S670 Exploratory Data Analysis Final Project Proposal

EXPLORING THE DOCTOR'S APPOINTMENT NO-SHOW DATASET

Group Name: Final Project Group 6

Team members: *Chhavi Sharma, Prahasan Gadugu, Supriya Ayalur Balasubramanian*

Project Motivation:

No-show at hospitals is an important national problem, the public healthcare sector is trying to cope with. A no-show in the healthcare sector is an appointment, where the patient or client did not show-up or try to call the hospital to cancel the appointment or reschedule the appointment. Missed appointments are associated with poorer patient outcomes and cost the hospitals dearly. Delayed testing potentially puts patients in danger. Missed screening or patient no-show may result in delayed disease detection. Reducing no-show rates can diminish cost and improve quality of health care delivery. Thus, no-shows reduce scheduling capacity, contribute to inefficiency, lower the quality of care, and negatively affect the working environment for providers and staff.

Therefore, it comes as no small surprise that reducing the rate of no-shows has become a priority in the United States and around the world. The first step to solving the problem of missed appointments is identifying why a patient skips a scheduled visit in the first place. What trends are there among patients with higher absence rates? Are there demographic indicators or perhaps time-variant relationships hiding in the data? Ultimately, these are the questions that drove us to take up this project for exploratory data analysis.

Dataset Description:

Dataset Source: Kaggle (https://www.kaggle.com/joniarroba/noshowappointments/version/3)

This dataset is drawn from 300,000 primary physician visits in Brazil across 2014 and 2015. The information about the appointment was labelled such as, when the patient scheduled the appointment and then the patient has either attended or not. The information about the appointment included demographic data, time data, and conditions concerning the reason for the visit.

We included a total of 15 variables from the original data. The variables and the description of the values are as follows,

- **Age**: integer age of patient
- **Gender**: M or F
- **AppointmentReservationDate**: date and time for which the appointment was made
- **AppointmentDate**: date of appointment without time

- **DayOfTheWeek**: day of the week of appointment
- **Status**: Show-up or No-Show
- **Diabetes**: 0 or 1 for condition
- **Alcoholism**: 0 or 1 for condition
- **Handcap**: 0 or 1 for condition
- **Tuberculosis:** 0 or 1 for condition
- **Smoker**: 0 or 1 for smoker / non-smoker
- **Scholarship**: 0 or 1 indicating whether the family of the patient takes part in the Bolsa Familia Program, an initiative that provides families with small cash transfers in exchange for keeping children in school and completing health care visits
- **Tuberculosis**: 1 or 0 for condition
- **SMSReminder**: 0 ,1 ,2 for number of text message reminders sent to patient about appointment
- **WaitingTime**: integer number of days between when the appointment was made and when the appointment took place.

Research Questionnaire:

Using exploratory data analysis methods on the Medical Appointment No Show dataset (from Kaggle), we aim to answer the following research questions:

- **What factors are most likely to determine whether a patient shows up to their scheduled doctor's appointment?**
- **How is the absence/no-show to a scheduled appointment dependent on the general characteristics and behavior patterns of the patient?**
- **Can we predict whether a patient would show up or not by taking the aforementioned variables as explanatory variables into consideration?**

References:

https://www.kaggle.com/joniarroba/noshowappointments/version/3

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4714455/

https://medium.com/@williamkoehrsen/exploratory-data-analysis-with-r-f9d3a4eb6b16