**Springboard Capstone Project Ideas**

**Date:** April 12, 2018

**Student Name:** Taesun Yoo

**Capstone Project Dataset Guideline:**

* Is it a real problem that someone would care about?
* Is there real data available?
* Is the data easy enough to acquire and clean?

**Project Proposal**:

* What is the problem you want to solve?
* Who is your client and why do they care about this problem?
* What data are you going to use for this? How will you acquire this data?

**Project #1: ED Admission Case Prioritization Modeling**

**Problem Definition**:

Emergency department (ED) visit is a fast way of getting hospital admission especially during weekends and holidays. However, not every patient requires an immediate hospital admission. Thus, prioritizing the patient cases are real challenges of any healthcare system in worldwide. The main goal of this project is to build a supervised machine learning (ML) classification model to predict whether incoming future patients need to get admitted or not admitted. This model will facilitate the management of patient cases in efficient and manageable ways.

**Target Audience (Clients)**:

Ultimately, hospital ED unit managers can utilize this model to predict and monitor incoming future ED patient cases for providing a quality of care for urgent patients who require more attention at a right time by case prioritization for ED admission.

**Data Source**:

This data source is a simulated proprietary dataset acquired from a graduate course offered at University of Toronto for “Big data analytics for healthcare”. This dataset is *simulated by random number generators* with baseline values were given for numerical features (i.e., average income, distance and number of visits) whereas other categorical features are also generated by random number but coded from existing mapping variable codes (i.e., diagnostic codes, ethnicity etc.). This dataset contains two sets of extract .csv files which include a training/validation set for train and validate a model and another test set for evaluating model performance with unseen data by a trained model.

**Project #2: Community Health Service Provider Peer Group Comparison**

**Problem Definition**:

In Toronto regional area, there are two types of health service provider (HSP) organization exists. They are hospital and community sectors. Hospital sectors have a peer group comparison group definition like (i.e., academic teaching, children specialty, mental health specialty etc.) to allow comparison on their performances (i.e., financial budget and service volume). However, unlike hospital health service providers, community health service providers do not have a formal definition of peer group to allow effective comparison on their key performance metrics. The goal of this project is to build a community health service provider segmentation model using unsupervised machine learning approach by utilizing features on the dataset. This will help to form a community peer group comparison based on numerical/categorical features or combinations of both features.

**Target Audience (Clients)**:

Main key stakeholder groups for this model will be community HSP’s managers/directors and government agencies responsible for distribution and allocation of funding for these community HSP. They can use this model to see some of important criterions are (i.e., budget spent, full-time employees, service type, population types and disease types etc.) and based on what peer group does each HSP fall into for effective resource and funding allocation.

**Data Source**:

This data source is a proprietary dataset acquired from my current work place. This dataset comes from joined queries of fact and dimensional tables from SQL server which will be merged with another sets of data coming from flat file sources (i.e., Excel spreadsheets). Any sensitive data fields or columns (i.e., name of health service providers etc.) will be de-identified for confidentiality purpose. Once data wrangling phase is done, data will be exported in a .csv format to be imported into data pipeline for building a machine learning model.

**NOTE**: if this project is finalized, I might require obtaining non-disclosure agreement and approval from my current workplace (i.e., data holder’s permission).

**Project #3: Employee Turnover Churn Modeling Predictions**

**Problem Definition**:

Company based in Spain “myhappyforce.com” developed an app where employees report daily happiness levels at their workplaces. The app is used by companies to track employee happiness or satisfaction level at workplace. The app will collect various kind of features (i.e., liked, disliked, happiness level, comments etc.) Main goal of this project is to build a churn model for making predictions which employees will churn in near future.

**Target Audience (Clients)**:

The main stakeholders of this model will be hiring companies team and human resource department managers. As this model will help them understand the causes and who should they target for reduce the potential risks of having high employee turnover (i.e., churn) rate. Thus, companies can improve their working environments/conditions, benefits and bonuses related to policies to improve their employees’ quality of life at work place (i.e., job satisfaction level).

**Data Source**:

<https://www.kaggle.com/harriken/employeeturnover/data>

Dataset consists of four tables of .csv files extracted from the app. These contain votes, comments, interactions and churn. A vote data reported when an employee answered questions about rating their happiness level on each weekday. Comments section of data stored their peers’ comments and like or dislike them. Finally, the churn section of data stored when an employee churned (quit or was fired).