Stack Education Data Science Industry Project
Fall 2020
Improving Healthcare Facilities' Operations for Better Patient Care

About Our Partners

Open Avenues Foundation works with talented foreign nationals and immigrants to lead programs that improve economic and educational outcomes for American citizens. In addition, we engage communities and institutions to create environments where global talent succeeds and advances the lives of others.

One of OAF's Entrepreneurs-in-Residence is Trang Nguyen, who is a data scientist at Tamr, where she is involved in creating new methods to use human-guided machine learning algorithms, accelerating data mastering projects and enabling the world's largest organizations to optimize their data operations, rapidly activate latent data, and increase the velocity of business outcomes through data-driven insights. The company has been collaborating with industry-leading companies and organizations to better aggregate and learn from existing datasets to resolve supply chain issues and operational failures.

Our Challenge & The Project

Healthcare transparency has been a constant challenge yet well-recognized need in the U.S., given the complexity of the process for the patients. One of the biggest pain points in the current system is the long wait time healthcare consumers have to undergo in the facility. This not only causes frustration but also discourages people from going to healthcare facilities for regular check-ups as well as office visits for illness.

Fortunately, healthcare facilities have proactively collected more and more operational data over time. Data about patient wait time, number of patients in line, number of exams and so on have been logged by multiple facilities. The challenge now becomes how we can utilize these operational data and machine learning to identify practical solutions to minimize patient wait time from an operational perspective, and therefore provide a better-quality and cost-effective care to patients.

Things to Consider

Healthcare as an industry has always dealt with vast amounts of patient data in all sorts of mediums and formats. This problem has only been increasingly confounded in the information age as digital records of patients are kept at every step of the way through their healthcare experience. As one can imagine, unless a standard is defined, datasets can be messy and may require cleaning prior to use for analysis. Data used in this project is no exception. In order to draw insights from the provided records you will have to learn the art of cleaning your data, and more importantly, learn how to evaluate the impact of data cleaning on the full pipeline.

Data summary:
The project data is part of the operational data challenge by Medical Analytics Group. You can download data and read more about data description here.

Key highlights about the dataset:

- The file contains 4 data subsets, each representing a different hospital facility.
    - The first 3 facilities (F1, F2, and F3) are scheduled while the last one (F4) is walk-in. You can exclude F4 data from this project.
    - Start with the F3 dataset, and if you have time, repeat the data processing and modeling for F1 and F2.
- Each data subset includes patient visit data for 600 to 1000 days.
- Target variable for prediction is *Wait*.

We expect the project to consist of the following stages:

1. Data processing and exploration – load F3 dataset, perform basic statistical analyses and create visualizations to highlight interesting trends and insights
2. Feature Selection - use the results from step 1 to select a set of features to be used for modeling.
3. Modeling and Validation – test a variety of models to find the best suited models through a traditional training/test/validation framework
4. Model Interpretation - extract actionable insights from the model
5. (If time allows) Repeat the process for F1 and F2 dataset

In Order to Be Successful, a Participant will Need the Following Skills

- An ability to creatively solve problems, read research papers and find new and interesting applications is essential
- R and/or Python as a core programming language
- GitHub for version control

Project Deliverables

At the end of the project we would like to have a detailed model that predicts patient wait time, and a presentation or white paper explaining the model and the data science pipeline.

Must Have Goals

- Create at least 1 new feature from the existing set of features in the dataset
- Complete a thorough feature selection process, and exclude any variables that should not be used in the modeling step
- Develop at least 2 different predictive models
- A presentation (maximum 10 slides and 10 minutes or less) explaining the data science pipeline and model conclusions.
- Github repository with all the code used for the project

Nice to Have Goals

- Develop a robust model pipeline including cross validation
- Repeat the data science pipeline for F1 and F2 datasets, and show comparison among the three facilities

The ultimate goal of this project is to create a model to predict patient wait times from hospital operational features, and therefore, provide actionable insights for hospitals to improve their operations.