**Capstone Project Proposal: Cardiac Stroke Patient Classification Model Predictions**

**Date: April 27, 2018**

**Student Name: Taesun Yoo**

**Problem Definition**

Hospital clients want to study one of critical disease which affects nearly 1 in 20 Americans. Stroke is a disease that affects arteries leading to and within the brain. A stroke occurs when a blood vessel that carries oxygen and nutrients to the brain is either blocked by a clot or ruptures. When that happens part of the brain cannot get the blood (and oxygen) it needs, so it and brain cells die.

Over the last few years, the client has captured several health, demographic and lifestyle data about their patients. This includes details such as age, gender, along with several health measurements (i.e., body mass index, hypertension) and lifestyle related variables (i.e., smoking status, occupation type). The main goal this project is to build a model that can predict how likely incoming future patients will develop stroke or no stroke conditions.

**Target Audience**

Ultimately, hospital client(s) can utilize this model to predict and monitor incoming future cardiac stroke patient cases. This will help physicians to take proactive health measures and target prevention by administering appropriate intervention for these patients from having a potential cardiac stroke.

**Data Source**

https://datahack.analyticsvidhya.com/contest/mckinsey-analytics-online-hackathon/

The data source is contributed by a chain of hospital clients to McKinsey (consulting firm) as a data science hack competition hosted at Analytics Vidhya. This dataset contains anonymized patients with their patient ID, gender, health conditions and other demographic features (i.e., residential type, occupation type etc.) about total of 12 features with mixed variables (i.e., categorical and numerical). 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. About volume of data, training set contains 43,000 observations whereas test set contains 18,600 observations.

**Methodology/Approach**

Firstly, data wrangling will be performed on the dataset. Data wrangling involves encoding categorical strings into meaningful numerical values and cleansing missing values with appropriate methods (i.e., value replacement, drop columns, drop observations etc.). Also data quality check on any features of outliers will be performed to avoid potential biases. Thus, Pandas data frame (i.e., tabular matrix with feature as a column and observation as a row) will be prepare for conducting exploratory data analysis. Secondly, exploratory data analysis will be conducted. This includes computation of summary statistics on features and plotting with visualization toolkits from Matplotlib and Seaborn. Finally, a supervised machine learning approach will be performed. Since the goal of project is to build a classification model which can predict whether how many patients would likely to develop a stroke or not as a binary outcome (i.e., yes or no). Anyways, before start building a model, feature importance algorithm will be executed to see which features are meaningful in terms of providing a boost for an accurate model. Thus, various classification models (i.e., logistic regression, naïve Bayes, decision trees etc.) will be tested and determine which model provides best performance (i.e., ROC curve, precision and accuracy metrics).

**Final Deliverables**:

As a part of capstone completion, I will be submitting following items. A capstone paper in word doc format with problem background, method, conclusion etc. A python source code will be uploaded in GitHub repository and a final presentation deck in PDF.