

Capstone Proposal - Predicting Missed Medical Appointments

Derek Samsom

This Capstone project explores the problem of missed medical appointments. Missed appointments result in lost revenue unless there is some over-booking to compensate. Over-booking can help reduce the lost revenue, but without being able to predict which appointments are expected to be missed and the probability of being missed, over-booking will result in times there are more patients than expected, reducing the quality of care.

The client is a medical provider that cares about this problem because they want to minimize revenue loss associated with missed appointments by introducing over-booking while minimizing any negative impacts on patient care caused by having too many patients show up at once. With this information, the client can decide to over-book when the probability of having too many patients show up at a given time is sufficiently low. The probability can be calculated by combining the individual probabilities of each appointment in a time slot and determining the probability that the over-booking results in having too many patients at a given time. The scope of this project is to determine the classification of new appointments (kept or missed), and the probability for each class. The use of the combined probabilities in determining over-booking thresholds is client specific and not included in this report

The data used for this project contains information on 342,862 appointments. Each observation is a single actual appointment. The variables include appointment date and time, patient gender and age, reminder call result, provider specialty, billing type, and the prior number of kept and missed appointments the patient has had. This data will be obtained by an anonymous medical provider.

The approach will be to examine the data and clean where necessary. The data will be explored and I will look for new features that can be added. The next step will be to create a train, validation, and test set for modelling. The train data will be used in both glm and random forest models using the caret package. The models will be compared based on their performance on the validation set, and the best one chosen. The planned performance metrics are F1 score and/or error. The chosen model will be tested on the test set to obtain the final performance estimate.

The deliverables include the code and narrative text of the report written in R Markdown, and a paper report generated by knitting the R Markdown file to pdf. I will also provide a slide deck of 10-12 slides targeted to a more general audience.