Predicting Appointment No Shows based on Available Patient Data

Rachel Holman, Benjamin Nikolai, Reanna Panagides

Problem Description

When a patient doesn't show up, or "No Show's", to a medical appointment, numerous negative consequences may arise, including wasted resources, profit, and time in the scheduling process, leading to delayed medical care that could possibly impact health outcomes (Liu et al., 2022). In order to mitigate the adverse effects of these occurrences for clinics and patients, it becomes crucial to predict patients with a higher probability of missing appointments. Identifying patients prone to "no-shows" could allow for tailored interventions such as more frequent reminders, assigning a social worker to coordinate transportation, or allocating home-health resources depending on individual needs and circumstances. These additional actions taken to prevent 'no-shows' could optimize clinic schedules, expanding care to more patients and reducing disparities among subgroups of patients.

Data Description

This project utilizes the "No-show Appointments Dataset" found on Kaggle which is a comprehensive collection of medical appointment records aiming to investigate factors influencing no-shows for scheduled appointments. The dataset, obtained from the healthcare system of a city in Brazil, contains information regarding appointments scheduled by patients including various demographic details and health-related attributes. Appointments in this dataset are limited to specifically 2016, and the key attributes of the file are listed in the appendix.

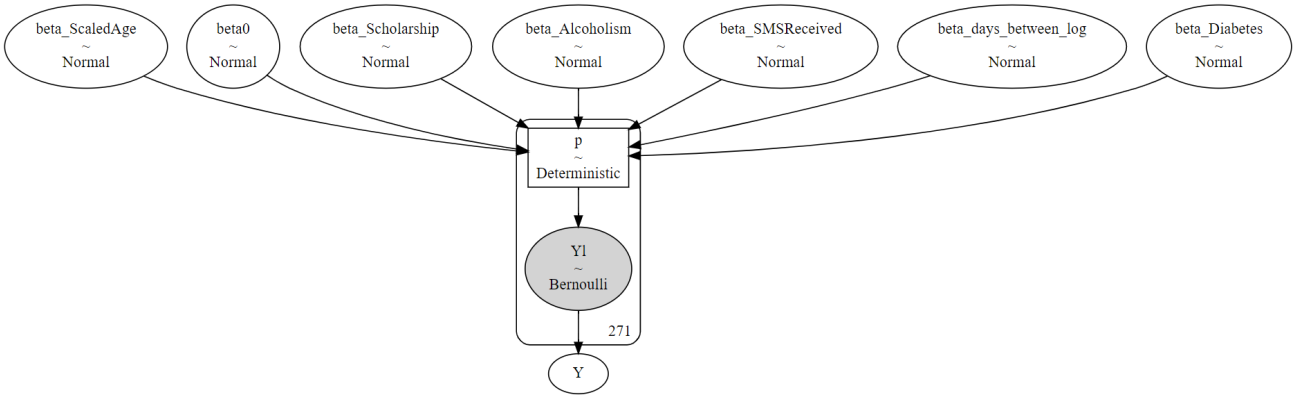
Before using this data for any modeling, we cleaned and preprocessed the variables in various ways. Firstly, a few of the variable names were misspelled ("Handcap" instead of "Handicap" for example) so we began by correcting these. Additionally, although there was no missing data, we did opt to remove one row which contained a negative value for age. Next we removed the ScheduledDay and AppointmentDay DateTime variables after computing the number of days between scheduling and appointment as a new variable. We then log transformed that new variable (called days_between) to non-linearly represent its relationship to the response variable. Finally, we saved the day of the week for appointments as a separate variable called app_weekday, scaled the age variable, and changed our response variable Noshow into a 0/1 binary rather than "No/Yes".

Probability Model

In order to identify patient "no-shows", we first used a hierarchical approach to model our data because we felt that it accounted for the grouped nature of the data collected based on the 'Neighborhood' of each clinic. There are 81 different neighborhoods represented in our data which all have different demographics and population sizes. Hierarchical models with partial pooling are built to account for the differences between groups such as these and would allow for neighborhoods with fewer observations to utilize a hyperprior distribution. Our predictors, however, did not align well with the hierarchical format and presented us with many divergences because there were not enough categorical variables to nest appropriately. After realizing this, we switched gears and determined that non-hierarchical generalized linear models (GLMs) would be the best fit for our data. Utilizing GLMs, we created both a full model that included all of the predictors in our dataset except for 'Neighborhood', and a reduced model that only included significant predictors. Finally, we opted for

under-sampling techniques to rebalance our outcome classes and generated our best model shown below.

Figure 1. Final Reduced GLM



Approach

Hierarchical Model:

To begin implementing our model, we split efforts into creating both our hierarchical model and standard GLM. Our hierarchical model was designed to include hyperpriors for each of our 81 different Neighborhoods as this allowed us to treat each Neighborhood as a separate variable while also allowing us to acknowledge some "underlying similarity" with their hyperpriors. We created multiple iterations of a varying slope and intercept Hierarchical model.

$$ln(\frac{P_{NoShow}}{1-P_{NoShow}}) = \alpha_{j,i} + \beta_{j,i}x_i + \epsilon_i$$

With our intercepts: $\alpha_{j,i} \sim N(\mu_{\alpha}, \sigma^2_{\alpha})$

Design matrix: $\beta_{j,i} \sim N(\mu_{k,\beta}, \sigma^2_{k,\beta})$ Where k represents a hyperprior for each individual predictor.

And finally our error term: $\epsilon_i \sim N(0, \sigma^2_{\epsilon})$

We include subscripts j to represent the differing slopes and intercepts we make per Neighborhood $(j = 1, 2, \ldots, 81)$ in $\alpha_{j,i}$ $\beta_{j,i}$.

During our model building phase, we tested different hyperpriors to use for the distributions for our intercept and for each β for our predictors in our design matrix. These specifically included the testing of setting our σ^2_{α}, $\sigma^2_{k,\beta}$ and σ^2_{ϵ} to Half Cauchy, Student T or Exponential. Attempts were also made on changing the number of predictors K in our design matrix.

Generalized Linear Models:

After realizing that our hierarchical model would not converge with the data we had, which mostly consists of binary coded columns with a single continuous, we decided to move on to building full and reduced generalized linear models. We chose a non-hierarchical GLM because it allows us

to create a linear relationship between our predictors and our non-continuous response variable. This particular GLM took the form of a generic logistic regression problem. We followed this formula:

$$ln(\frac{P_{NoShow}}{1-P_{NoShow}}) = \alpha + \beta x_i$$

With our intercept: $\alpha \sim N(0, 1)$

And design matrix: $\beta \sim N(\vec{0}_k, \vec{1}_k)$ with k $0's$ and k $1's$, k being the number of predictors.

Full and Reduced Generalized Linear Models:

In our full GLM, we included all of our predictors except for 'Neighborhood'. After running our full GLM, significant predictors of "no-show" included: Scaled-Age, Scholarship, Diabetes, Alcoholism, SMSreceived, and Days-Between. Predictors that were not significant in the presence of other predictors included Gender, Hypertension, Handicap, and App_Weekday. We included the significant predictors from our full GLM in our reduced GLM. Finally, we compared the within-sample performance using WAIC and the out-of-sample performance using testing and training datasets to determine that the reduced model performed the best.

Under-Sampled, Reduced Generalized Linear Model:

In order to further improve performance of our model, we utilized an under-sampling technique that rebalances unequal outcome classes within our data. Prior to under-sampling, we had 88,208 patients who showed up for their appointments and 22,319 who did not show up for their appointment. We under-sampled by removing a random selection of observations to allow for more equal distributions of the outcome classes.

Log-Transformed, Under-Sampled, Reduce Generalized Linear Model:

Our final effort to improve model performance consisted of replacing days_between with a transformed version to account for its non-linear relationship with the response. This ultimate model contained the following predictors: Scaled-Age, Scholarship, Diabetes, Alcoholism, SMSreceived, and log(Days-Between).

Results

To test our models, we used Frequentist and Bayesian metrics to compare the validity of their outputs for our problem. For strict classification evaluation, we used metrics like the ROC curve and its AUC value. We also paid close attention to the model's Precision score as we wanted to emphasize its ability to predict true positives, or patients who would not show up to appointments. To measure uncertainty we used Bayesian metrics like highest density interval and posterior predictive checks plots to observe the distributions of each parameter and see how well they fit our data.

Our initial full model performed rather poorly, generating predictions with an AUC value of 0.503 and a precision score of .0152. This essentially tells us that our model is no better than a classifier that predicts at random. Looking at the plots of each parameter's posterior, we noticed a wide range of different variances for each distribution. While this model does not perform well, this uncertainty measurement does help inform our next steps of the model building process.

Next, we reduced our model keeping only the predictors found to be significant. This simple model had an AUC score of 0.504 and a precision of .0152 which showed little to no improvement to

our full model. Similarly, its posterior plots showed a large difference in variances between each parameter's distributions; this proved to us that modeling with predictors that had no transformation performed on them would not yield our desired result. By undersampling the data used to train this model, we improved its AUC to 0.558, and precision score to 0.2318. Posterior predictive checks show that we have a model that is capable of explaining the observed data. Following this, our best model was created using the reduced set of predictors and a transformed version of the days_between variable. This fit generated a model that could predict with an AUC of 0.612 and precision score of 0.5649. Its posterior predictive check did not align well with the observed output for patients categorized as "No-shows". This emphasizes a lack of confidence and high level of uncertainty in our model.

Model	AUC	Precision
Hierarchical Model	None*	None*
Full GLM	0.503	1.52%
Reduced GLM	0.504	1.52%
Reduced GLM, Undersampled	0.558	23.18%
Reduced GLM, Undersampled, Transformed	0.612	56.49%

Conclusions and Limitations

Overall, we were disappointed by the predictive performance of our best model. Although we attempted to predict whether a patient would be a "no-show", our best model only performed slightly better than random guessing. Ultimately, we would not recommend the utilization of this algorithm in the clinic setting to predict "no-shows". We conclude that different patient characteristics that were not included in the dataset may better predict "no-shows". Given more time and resources, we may have been able to improve our predictive accuracy further by exploring interaction terms, using different preprocessing regularization, and adding more quantitative predictors to our dataset to improve predictive performance.

This study faced a few limitations. Firstly, all of the patients included in this data set live in and receive care in Brazil, so our predictions may not be generalizable to other populations. As a result, this model should only be used with regard to this population of patients to predict probability of 'no-show'. Secondly, the outcome class of "no-show" is imbalanced with only 20% of the patients being classified as 'no-show' which could impact the accuracy of predictions. This limitation can be combated by subsetting the data to include equally distributed outcome classes or by using rebalancing sampling techniques such as random over/under sampling like we did in our analysis. Thirdly, we hoped to use a hierarchical model for the dataset but it ultimately proved to be a limitation in our modeling approach since the predictors included in the dataset were seemingly not appropriate for a hierarchical model. Lastly, Simple binary predictors representing a patient's medical history may not be appropriate for finding a solution to this problem.

References

Liu, D., Shin, W.-Y., Sprecher, E., Conroy, K., Santiago, O., Wachtel, G., & Santillana, M. (2022). Machine learning approaches to predicting no-shows in pediatric medical appointment. *Npj Digital Medicine*, *5*(1), Article 1. https://doi.org/10.1038/s41746-022-00594-w

Appendices

Data Description:

- PatientId: Identification of a patient
- AppointmentID: Identification of each appointment
- Gender: Male or Female. (M/F)
- ScheduledDay: is the day someone called or registered the appointment, this is before appointment
- AppointmentDay: is the day of the actual appointment
- Age: How old is the patient.
- Neighborhood: Where the appointment takes place.
- Scholarship: True or False. (0/1)
- Hypertension: True or False. (0/1)
- Diabetes: True or False. (0/1)
- Alcoholism: True or False. (0/1)
- Handicap: Level of Handicap Status (0-4)
- SMSReceived: 1 or more messages sent to the patient.
- Noshow: True or False. (Yes/No)
- App_weekday: Day of the week of appointments
- Days_between: Number of days between ScheduledDay and AppointmentDay
- Days_between_log: Log transformation to the Days_between variable
- ScaledAge: scaled Age variable (centered at 0 with standard deviation of 1)

Github Link:
https://github.com/rachel-holman/Bayesian-Analysis-of-No-Shows-in-Hospital