**Capstone Milestone Report: Cardiac Stroke Patient Classification Model Prediction**

**Date:** May 28, 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.

**Potential Clients/Stakeholders:**

Ultimately, hospital client(s) such as cardiac intensive care unit managers and physicians will be interested in utilizing this model. As a classification model with high accuracies and minimum classification errors were able to predict and monitor income stroke patient cases. This will help them to take proactive health measures and target prevention for patients with high risks of developing a stroke condition.

**Data Source**

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

The data source is contributed by a chain of hospital clients and deliver to McKinsey (consulting firm) as a data science hack competition hosted at Analytics Vidhya. In total, dataset comes with 11 features with one output label for predicting stroke patient cases. These anonymized patient cases contain information about their patient ID, gender, health conditions and other demographic (i.e., residential type, occupation type etc.) with 8 categorical and 3 numerical features. The training set contains about 43,000 observations.

Potentially, if dataset contains geographic locations (i.e., states, cities) of hospital clients’, the author was able to integrate other demographics census data such as income, population size etc. However, the dataset does not provide any geographic identifiers. Thus, the feasibility of integrating current data with other data is not possible and applicable.

**Data Wrangling and Transformation**

A stroke training set data imported into Python as a data frame. The author computed and checked number of missing value counts and its percentages. If any feature is missing more than 50% (i.e., threshold level), these features get removed from dataset. According to the computation, none of features were above 50% so all features were kept. After missing value checks, any outliers on numerical features (i.e., age, avg. glucose level and body mass index) were handled using a traditional statistic interquartile (IQR) range method. Using IQR method, lower bound and upper bound thresholds were computed. Thus, the author was able to determine how many outliers were present upon selection of threshold on each numerical feature (i.e., where upper bound of body mass index (bmi) and avg. glucose level) had outliers.

Once outliers were handled on dataset, feature imputation step is performed. Feature imputation is a step where missing values on each feature (i.e., variable) gets replacement by measures of central tendency (i.e., mean, median and mode) depends on type of variables being a categorical or a numerical.

After successful execution of data pre-processing steps, feature transformation is applied on dataset. Typical feature transformation includes feature scaling and feature encoding steps. First, feature encoding is a process where a categorical variable (i.e., gender) is converted in a numerical format (i.e., binary “0” or “1”). This can be handled properly by any classification algorithms. Thus, all categorical features on a training set gets converted into numerical format. Once all categorical features get encoded, feature scaling is applied. Feature scaling is a step where a scaler function (i.e., min-max, Z-score, etc) transforms the original scale (i.e., range) of dataset on various features. This help each range of feature to be in a closer range which helps distance-based classification algorithm(s) or models to make predictions accurately.

**Initial Data Exploration**

As for initial data exploration, author counted frequencies of stroke vs non-stroke patient cases for obtaining a prevalence of stroke condition within patient population:

* Pie chart showed 98% of patients were non-stroke (i.e., healthy) and only 2% were stroke patients. Thus, this suggested need for class imbalance adjustment before training a machine learning (ML) model.

For categorical variables interaction with a stroke condition, various frequency counts are computed and grouped by lifestyle and health indicator factors:

* **Gender**: stroke is not a gender specific condition. However, among patient populations with a stroke condition there are more female patients (55%) with a stroke condition than man (45%).
* **Marital Status**: interesting enough, there are more stroke patients with a married status (90%) than a single status (10%).
* **Residence Type**: residential area is not an important factor as almost equal proportion patients have stroke condition who reside in rural (49%) or urban area (51%).
* **Occupation Type**: work type is a quite interesting feature. As it demonstrates that there is high tendency for patients work in private sector (56%) or a self-employed (32%) have a stroke condition than patients work in government sector or children (i.e., kindergarten).
* **Smoking Status**: smoking status seems weekly associated with a stroke condition. As it shows that group of non-smoking patients (55%) have higher chances of having a stroke than group of smoking patients (45%).
* **Hypertension**: hypertension is not a significant determinant factor on stroke condition. It clearly shows that patients with no hypertension (74.5%) have more strokes than a group of patients with hypertension (25.5%).
* **Heart Disease**: heart disease is not a significant factor as well. As group of patients with no heart disease have a stroke (77.4%) than patients with heart disease condition (22.6%).

For checking distribution on numerical variables, histograms were plotted on age, body mass index (bmi) and avg. glucose level:

* **Age**: a histogram showed that age distribution of patient population is non-uniformed. Also, it showed some tri-modal characteristics (i.e., two peaks on lower and upper end) suggesting that decent number of stroke patients from infants to child (i.e., less than a year to 5-year-old) and seniors (i.e., 80-year-old). However, most of stroke patients were around mid-30s to 60s.
* **Body mass index**: a histogram showed that quite a normally distributed patient population. Most patients bmi was centralized around range of 25 to 30.
* **Average glucose level**: a histogram suggested quite a positive skewed (right) distribution of patient population for average glucose level. Many stroke patients were on lower end (i.e., 75 to 100) whereas some patients had extreme high level of average glucose than entire patient population.

Correlation matrix plot on entire training set

* There were some interesting correlation patterns did exist among features. First, age and bmi showed a positive moderate correlation (i.e., 0.35). There was increased in bmi from younger age to mid age range of patient population. Second, age and glucose showed a positive weak correlation (i.e., 0.24). Although trend is not so clear but as patients get older, their average glucose level becomes higher within certain age range. In addition, there was a weak positive correlation with age and stroke condition (i.e., value of 0.16) in comparison to other numerical features’ (i.e., bmi and glucose level) correlations on a stroke condition. Therefore, it was hard to make any generalization on trends or patterns about health monitoring features (i.e., age, bmi, glucose level) and a stroke condition.