Modeling the Count of Rental Bikes in the Greater Washington D.C. Metropolitan Area

Gianni Spiga, B.S.
University of California Davis

ABSTRACT

Modern bike sharing services provide data on daily users, including the number of bikes that are rented each day, which this analysis attempts to build an accurate model for. The data sourced from the UCI Machine Learning Repository is cleaned, visualized, and used as the foundation of multiple count regression models. Model reduction and variable encoding lead to a final set of predictors to best describe the relationship between weather behavior, time of year, seasons, and the number of rented bicycles for any given day.

INTRODUCTION

- In the greater Washington, D.C. metropolitan area, including Virginia and Maryland, a common method of transportation for both residents and visitors is bike sharing, with over 700 stations and 5,400 bikes to date.
- The data is mapped by the Laboratory of Artificial Intelligence and Decision Support (LIAAD) at the University of Porto, and made accessible by the UCI Machine Learning Repository.
- Our data on bicyclists' behavior is provided by the company Capital Bikeshare, the sharing system in the D.C. area. Weather data is provided by i-weather.com involving humidity and forecast. Lastly, holiday and working day schedules are provided by the D.C. Department of Human Resources.
- Recordings are everyday between 1/1/2011 and 12/31/2012, leaving us with 731 continuous days of information.
- **Goal**: Build an accurate model for the counts of bikes regressed on weather conditions, time, and calendar events.

Column	Description
instant	Record Index
dteday	Date
season	Season (1:winter, 2:spring, 3:summer, 4:fall)
yr	Sear (0: 2011, 1:2012)
mnth	Month (1 to 12)
hr	Hour (0 to 23)
holiday	Whether Day is Holiday or Not
weekday	Day of the Week
workingday	If Day is Neither Weekend nor Holiday is 1, Otherwise is 0.
weathersit	Weather Conditions[1]
temp	Normalized temperature in Celsius.
atemp	Normalized feeling temperature in Celsius.
hum	Normalized humidity. The values are divided to 100 (max)
windspeed	Normalized wind speed. The values are divided to 67 (max)
casual	count of casual users
registered	count of registered users
cnt	count of total rental bikes including both casual and registered

METHODOLOGY

- Data cleaning requires dropping highly correlated columns, indexing columns, and factor transformation, leaving a remaining nine predictors.
- The first approach is using the Poisson model, the classical model for count data; however this requires that the variance of response is approximately equal to the mean, in other words, no *overdispersion.*
- However, building a Poisson model for the data has an dispersion value of 165.35.
- Negative Binomial model, which adjusts for overdispersion, has promising results. The AIC is 10 times smaller than the Poisson model as well as a dispersion of only 1.06.
- Performing a forward and backward stepwise regression, in hopes for a minimized BIC, leaves us with predictors *season, year, holiday, weather conditions, temperature, humidity,* and *wind speed.*
- Following this, we aim to fit the data better by adding interactions. We perform a backward stepwise regression with only two-way interactions since three-way interactions would be computationally expensive and challenging to interpret.
- Doing this, we keep all main effects with the addition of interactions season with temperature, year with temperature, weather conditions with humidity, and weather condition with wind speed. This gives us the following model:

$$\begin{aligned}E(Y|X) = \beta_0+\beta_1X_{season2}+\beta_2X_{season3}+\beta_3X_{season4}+\beta_4X_{yr1}+\beta_5X_{holiday1}+\beta_6X_{weathersit2}\\ +\beta_7X_{temp}+\beta_8X_{hum}+\beta_9X_{windspeed}+\beta_10X_{season2:temp}+\beta_11X_{season3:temp}+\beta_12X_{season4:temp}\\ +\beta_13X_{yr1:temp}+\beta_14X_{weathersit2:hum}+\beta_15X_{weathersit2:windspeed} \quad (1)\end{aligned}$$

- Both Likelihood Ratio Test and plots of Deviance Residuals vs Fitted Values show the interaction model is the better fit.
- The dataset overall has 3 large outliers, observation 65, 668, and 725. While the reasoning for observation 65 being an outlier an unclear, we have clear evidence for the other two, Hurricane Sandy and Christmas, causing much lower than expected counts.
- We drop observation 65 since it has high leverage and build a final model.

RESULTS

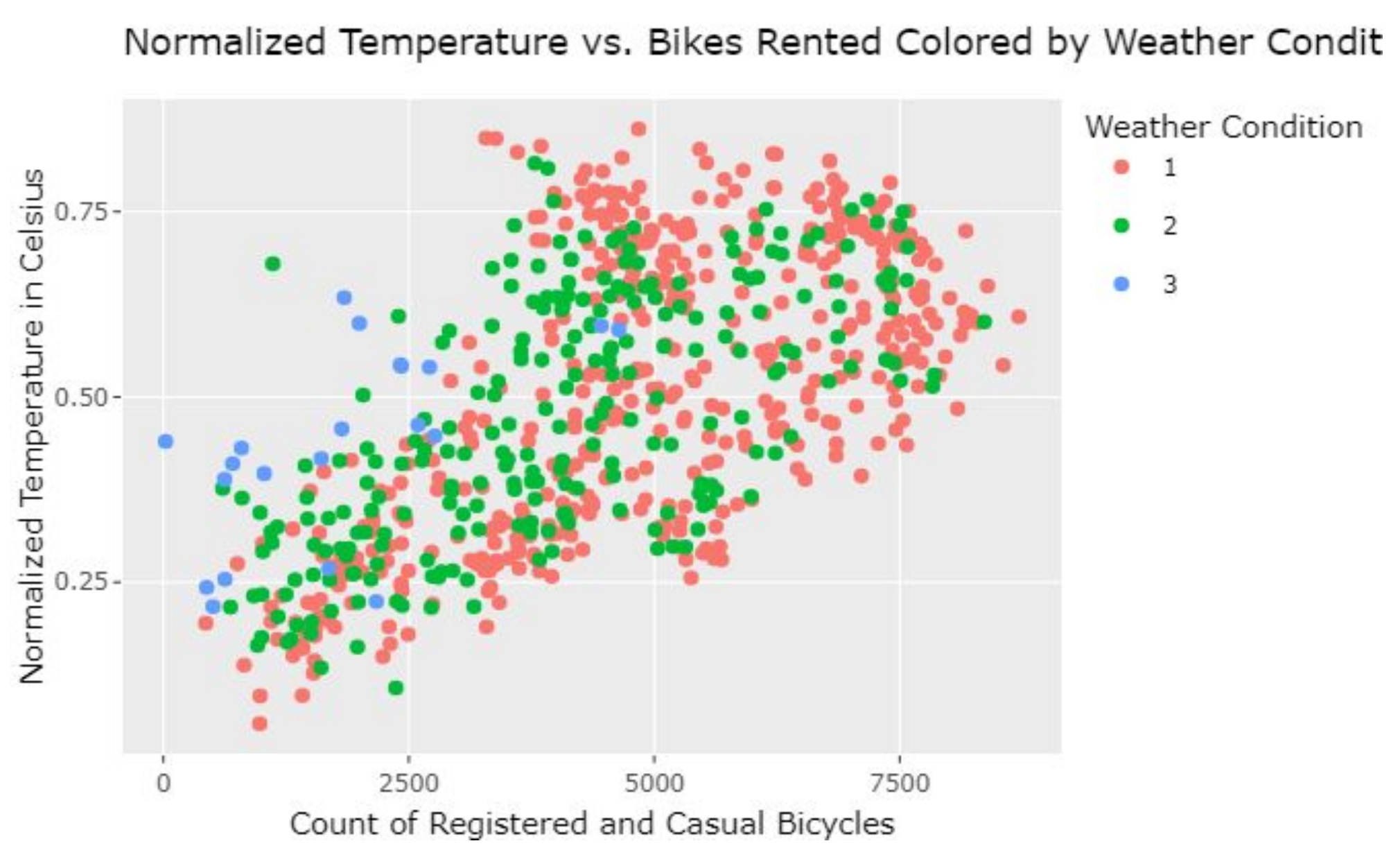
	Coefficient	95% CI Lower	95% CI Upper
(Intercept)	6.89	6.71	7.08
season2	0.85	0.66	1.03
season3	2.29	1.93	2.65
season4	0.83	0.65	1.00
yr1	0.64	0.54	0.74
holiday1	-0.19	-0.29	-0.09
weathersit2	0.74	0.49	0.99
temp	3.02	2.67	3.36
hum	-0.24	-0.46	-0.02
windspeed	-0.46	-0.78	-0.15
season2:temp	-1.50	-1.94	-1.07
season3:temp	-3.71	-4.29	-3.12
season4:temp	-1.21	-1.68	-0.74
yr1:temp	-0.35	-0.53	-0.16
weathersit2:hum	-0.95	-1.26	-0.63
weathersit2:windspeed	-1.14	-1.62	-0.66

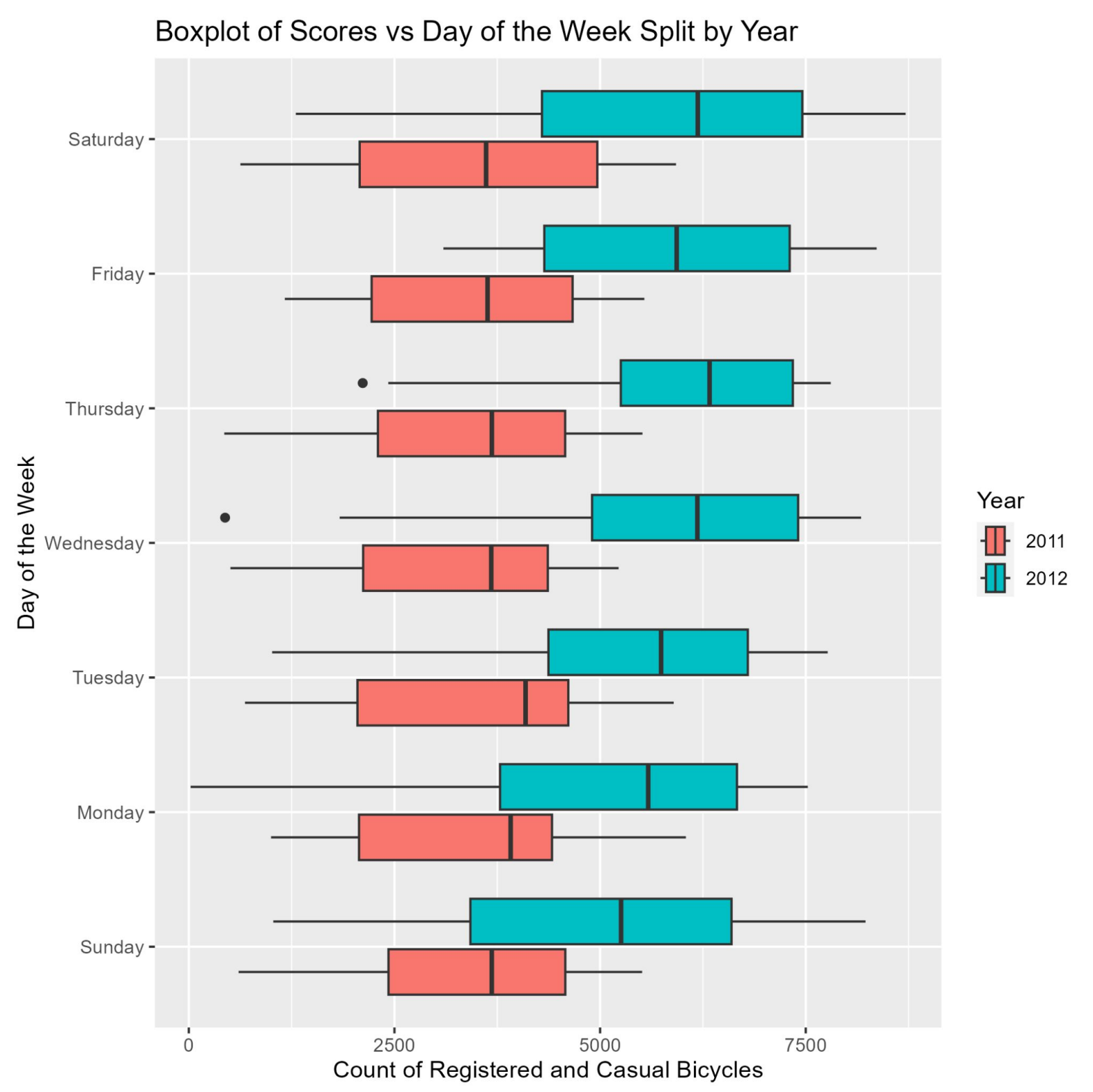
- The table above displays all the calculated coefficients with respective 95% confidence intervals for the final model.
- We can see that season and temperature have the greatest positive effect on bikes counted, however, their interaction will have a negative effect.
- Whether we are modeling 2011 or 2012 has a significant difference, as more people rode in 2012 on average (see boxplot in Recommendations).
- Humidity and whether or not the given day is a holiday decrease the expected count, however, their effect is not large.

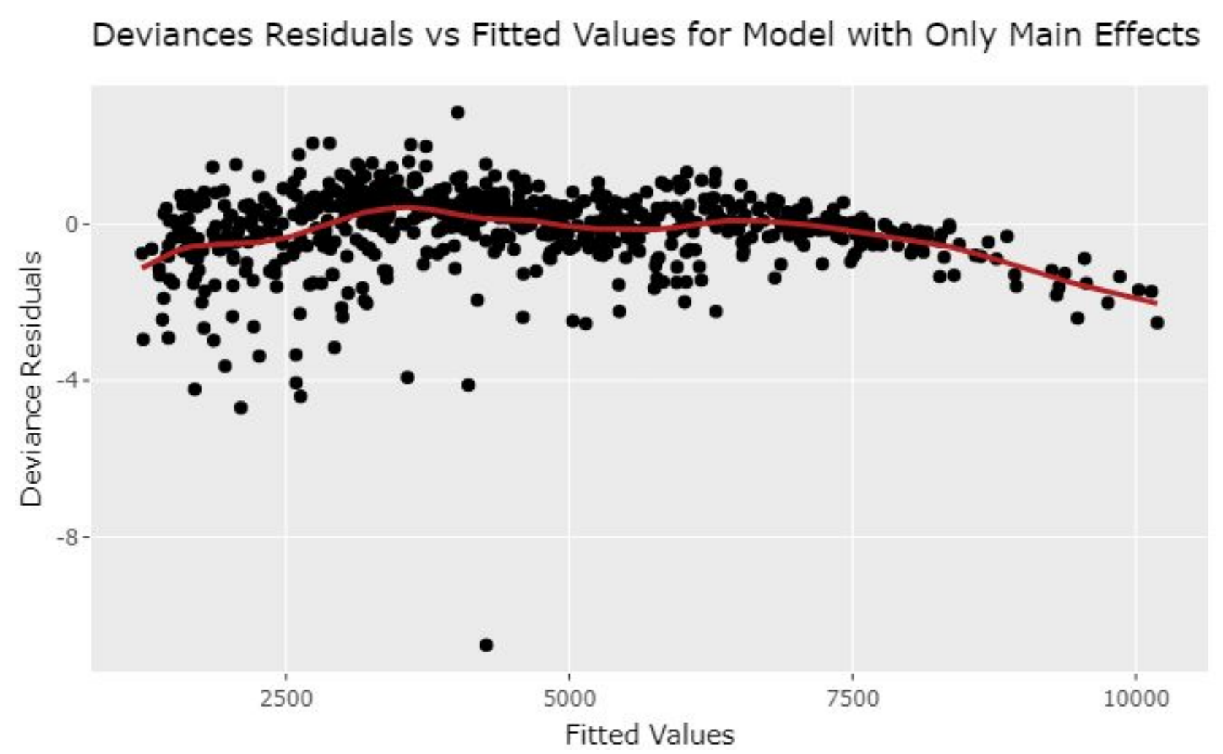
CONCLUSION

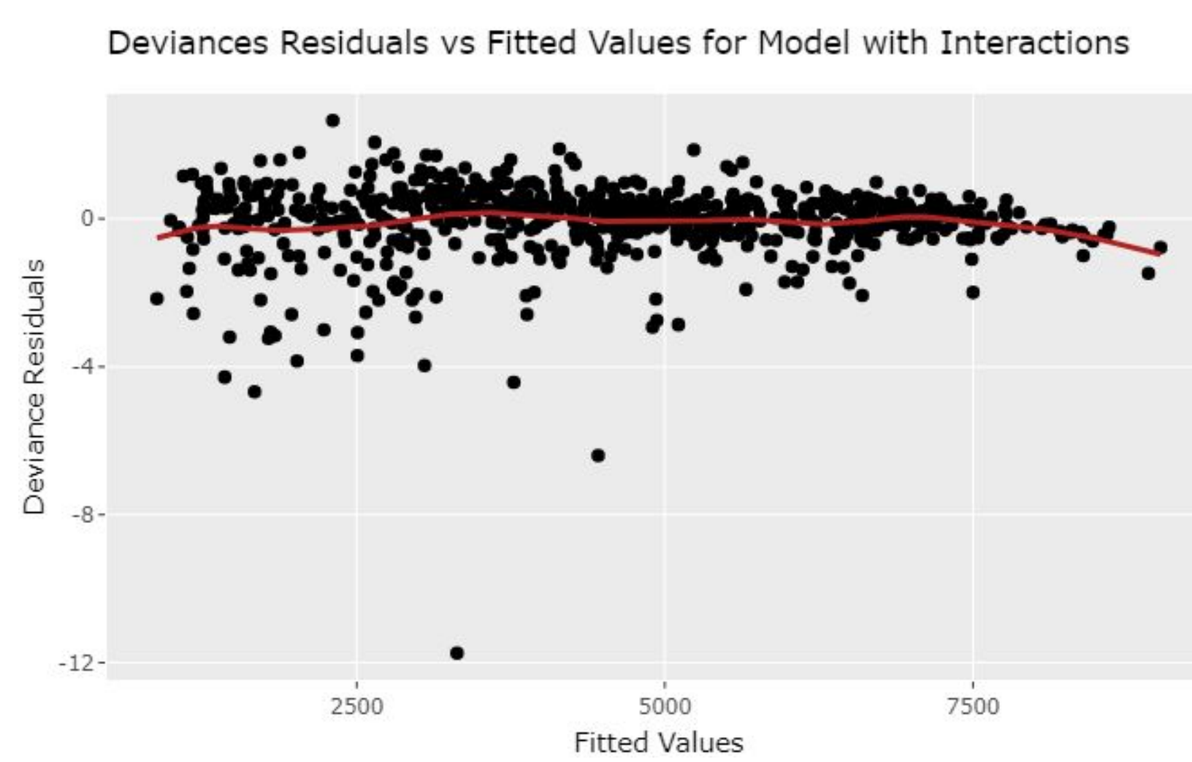
- As the amount of shared bikes increases on the road, it is important for the number of bikes to be modeled accurately.
- In our analysis, we found that the best model was a Negative-Binomial model, which accounted for the over-dispersion of the response.
- We found predictors which involve the season, year, holidays, weather conditions, temperature, humidity wind speed along with multiple interactions involving these variables accurately modeled the count of bikes on the road on any given day.

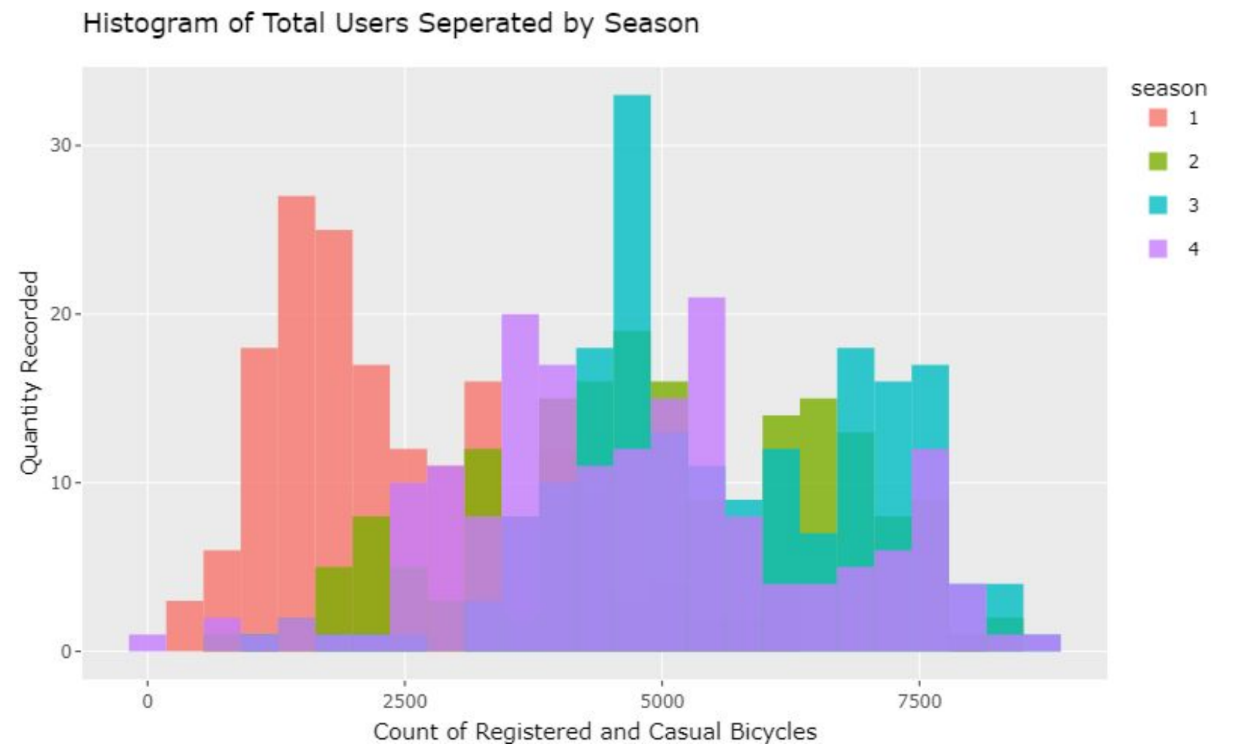
RECOMMENDATIONS











ACKNOWLEDGEMENTS

Thank you to Prof. Hans-Georg Müller and Han Chen for their guidance and assistance in this analysis. I would also like to acknowledge the Laboratory of Artificial Intelligence and Decision Support at the University of Porto as well as the UCI Machine Learning Repository for making this analysis possible by providing the data.