Summary of "Modeling bike counts in a bike-sharing system considering the effect of weather conditions"

This paper identifies a method to quantify the effect of weather conditions on bike sharing counts in San Francisco Bay area with the aim of improving bike sharing systems given their wide benefits such as decreasing transportation pollution and increasing mobility efficiency in cities, to name a few. The benefit of this model specifically will be to decrease the environmental costs and time consumption and other complications associated with the rebalancing operation of bikes between stations; which is important to ensure that each station has enough number of bikes to satisfy the demand, especially given the limited number of docks at station.

The methodology used to create the bike count model includes several steps. First the effect of various variables is quantified (month of the year, day of the week, time of the day and different weather conditions), then these predictors were ranked by Random Forest technique and were used to predict a regression model using a guided forwarded step-wise regression. More than one model was created then the Bayesian information criterion was used to evaluate the models. In the first step, the count models employed generalized linear models, specifically two models were used Poisson, which condition to apply is that the mean and variance must be equal, and negative binomial regression, which uses same condition as Poisson except that another parameter is involved that loosens the initial condition and adjusts the variance independently, so it is considered to accommodate more dispersion. The second step after that is to apply machine learning to avoid overfitting of predictors, using Random Forrest method. The RF method randomly constructs a group of trees, where each tree is a subset of features, so trees are not correlated, then the ranking of features is obtained based on majority of votes from all trees, after that forward step-wise regression is applied, and finally a model is selected BIC after computing the log-likelihood of each model, and the model of the lowest BIC is to be selected.

This methodology was applied on a dataset of bike-sharing for San Francisco Bay Area between August 2013 to August 2015, where incidents were documented every minute for 70 stations in the area, which led to a large dataset, and another dataset was used which included weather conditions during these 2 years, and it included the following attributes: "date (in month/day/year format), ZIP code, temperature, humidity, dew level, sea level pressure, visibility, wind speed and direction, precipitation, cloud cover, and weather description for that day (i.e., rainy, foggy or sunny)."
For the first step, the histogram of new counts frequency for all stations showed dispersion, which gave a hint on better fitness of NBRM. However, both PRM and NBRM were applied at first to generate a full model of all available predictors. Then RF was used to rank the predictors in the full model based on the OOB error. Forward stepwise regression was then used to fit several models that were constructed by RF, then BIC was applied to select the best subset of predictors to construct this model.

At first it was assumed that there is no interaction between the 70 stations, to quantify effects fast and effectively and in an attempt to create one model for all variables rather than a model for each station, and this approach was described to satisfy the level of accuracy needed. And the results showed a logarithmic mean of bike counts at each station following parallel hyperplanes, which shows no interaction between stations. And in order to construct one model instead of 70 models, one for each station: 69 indicators were used with one reference, a similar approach was used for months of year with 11 indicators and January as a reference, and 6 indicators for the days of the week, and so on for all data attributes. If there was no significant difference between each pair of parameters, for example

between 2 stations, it was assumed bike count was the same for the two stations to an acceptable level of accuracy.
The results showed that different stations, month-of-the-year, day-of-the-week, and time-of-the-day were all shown to influence the model. And the following weather attributes were selected for additional exploration: mean temperature, mean humidity, mean visibility, mean wind speed, precipitation, and weather description. And for the second time, RF and forward step-wise regression were employed and the resulting models were compared by BIC. And the model with the trade-off between the minimum BIC value and the consideration of the effective parameters was selected. And as was shown because of dispersion, indeed NBRM was shown to be better than PRM, and it was selected for the rest of the modeling steps.

Among 111 models created, the results showed that bike counts are significantly influenced by the month-of-the-year, day-of-the-week, time-of-the-day, and some weather variables, mainly temperature and humidity level, which is also dependent on geographic location. And the most significant variables affecting bike counts are available number of bikes at time t-1 and the time-of-the-day.

This paper holds many strengths in meeting our project, first the factors used in this paper match our dataset, and the paper justifies the use of these factors among others used in other studies. And this paper was the first to study the effect of humidity which was shown to have a significant effect on the model. The methodology and tools used match our set of expertise, which is to be developed through this course, and the tools and methods used were all explained and justified. This paper can be used as a reference for us to build our model and compare the results given different factors, especially difference in geographic location, while also contributing to this research with the insights we obtain.

However, a few weaknesses of this paper were detected. First, the paper targets only docked bike-sharing model and it's applicable to only certain geographic areas with certain weather conditions, as for another location, different parameters might be additionally considered. The final stages of refining the model and detecting possible errors was merely systematic and rather relied on observations and experts opinions, so in addition to the scientific and sequenced steps, there was some subjectivity in the methodology when it comes to certain decisions like the number of trees in RF and the selected factors for each tree and their number which wasn't discussed and explained enough, so it might be hard for us to follow the same methodology at these stages. The paper only used Poisson and negative binomial models at the first stage and didn't attempt more complex distributions. And the paper chose the final model prioritizing simplicity, designated by a smaller number of predictors, among the last two proposed models, without explicitly comparing their levels of accuracy. And some standard values of comparisons for example minimum log-likelihood and typical BIC measures weren't shared in the paper.

Reference:

Ashqar, H. I., Elhenawy, M., Rakha, H. A., Road, V. P., & Qld, K. G. (2019). Case Studies on Transport Policy Modeling bike counts in a bike-sharing system considering the e ff ect of weather conditions. *Case Studies on Transport Policy*, *7*(2), 261–268. https://doi.org/10.1016/j.cstp.2019.02.011