Capstone Project Proposal

Building a Vetting Process for Better Scheduling Patients

Introduction

National Center for Health Statistics reported that U.S. healthcare spending grew 4.3 percent reaching $3.3 trillion, about 17.9 percent of the U.S. Gross Domestic Product (GDP) in 2016. To confront rising costs, limited resource capacities, and burgeoning demands facing U.S. healthcare providers, effectively using clinical resources is critical. A major obstacle in cost-effective healthcare delivery and patient safety is no-show. Patient no-show is defined as a patient who does not appear for the scheduled appointment.

The extent and the cost of no-shows are widely studied. Across all specialties, the average no-show rate is of the order of 23%. Researchers found that no-shows and cancellations represented 31.1% of overall scheduled appointments among approximately 45,000 patients per year at a large family practice center with an estimated total annual revenue shortfall of 3% to 14%. Further, several studies found no-show rates of as low as 3% and as high as 80%.

Problem Statement

Missed appointments can cost the U.S. healthcare system hundreds of billions each year. Providers use different methods to reduce the patient no-show, including reminder procedures, penalization, and overbooking. The success of the methods is not clearly determined. On the patient end, 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 healthcare delivery.

Possible Clients

Not only private healthcare practitioners but also public healthcare providers, such as VHA, CMS, research on why patients don't come to their medical appointments and the predictors of no-shows. Using a neutral dataset, which is not from the U.S. healthcare system, the project aims to provide possible clients an overview of how to apply data mining approach to build a vetting process for better scheduling patients.

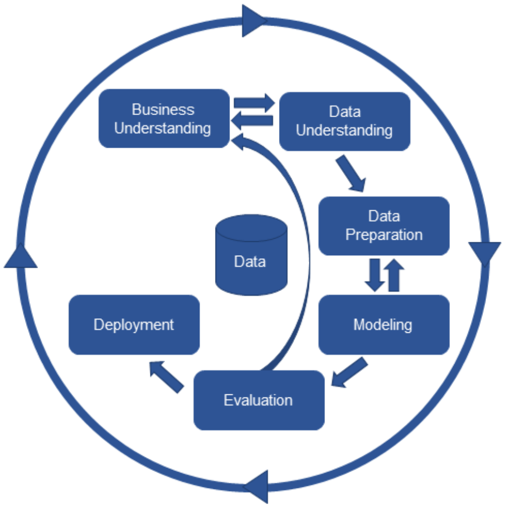
The outcome of the projects can help the clients plan for a data science project to predict medical no-shows using their own data sets. They can foresee what should be done for the data collection and/or selection, data preparation, and the process to build a similar vetting process that work best in their current setting. Having the details, the clients can estimate the expenses needed from both data end and personnel end, including skill sets required for the data scientist team working on the project.

Dataset Overview

The dataset included more than 110 thousand of medical appointments in Brazil, from November 2015 to May 2016. The dataset was published on Kaggle along with details on data characteristics and their context. The dataset included 14 columns and the no-show rate is about 20%.

Approach & Goal

I plan to follow the Cross-Industry Standard Process for Data Mining (CRISP DM) methodology for this project. This is made up of six iterative phases as being showed in the diagram.



1. Business Understanding:
   I will learn what similar research have found, and how the different types of data can affect the outcome.
2. Data Understanding (EDA):
   I will perform exploratory data analysis, generate summary statistics and data visualizations, to look for interesting trends or patterns from the data set.
3. Data Preparation:
   I will transform and impute data to resolve data quality issues in this phase. I will create derived and dummy variables and convert them into multiple binary flag fields.
4. Modeling:
   To predict no-show, a binary target, I plan to start with a decision tree model, then move on to build more complex models such as, random forests and Bernoulli Naïve Bayes.
5. Evaluation:
   After building models using a training dataset, I will run them against a testing data set, a subset of the clean data, which will be set aside prior to the modeling phase. I will calculate the sensitivity of each model and determine which model performs the best.
6. Deployment:
   I will build an interactive tool including some useful functions that can help my possible client utilizing the above phases to select their best model with the optimal sensitivity.

Deliverables

The deliverables for this project will include:

- ➢ Final project report
- ➢ Power-point presentation
- ➢ Jupyter Notebooks in Python 3.6 containing all codes related to the above 6 phases.