How to Predict a No-Show

An Analysis of Appointment No-Show Rates in Brazilian Heath Clinics

By Dovid Burns

Why do patients miss their appointments? This simple question underpins one of the most disruptive challenges facing medical offices around the globe. Data analytics lends powerful tools to be able to predict the likelihood of a patient no-show. This information could be adapted by individual practices and hospitals around the globe, allowing for local healthcare networks to make significant staffing adjustments.

HOW TO REDUCE HEALTHCARE COSTS: STAFFING APPLICATION

An example of improved staffing practices includes overbooking certain time blocks or days with larger numbers of patients who have higher probabilities of failing to present for their appointments. This would allow the healthcare provider to schedule fewer staff members on days or time blocks with patients who are likely to arrive for their appointments, thereby potentially averting significant financial losses to healthcare institutions due to a problem as simple as patient no-shows. This study uses data from over 110,000 public healthcare appointments in Vitória, Espírito Santo, Brazil (acquired from kaggle.com datasets), and attempts to indicate the top factors predicting patient no-show.

HOW TO PREDICT A NO-SHOW

Examination of the data showed that the leading indicators of missing an appointment are appointment day, patient handicap, hypertension, diabetes, scholarship receipt, text message reminder, age binned, time between appointment scheduled to appointment day, and neighborhood (see Fig. 1). The strongest indicators based on smallest p-value and largest difference in no-show rates are received a text, age binned, time between appointments scheduled to appointment day binned and neighborhood. Our initial survey indicated that there did not seem to be a correlation between missing an appointment and alcoholism or gender. To further verify these claims, we ran t-tests for independence and chi-squared tests between these variables. To check for correlation between the eight indicator independent variables, we developed a correlation heat map of all the numerical variables (see Fig. 2). Because the highest correlation found between these variables is only 0.24, we confidently analyzed all of the significant variables without concern of a correlation bias. We found the best predictions using logistic regression, followed by random forest, then, followed by k-nearest neighbors.

Feature	**Percent Difference No- Show Rate**
appointment day	<= 1.9%
hypertension	3.6%
diabetes	2.4%
scholarship	3.9%
received a text reminder	10.9%
age binned	<= 9.8%
time between appointment scheduled to appointment day	<= 27.8%
neighborhood	<= 14%

Figure 1. The data identified eight variables that predict patient no-show. Of these, the strongest indicator that a patient will fail to arrive for his appointment is the time that elapses between the scheduling and the actual appointment date. The appointment day – meaning the day of the week of the appointment – shows the least impact on patients' arrival for their appointment.

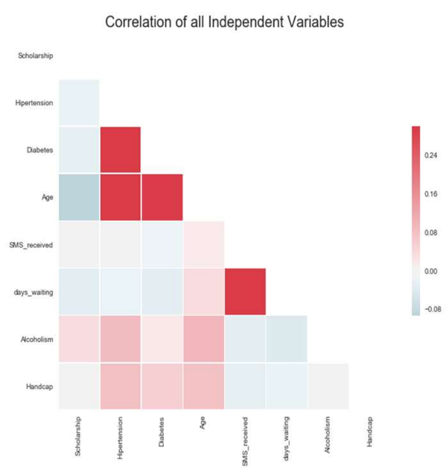

Figure 2. The heat map indicates a positive correlation between diabetes and hypertension. Likewise, we also note a positive correlation between age and hypertension, age and diabetes and a slight negative correlation between age and scholarship. Lastly, the time that elapses between scheduling and the appointment date is correlated with receiving a reminder text.

The models we created with this data were reasonably predictive. However, more information would likely have made our predictions much stronger. The maximum f1-score we found was 0.446 with an AUC of 0.724.

HOW TO USE THIS INFORMATION

Healthcare groups and physicians' offices will be able use this information to schedule more patients, or to bring in fewer practitioners on days when more patients are predicted to fail to arrive for their appointments. Additionally, medical offices can implement effective reminder methods for patients who are predicted to no-show in order to increase the likelihood that these clients do not actually miss their appointments. Another recommendation to medical office staff would be to call likely no-show patients early on the date of their appointment to find out if the patients anticipate arriving to their appointments. The appointments freed by clients who cancel in this manner could then be offered to a back-up waiting list of patients who would like to take last-minute appointments. Alternatively, if there are days where the algorithm predicts very unlikely no-shows for the scheduled appointments, healthcare offices would be able to staff accordingly with greater reliability than previously.

HOW TO IMPROVE THE ACCURACY OF THE ALGORITHMS: FURTHER INQUIRY

While this data set does contain valuable data, there is further information that would most likely make our analysis stronger. It would be helpful to know the type of appointment, e.g. routine physical, sick-visit with primary doctor, specialist visit or follow-up. Additionally, knowing the time of the appointment could also be a useful factor in predicting likelihood of no-show. Further health background data, such as patient pregnancy, AIDS, cancer, etc. would potentially play a role in patient no-show as well. Knowing the type of insurance that the patient has could also be a predictor in no-show. Further, knowing the approval ratings of the hospitals and/or doctors might be helpful in predicting no-show. Lastly, having a longer time frame of appointment data to examine would most certainly increase the predicting power of our algorithms. View the full report and our Jupyter notebooks in the github repository.