ISYE-6420: Predicting Bike Rentals Using Bayesian Regression

Zachary Burns

December 3, 2023

Abstract

In this paper, we explore the use of various Bayesian regression analysis methods such as Linear Multivariate Regression and Gaussian Random Walk methods to predict bike rentals in Washington D.C using the Capital Bikeshare data set.

1 Introduction

Bike sharing services have gained considerable traction across North American in recent years and have become a common method of transportation for people commuting to and from work. Heavy street traffic in busy cities and a desire for an environmentally friendly form of transportation make biking an attractive alternative to traveling by car. In this paper, we examine the Capital Bikeshare program using daily data [1] collected from Washington D.C. We explore several approaches utilizing Bayesian methods for predicting bike rentals on a daily basis over the span of 2 years as well as evaluate and compare the performance of each model.

2 Data

The data set contains daily counts of bike rentals (which includes both registered and casual users) between January 1, 2011 and December 31, 2012 in the Capital Bikeshare system along with corresponding weather and seasonal information. The daily data contains 731 rows and 16 columns. However, through some initial exploratory investigation, I only selected the following six features: (1) Temperature, (2) Humidity, (3) Windspeed, (4) Holiday or Not, (5) Working Day or Not, and (6) Weather Situation. I removed five of the features due to high correlation between these six selected features. For example, temperature and air temperature are nearly perfectly correlated and does not add additional information to the model. The other removed features are described in the Data Dictionary section of the notebook I included with my submission. The latter 3 features (holiday, working day and weather situation) are categorical features with the former two being binary features and weather situation described as a nominal variable. The weather situation feature has been discretized into 3 weather patterns described by: 1) Clear, Few clouds, Partly cloudy, 2) Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist, and 3) Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds. A summary of the data used in this study can be seen in Table 1 below.

2.1 Exploratory Data Analysis

Before we explain the model structures, let's first take a look at the features included:

	intercept	holiday	workingday	weathersit	temp	hum	windspeed	season	rentals
count	731.00	731.00	731.00	731.00	731.00	731.00	731.00	731.00	731.00
mean	1.00	0.03	0.68	1.40	0.50	0.63	0.19	2.50	4504.35
std	0.00	0.17	0.47	0.54	0.18	0.14	0.08	1.11	1937.21
min	1.00	0.00	0.00	1.00	0.06	0.00	0.02	1.00	22.00
25%	1.00	0.00	0.00	1.00	0.34	0.52	0.13	2.00	3152.00
50%	1.00	0.00	1.00	1.00	0.50	0.63	0.18	3.00	4548.00
75%	1.00	0.00	1.00	2.00	0.66	0.73	0.23	3.00	5956.00
max	1.00	1.00	1.00	3.00	0.86	0.97	0.51	4.00	8714.00

Table 1: Summary statistics for bike sharing data which contains 731 rows collected from Capital Bikeshare in Washington D.C

1. Temperature is provided as a normalized temperature in Celsius. The values are derived via $(t - t_{min})/(t_{max} - t_{min})$ where $t_{min} = -8, t_{max} = +39$. The temperature is provided as the average daily temperature. In our exploratory data analysis section in the provided notebook, we can see a direct relationship between daily temperature and the number of bike rentals.

2. Humidity is also provided as a normalized value (values have been divided by 100). Humidity seems to have less of an impact on bike rentals given the relatively stationary distribution of the feature. Washington D.C is typically quite mild given it's relative physical location and as a result has above average humidity.

3. Windspeed has also been normalized by dividing values by 67. Windspeed doesn't seem to have a cyclical nature (varying from season to season) or a clear connection to the number of bike rentals. We can notice from Table 1 that the mean windspeed is relatively low (0.19). My intuition says that unless the weather is extremely windy, this feature will have minimal impact on the number of bike rentals.

4. Working Day is a binary feature which is 0 if the date is a weekend or a holiday and 1 if a weekday and not a holiday. We can see a pretty even distribution among these two values indicating no clear relationship between working days and the number of bikes rented.

5. Holiday is a binary feature which indicates the presence of a holiday (when equal to 1). We can see a positive relationship between the number of bikes rented during non-holidays likely due to users commuting to work.

6. Weather situation shows a much more direct relationship between nicer weather (weathersit equal to 1) and an increase in the number of bikes rented which matches our intuition.

3 Approach

To simplify, I started with an extremely simple model using temperature as a single covariate with an intercept. We then develop a slightly more complex model using the remaining chosen predictors to see if we can lift predictive performance using additional covariates. Finally, we experiment with using a Gaussian Random Walk model using temperature as a time-varying coefficient.

4 Bayesian Regression Analysis

4.1 Bayesian Linear Regression: Baseline Model

First, we create a simple baseline model with a single predictor variable (temperature) and an intercept. The model structure is defined as:

$$\begin{aligned} \alpha &\sim N(0,1) \\ \beta_1 &\sim N(0.44, 0.2) \\ \nu &\sim Ga(2, 0.1) \\ \sigma &\sim \text{Half Normal}(\sigma = 1) \\ \mu_i &\sim \alpha + \beta_0 \cdot X_i \end{aligned} \tag{1}$$

When choosing the model structure, I decided to use more informative priors. As I was experimenting, I researched average daily temperature for Washington, D.C which was approximately 12.67 degrees Celsius [2] which when scaled using the Min-Max scaling method results in a value of ~ 0.44. This seems reasonable given a fairly mild climate with slightly high variance. I chose to include an intercept as well because there does seem to be a positive linear trend over time with respect to the number of bikes being rented on a given day. As for ν, I selected a Gamma prior for the degrees of freedom parameter (ν) in the likelihood because we saw from the data that the number of bike rentals can fluctuate significantly day-to-day. As a result, I chose a low degrees of freedom parameter (ν) resulting in heavier tails, as suggested by [3] to determine if this allows our model to be more robust to outliers. In practice, I found this didn't substantially impact the model in any meaningful way when compared to experimenting with higher degrees of freedom values. Additionally, I chose to use a Student's t-distribution over the typical Standard Normal Distribution for this use case because according to [4], the Student's t-distribution is often used to model the errors in the data when the residuals are not normally distributed and have heavier tails. The Student's t-distribution can be more robust to outliers than the normal distribution, which makes it a better choice when the data has extreme values. In my analysis, I found the number of bike rentals can fluctuate significantly depending on the season, so using this approach seemed reasonable. Note that I didn't include this baseline model with the Normal Likelihood in this analysis for brevity but can be found in my notebook. When running the same model structure with a Normal likelihood, I found the results to be almost identical in both regression coefficient values and model fit. This is likely because the Student's t-distribution is a generalization of the the normal distribution as $\nu \to \infty$.

	mean	sd	hdi 2.5%	hdi 97.5%	mcse mean	mcse sd	ess bulk	ess tail	r hat
intercept	0.13	0.02	0.10	0.17	0.00	0.00	20991.85	23310.37	1.00
β_{temp}	0.77	0.04	0.70	0.83	0.00	0.00	20779.66	23287.66	1.00
ν	10.52	1.74	7.29	13.98	0.01	0.01	26722.97	26427.23	1.00
σ	0.16	0.01	0.15	0.17	0.00	0.00	26790.74	25262.86	1.00

Table 2: Base Model with a single predictor (Temperature)

We can see in Table 2 that there does seem to be a slight positive trend indicated from the intercept coefficient (~ 0.13). Further, Table 2 also shows the 95% HDI credible set which is quite narrow indicating confidence in the coefficient value nor does it contain 0. The coefficient value for β_{temp} is ~ 0.77 with a relatively tight 95% HDI credible set. This indicates the model is able to interpret a reasonably strong relationship between temperature and our outcome variable. This aligns with my expectations noted above where users are more likely to rent bikes for transportation when the weather is nicer.

Using this model structure, we can see the posterior predictive fit seems reasonable in Figure 1. We notice that our posterior mean seems to over predict our data in summer 2011 and mirrors the observed data reasonably well going into winter 2012. However, we also observe increased variability in our data for summer 2012 which our model does not handle well. Further, Appendix A gives another perspective on how well our model fits the data by illustrating the mean response and also highlights the non-linear relationship between temperature and bike rentals.

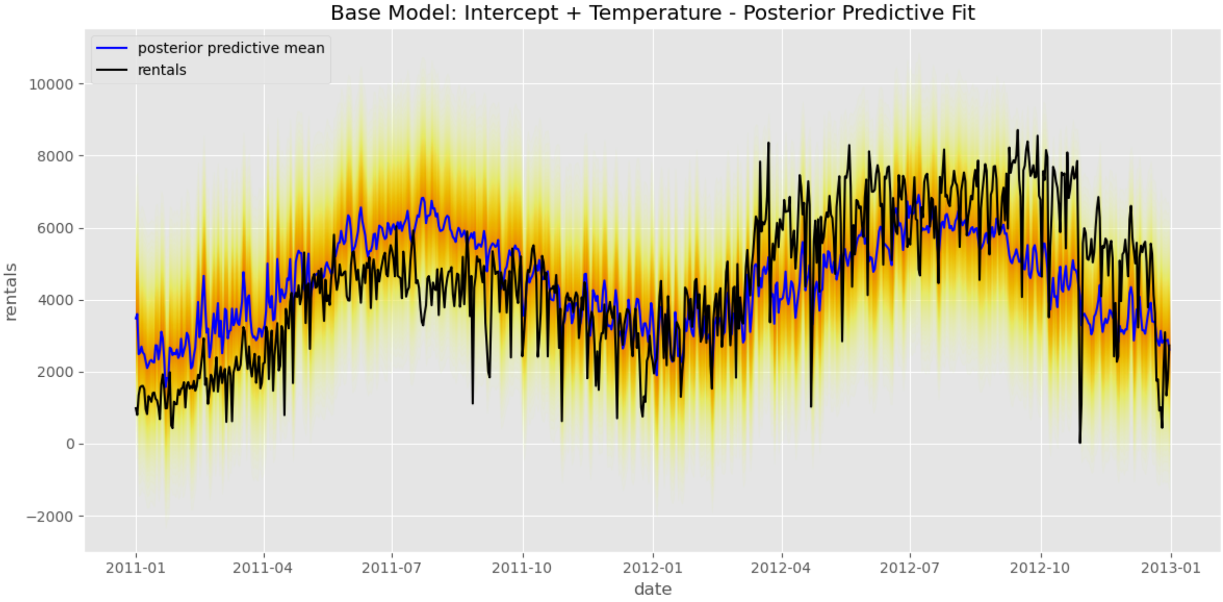

Figure 1: Base Model Posterior Predictive Fit

4.2 Bayesian Linear Regression: Full Model

Now, we introduce a slightly more complex model which includes the following five additional features: 1) humidity, 2) windspeed, 3) workingday, 4) holiday, and 5) weathersit. The model structure is defined as:

$$\begin{aligned}
\alpha &\sim N(0,1) \\
\beta_{temp} &\sim N(0.44, 0.2) \\
\beta_{hum} &\sim N(0.5, 0.1) \\
\beta_{windspeed} &\sim N(0.2, 0.2) \\
\beta_{categorical} &\sim ZSN(\sigma = 1) \\
\nu &\sim Ga(2, 0.1) \\
\sigma &\sim \text{Half Normal}(\sigma = 1) \\
\mu_i &\sim \alpha + \sum_{k=1}^{6} \beta_k \cdot X_{ik}
\end{aligned} \tag{2}$$

When structuring this model, my main experiments were focused on the priors for the categorical variables ($\beta_{categorical}$). During my research, I found three different approaches for choosing priors on the categorical predictors which were: 1) pm.Categorical, 2) pm.ZeroSumNormal, and 3) One-Hot-Encoding with uninformative normal priors. My analysis quickly revealed that option 1 wasn't suitable and resulted in extremely high (or NaN) r_{hat} values. Supposedly this is because PyMC doesn't interact well with categorical features when mixing continuous variables with the pm.Categorical distribution which is discrete. Using pm.ZeroSumNormal (ZSN) seemed to provide the most reasonable and intuitive coefficient values which provides one value for each unique value in the corresponding feature. With the Zero Sum Normal approach, we notice that the coefficient values for each categorical variable sum to zero. This suggests counter acting effects for each value the binary features could take on. However, we notice that since the weathersit feature has more than two categories, this doesn't hold but the coefficients still sum to zero. Using the normal uninformative priors, the interpretation is that each unique value for a given feature contributes differently to the outcome variable.

As for the priors on humidity and windspeed, I took a similar approach as with temperature. I researched average windspeed [5] and humidity [2] and scaled the calculated averages (average values by month) to derive our priors.

Figure 2 highlights the posterior predictive fit using the above model structure which seems to suffer from similar problems as our baseline model. At first glance, we can see the model generally over predicts in 2011 while under predicting in 2012. This is likely due to our data being non-linear and so as temperature crosses a threshold in the summer (i.e. when it is too hot) we witness a decrease in demand rather than an expected increase based on our temperature coefficient. Further, the model summary can be seen Appendix B. When looking at parameter estimates, we can see that both $\beta_{holiday}$ and $\beta_{workingday}$ 95% credible sets (HDI) contain 0 while the $\beta_{weathersit}$ coefficients do not.

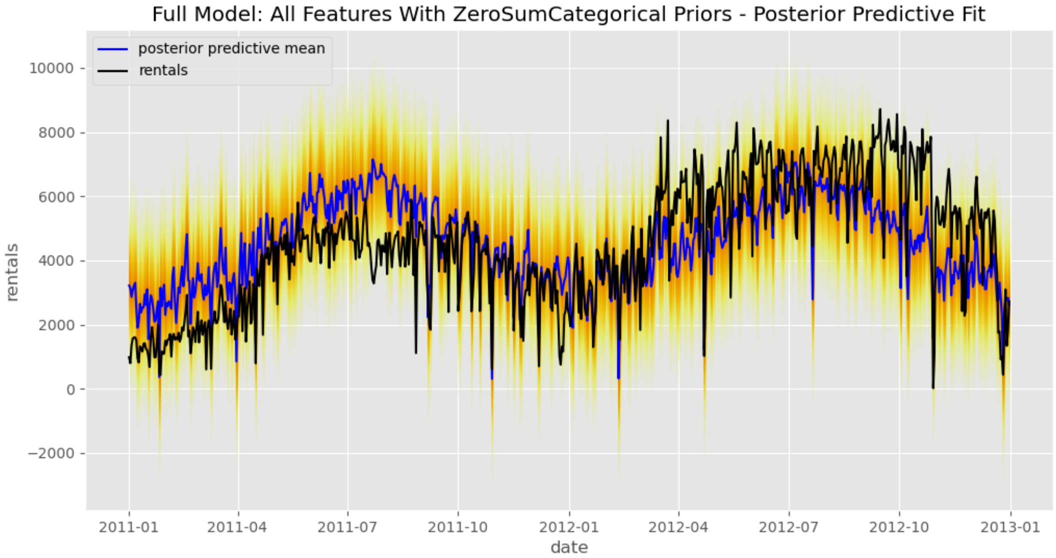

Figure 2: Full Model Posterior Predictive Fit using ZeroSumNormal priors

4.3 Gaussian Random Walk

For the final model, I studied a Gaussian Random Walk process where we allow the temperature coefficient to vary over time. The main idea here is that in a time-series model, rather than assuming that a parameter has a constant impact on the outcome variable over time, we allow it to change gradually. By taking this approach, we are assuming that our data is not independent and identically distributed and thus is changing over time which requires learnable parameters for every observation:

$$Y_i = f(\beta_i X_i) \tag{3}$$

In the context of bike rentals, this seems like a reasonable assumption to make given that the temperature seems like more of a deterrent when extreme values are observed, hot or cold. I structured this model by creating one temperature coefficient for every observation in our data. I also attempted (for quite a long time) to create a second Gaussian Random Walk model which contains 12 temperature coefficients, one for each month. My reasoning here is that one coefficient for every day is likely overkill and almost certainly over-fits the data. However, I couldn't get PyMC to cooperate when creating these monthly coefficients on daily data unfortunately. As noted above, the temperature varies significantly between portions of the year and some months seem to have a larger impact on the number of bike rentals on a given day than others.

The model structure is the same as the linear models created above with the only difference being in how we structure the temperature coefficient which is described as:

$$\beta_t = N(\beta_{t-1}, \sigma^2) \tag{4}$$

The main choice here in terms of prior selection was focused on σ. This represents the step size and constrains how large of a step β_t can vary from β_{t-1}. The daily temperature differences can be seen in the EDA section of the notebook. Since this parameter represents the daily difference, we can see that most of the time the values lie in the +/- 0.1 range which is how we selected our σ value for the step size. Further, we choose Normal distributions for both σ and the initial distribution of the Gaussian Random Walk because temperature should be able to fluctuate in both directions.

Figure 3 highlights the posterior predictive fit for our Gaussian Random Walk model. At first glance, we can see the model clearly has a better fit to the data than our other models noted above. Additionally, Figure 4 illustrates how our temperature coefficient varies over time. Interestingly, there seems to be three to four major change points in the temperature coefficient suggesting that one parameter value for each season seems like a reasonable approach. We can also see that temperatures above 0 degrees Celsius contribute positively to demand and temperatures above 20 degrees have a diminished impact on demand which is interesting. This could potentially suggest that a high number of users are commuting to and from work, so hot weather deters them to avoid reaching the office sweaty.

5 Model Results

To evaluate the performance of the above models, I decided to compare them using Deviance, R^2 and the Laud-Ibrahim critierion in order to see if they all agree on which model is best. First, Table 3 illustrates the Deviance scores using Pareto smoothed importance sampling leave-one-out cross-validation. We immediately notice that the Gaussian Random Walk model has a significantly lower Deviance score and is clearly the preferred model according to the *weight* metric. Further, Table 4 highlights the R^2 scores for each regression model. Again we see that the Gaussian Random Walk

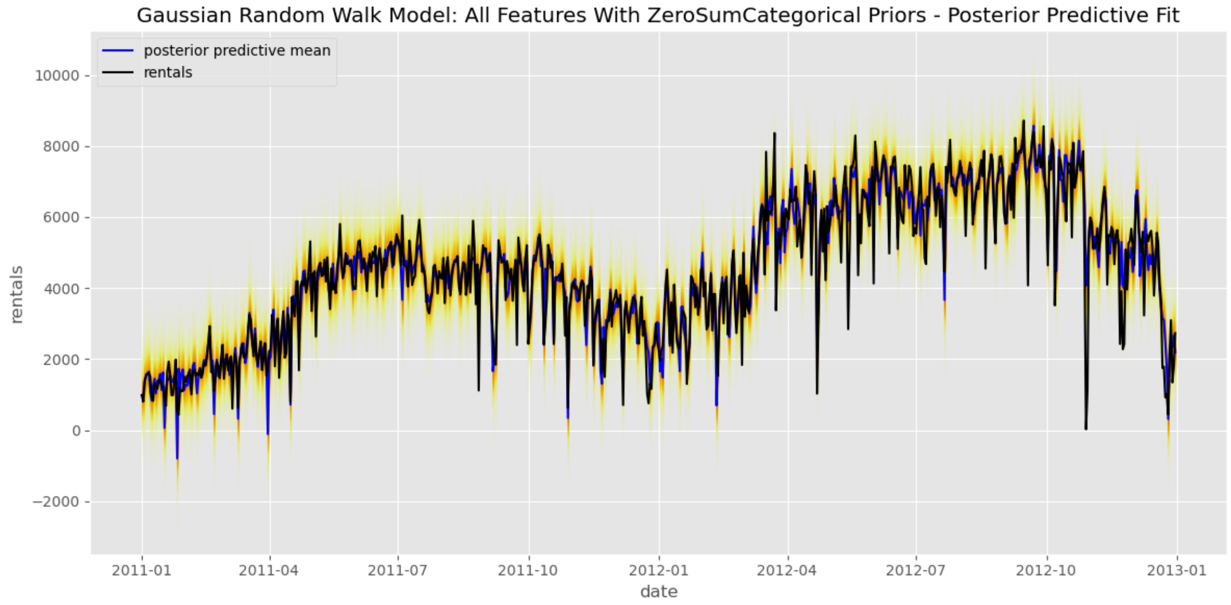

Figure 3: Gaussian Random Walk Posterior Predictive Fit using ZeroSumNormal priors

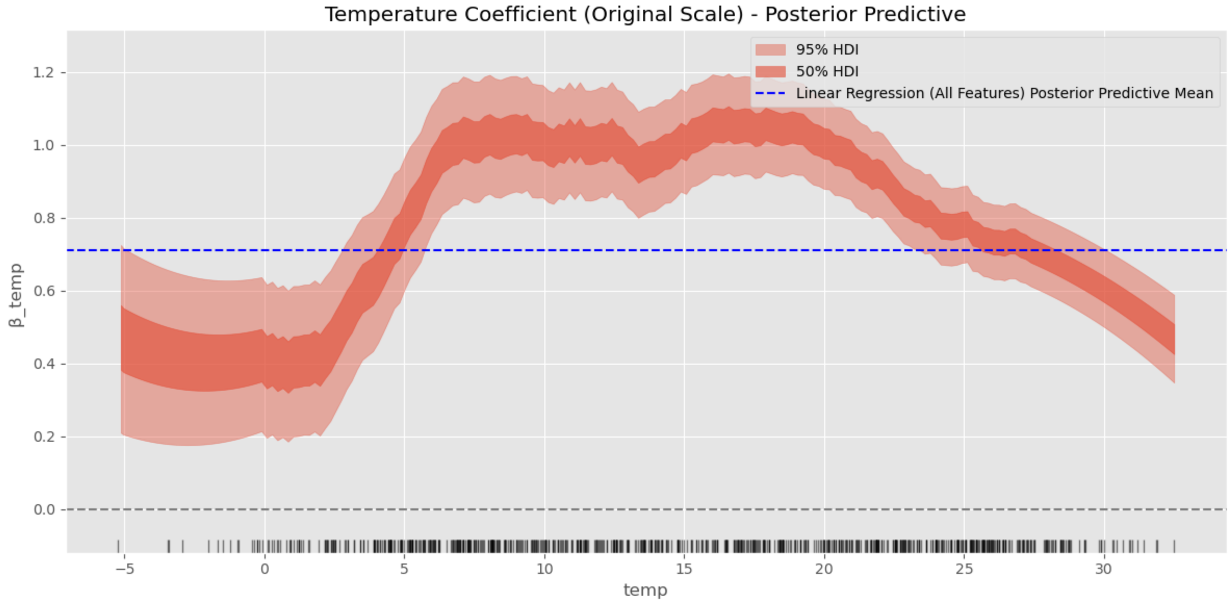

Figure 4: Gaussian Random Walk Model: Temperature Coefficient Changing Over Time

model seems to explain much more of the variance in the data than the other models studied. This is not surprising given the fact that our data appears non-linear when looking at the relationship between temperature and bike rentals. Lastly, Figure 5 which shows the L-Weighted measure scores also suggests that the Gaussian Random Walk model fits the data much better. Interestingly, there is a very small difference between our baseline (using only temperature) and the full regression models further suggesting that temperature is the most important parameter by far.

	rank	elpd_loo	p_loo	elpd_diff	weight	se	dse	warning	scale
gaussian_rw	0	-1650.641	219.564	0.000	0.959	56.161	0.000	False	deviance
full_model_norm	1	-571.490	7.552	1079.152	0.041	32.156	59.340	False	deviance
full_model_zsn	2	-571.236	7.581	1079.406	0.000	32.142	59.324	False	deviance
baseline_normal	3	-481.438	2.537	1169.203	0.000	32.504	58.349	False	deviance
baseline_student_t	4	-458.049	2.673	1192.592	0.000	33.419	59.216	False	deviance

Table 3: Model Comparison: Deviances

	baseline_student_t	baseline_normal	full_model_norm	full_model_zsn	gaussian_rw
R2 Score	0.455	0.450	0.486	0.486	0.803

Table 4: R^2 Scores for each model examined in this study

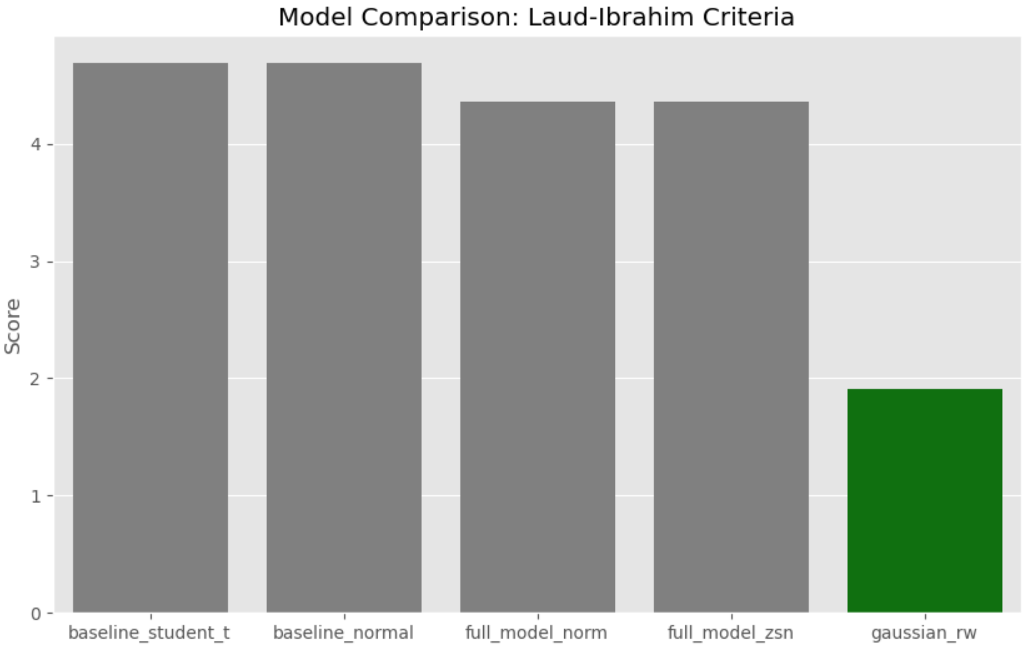

Figure 5: Model comparison using the Laud-Ibrahim method

6 Future Work

There are several potential improvements that could be studied to see if we can improve model performance. One possibility would be to decompose the data into casual and registered riders. I believe that additional domain knowledge could be embedded into our above models to create a better fit on the data. My intuition says that these two groups of users come from different distributions and modelling as one single distribution isn't necessarily the right approach. Further, there is an hourly version of this data set which could provide more insight into different patterns and structures within the data. I would also like to take a look decomposing the data into trend, seasonality and day-to-day differences. This approach is a more traditional time series forecasting technique, but seems like an interesting avenue to explore. Additionally, we can see from our exploratory analysis that the data seems to be non-linear and as a result, our linear models explored above can only take us so far. Some possible options would be to look at other Gaussian Process models to model the different components of the data such as seasonality, trend and the different types of users (registered and casual). Additionally, I attempted a Negative Binomial regression, but I couldn't find a suitable model structure that gave reasonable results. The results of this can be seen at the end of my notebook. Additionally, we saw from the analysis that outliers seem to be present in the dataset. Identifying and removing some of these could help improve performance. Lastly, more extensive feature engineering could be performed to address the non-linearity in our data.

References

[1] Fanaee-T, Hadi. Bike Sharing Dataset. UCI Machine Learning Repository, 2013. DOI.org (Datacite), https://doi.org/10.24432/C5W894.

[2] Washington, D.C. Climate, Weather By Month, Average Temperature (Washington, D.C.; United States) - Weather Spark. https://weatherspark.com/y/20957/Average-Weather-in-Washington-D.C.;-United-States-Year-Round. Accessed 24 Nov. 2023.

[3] Juárez, Miguel A., and Mark F. J. Steel. 'Non-Gaussian Dynamic Bayesian Modelling for Panel Data'. Journal of Applied Econometrics, vol. 25, no. 7, Nov. 2010, pp. 1128–54. DOI.org (Crossref), https://doi.org/10.1002/jae.1113.

[4] Fonseca, Thaís C. O., et al. 'Objective Bayesian Analysis for the Student-T Regression Model'. Biometrika, vol. 95, no. 2, 2008, pp. 325–33. JSTOR, https://www.jstor.org/stable/20441467.

[5] Climate: Washington, D.C. in the United States. Worlddata.Info, https://www.worlddata.info/america/usa/climate-washington-d-c.php. Accessed 25 Nov. 2023.

Appendix A Baseline Linear Regression Model: Model Fit

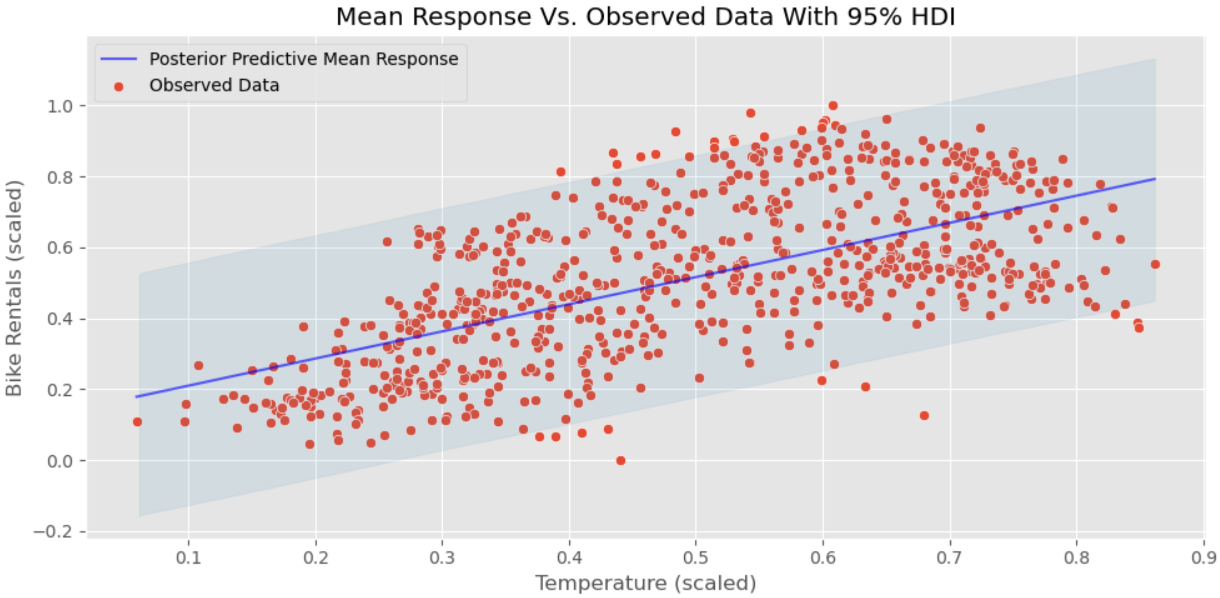

Figure 6: Plotting the posterior predictive mean against observed data and the 95% HDI (in blue)

Appendix B Multivariate Linear Regression: All Features

	mean	sd	hdi_2.5%	hdi_97.5%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
intercept	0.11	0.05	0.02	0.21	0.00	0.00	24358.68	26350.49	1.00
beta_temp	0.71	0.03	0.65	0.78	0.00	0.00	45202.44	28620.69	1.00
beta_hum	-0.03	0.05	-0.12	0.07	0.00	0.00	31758.06	29824.46	1.00
beta_windspeed	-0.29	0.08	-0.44	-0.14	0.00	0.00	39393.95	30517.05	1.00
beta_holiday[0]	0.03	0.02	-0.00	0.07	0.00	0.00	43967.06	28920.65	1.00
beta_holiday[1]	-0.03	0.02	-0.07	0.00	0.00	0.00	43967.06	28920.65	1.00
beta_workingday[0]	-0.01	0.01	-0.02	0.01	0.00	0.00	52162.21	29467.68	1.00
beta_workingday[1]	0.01	0.01	-0.01	0.02	0.00	0.00	52162.21	29467.68	1.00
beta_weathersit[2]	0.05	0.01	0.02	0.08	0.00	0.00	37836.00	31485.80	1.00
beta_weathersit[1]	0.11	0.02	0.08	0.15	0.00	0.00	26606.16	29729.51	1.00
beta_weathersit[3]	-0.16	0.03	-0.21	-0.11	0.00	0.00	27604.75	26454.21	1.00
nu	48.23	17.75	18.83	83.39	0.08	0.06	51046.92	29173.24	1.00
sigma	0.16	0.00	0.15	0.17	0.00	0.00	46896.26	28663.95	1.00

Table 5: Full Model with all predictors and ZeroSumNormal categorical priors