

Capstone Project Proposal

Dovid Burns

12/12/17

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. 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 300,000 public healthcare appointments in Vitória, Espírito Santo, Brazil (acquired from kaggle.com datasets), and attempts to indicate the top factors predicting client no-show. To meet this goal, the appointment data will also be analyzed over certain time periods to create new features. Additional binning of continuous variables like age and calculating the effect of different variables on the no-show rate will show variable importance. Once the principle features have been identified, the dataset will be split into train and test subsets, and predictive algorithms will be trained to use these variables to determine whether or not a patient is likely to arrive for their appointment. Test trials with various

algorithms will indicate which algorithm does the best job of predicting patient no-shows. The deliverables will be the code used for both data exploration and for training and testing the predictive algorithm. The findings of this study will be submitted as a paper, including recommendations for medical scheduling changes as a result of the key features found. Additional future steps for continued research and data exploration of this problem will also be presented.