riders that are present in this data. Indeed, one way of looking at the range of probabilities of giving a negative rating different riders have (demonstrated in figure Figure 5) is that some are more likely to offer useful information. In the next chapter, we will attempt to pick apart different groups of riders, and attempt to characterize their patterns of when and how long they ride.



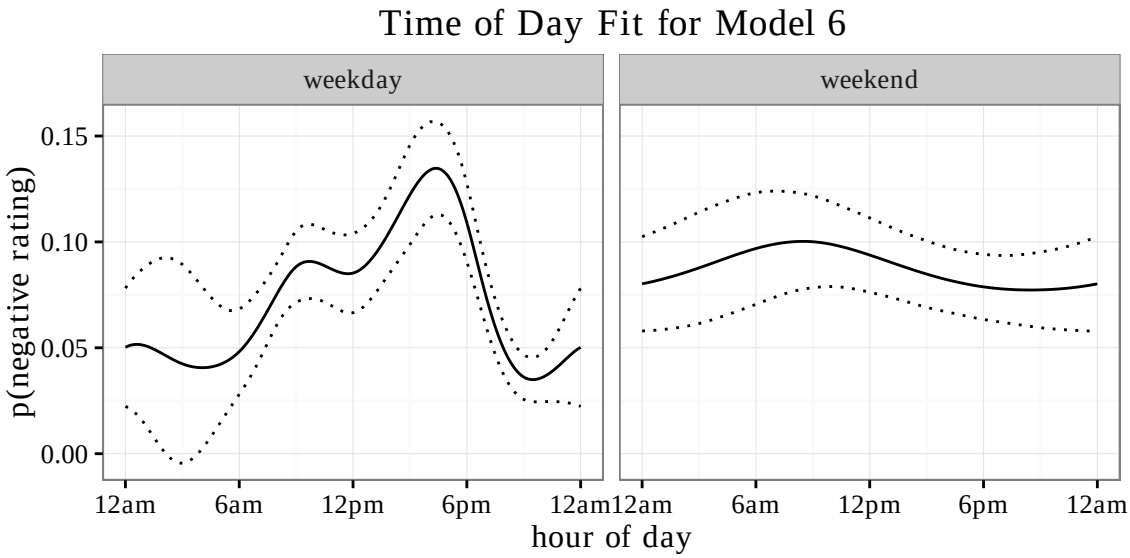


Figure 3: Predicted probabilities of a negative rating by time for a typical ride. The rider was chosen so the intercept was closest to the mean intercept for model 5. The median length and average mean temperature were used, and all other predictors were set to zero. The dotted lines show ±2 standard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

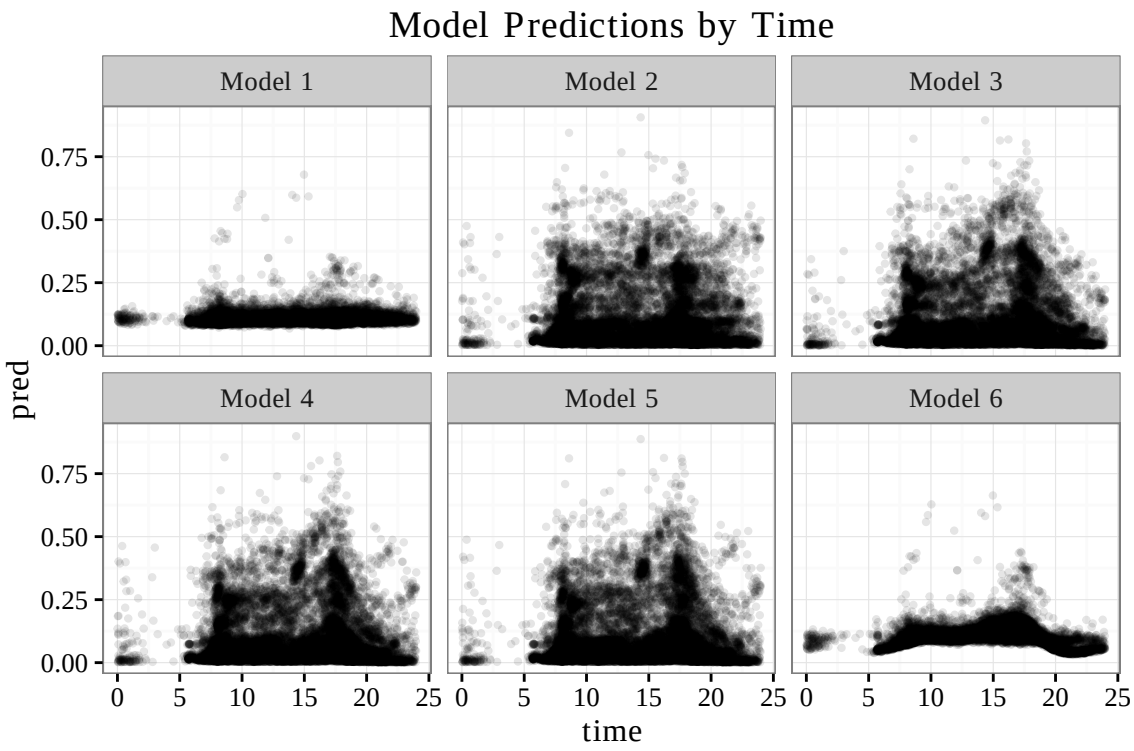


Figure 4: Each of the models predictions for all rides with time of day on the x-axis. Notice how starting with model 2, daily trends start to emerge. This indicates that the rider intercepts are picking up on time of day trends, which must be reflected in riders typical ride.

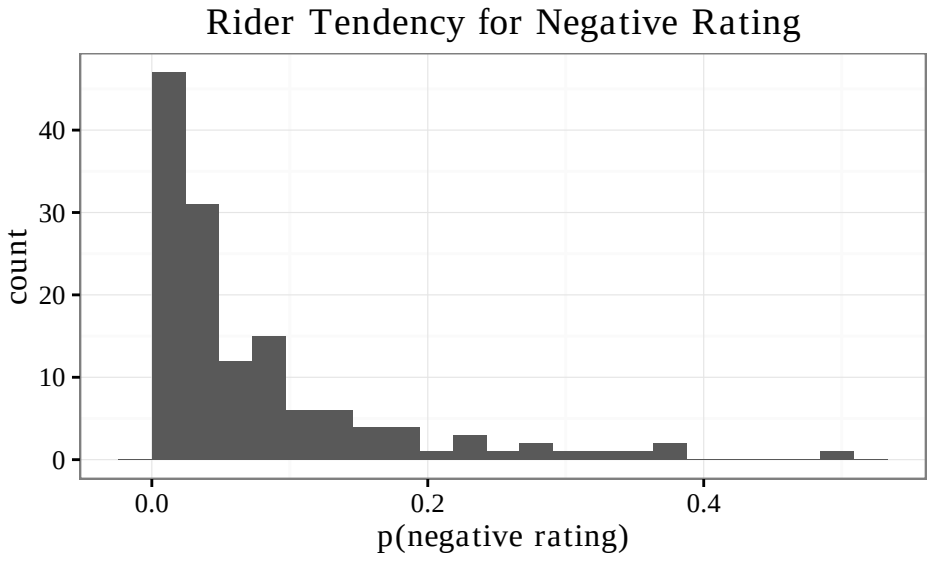


Figure 5: Predicted probability of a negative rating for a typical ride for each rider, for Model 2 and Model 4. The typical ride is a ride at noon on a weekday with median length, mean temperature, and other variables at zero.