Capstone Final Report: Cardiac Stroke Risk Stratification Model

Student Name: Taesun Yoo
Date: June 12, 2018

1. INTRODUCTION

1.1 Problem Definition

Stroke is one of critical disease which affects nearly 1 in 20 Americans and 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 the brain cells in the affected portion die. As stroke is a preventive condition, a risk prediction model is of interest to hospital clients. 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 of this project is to build a model that can predict how likely incoming patients will develop stroke.

1.2 Target Audience

Ultimately, hospital client(s) can utilize this model to predict and monitor future incoming patient cases. This will help physicians to take proactive health measures and target preventions on patients with high risks of developing a stroke condition.

1.3 Data Source

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 a total of 12 features on anonymized patients including mixed variables (i.e., categorical and numerical) such as the patient ID, gender, health condition and other demographic features (i.e., residential type, occupation type etc.). This dataset contains two sets of .csv files which include a training/validation set for model training and another test set for final model performance. About volume of data, training set contains 43,000 observations whereas test set contains 18,600 observations.

2. DATA WRANGLING

2.1 Data Pre-processing

Initially, dataset required data cleansing and metadata formatting (i.e., data types). Dataset was loaded as data frame and a few observations are printed to check data type on each column. Currently, dataset posed a couple of problem as shown in Table 1.

Index	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	30669	Male	3	0	0	No	children	Rural	95.12	18	nan	0
1	30468	Male	58	1	0	Yes	Private	Urban	87.96	39.2	never smoked	0
2	16523	Female	8	0	0	No	Private	Urban	110.89	17.6	nan	0
3	56543	Female	70	0	0	Yes	Private	Rural	69.04	35.9	formerly smo…	0
4	46136	Male	14	0	0	No	Never_worked	Rural	161.28	19.1	nan	0
5	32257	Female	47	0	0	Yes	Private	Urban	210.95	50.1	nan	0

Table 1. Top 5 observations of cardiac stroke training set data frame before data cleansing.

First, there were missing values presented in two features smoking_status and body mass index (BMI). Smoking_status is a categorical feature and it was missing 30.6% of observations. Followed by BMI is a numerical feature and it was missing about 3.37% as shown in Table 2. Missing values were imputed using different measures of central tendency (i.e., mean, median and mode). For numerical features like BMI was replaced by the median of BMI. Alternatively, it could be replaced using a mean value, but median value is better as it is less subjective to outliers (i.e., due to presence of extreme values on that feature).

Feature Name	Data Type	Missing Value Counts	Missing Percent (%)
smoking_status	Object	13292	30.6
Bmi	Float	1462	3.37
hypertension	Int	0	0.0
heart_disease	Int	0	0.0
ever_married	Object	0	0.0

Table 2. A table showing data type, missing value counts and percentage on five features

Second, outliers were required to be managed properly on numerical features. In a training set, there were three independent features (i.e., continuous). These included "age", "BMI" and "avg_glucose_level". Interquartile range (IQR) method applied here. For example, if any value of a feature sits below lower and above upper bounds of IQR, these observations will be removed from dataset. IQR is defined as: IQR = $Q_3 - Q_1$ in which Q_3 is 75^{th} percentile and Q_1 is 25^{th} percentile of a feature. Lower bound (LB) equals to $Q_1 - (1.5*IQR)$ and upper bound (UB) equals to $Q_3 + (1.5*IQR)$.

The results of outlier detection based on IQR method is shown in Table 3. There were outliers on bmi and avg_glucose_level since max. value of these features are greater than their defined UB values. Filtering with UB value of 'avg_glucose_level' resulted in too many observations loss (i.e., reduction in 10% of sample size) compared to using UB value of BMI to remove (i.e., reduction in 2% of sample size) observations. Thus, data frame was filtered by using computed UB of 'bmi' rather than 'avg_glucose_level'.

Feature	Min	Max	Lower Bound	Upper Bound
Age	0.08	82.0	-30.0	114.0
Bmi	10.1	97.6	8.65	47.4
Avg_glucose_level	55.0	291.1	25.7	163.8

Table 3. A summary table on continuous measures for outlier detection using IQR method.

2.2 Feature Transformation

After data pre-processing, feature transformation was required for building a good classification model. Mainly, there are three kinds of transformation techniques for machine learning problems. First is a feature scaling where different ranges of feature inputs feed into a scaler function (i.e., min-max, logarithmic, etc.). This function will re-scale them into a similar range on each feature. This helps some classification model(s) to handle importance of features in a normalized fashion. Feature scaling was applied before the classification models were constructed.

Second is a feature encoding. Feature encoding is a process in which categorical feature gets split into multiple column(s). Then original string values ("Yes or "No") of feature gets re-encoded as a binary integer (0 and 1). For example, nominal categorical feature like 'ever_married' will be processed as ever_married_Yes and ever_married_No using a dummy variable encoding. Note that while performing a dummy variable encoding, first dummy variable on each converted feature needs to be removed. This prevents a model from having high-multi collinearity (i.e., high correlation) among same kind of features. In addition, an ordinal feature like smoking status required to be re-encoded as numeric with a unique order. For example, original string values like never smoked, formerly smoked and smokes were encoded as 0, 1 and 2.

Third type is feature engineering. Feature engineering is a common technique in ML problem where an existing feature get transformed or multiple features combined to generate new feature(s). In this problem, discretization is applied on age feature to create a two set of features. One is an age group (i.e., age <30, 30-40, …, age > 80) and the other one is an age group label (i.e., 1, 2, 3…, n).

Finally, dataset was properly cleaned, and each feature was formatted with correct metadata type. Metadata formatting ensures nominal, ordinal and numerical features to be encoded as a category, an integer and a float data type. Dataset was prepared for conducting exploratory data analysis and building a set of classification models.

3. EXPLORATORY DATA ANALYSIS

3.1 Analysis on Entire Training Set

Initial data exploration started with entire training set

Pie chart on prevalence of developing a stroke condition

First, pie chart was constructed to investigate the proportion of non-stroke vs. stroke cases. Pie chart showed that 98% of patients were healthy (i.e., non-stroke) and only 2% of patients had a stroke condition (Figure 1). Thus, this suggested class imbalanced on this data must be addressed to build a good classification model.

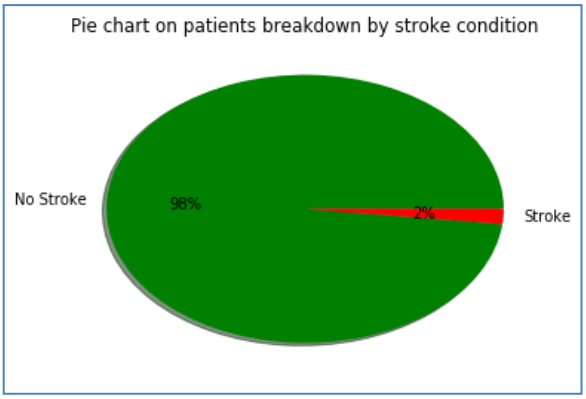

Figure 1. Pie chart breakdown by proportion of non-stroke and stroke patient cases

Frequency counts on a group of stroke patient population

Frequency counts analyses were performed on features related to lifestyle and health monitoring in the group of stroke patients. Here is a summary of observed insights from frequency counts:

- **Gender**: among patient population with a stroke condition there were more female patients (55%) than male patients (45%).
- **Marital Status**: there were more patients with a married (90%) than a single status (10%).
- **Residence Type**: Almost equal proportion of patients with a stroke condition resided in rural (49%) and urban area (51%).
- **Occupation Type**: there were more patients with a stroke condition working in the private sector and were self-employed than ones working in the children and government sectors.
- **Smoking Status**: smoking status seems weekly associated with a stroke condition. As it showed that group of non-smoking patients (55%) had a higher chance of having a stroke than group of smoking patients (45%).
- **Hypertension**: it clearly showed that a group of patients with no hypertension (74.5%) had more strokes than a group of patients with hypertension (25.5%).
- **Heart Disease**: among the stroke patients, majority of them had no sign of heart disease and only minority had heart disease

Distributions of patient population by age, body mass index and glucose level

A histogram analysis was done on three features to show distribution and two different groups of population were plotted. Blue and orange labels showed non-stroke and stroke populations respectively. Here is a summary of observed insights from a histogram:

- **Age**: a histogram showed non-uniform distribution on both populations. Distribution of age on stroke population indicated that majority of patients were senior (i.e., skewed to left).
- **BMI**: a histogram showed uniform distribution on both populations. Most of stroke patients were in BMI range between from 25 to 27.
- **Avg_glucose_level**: a histogram showed non-uniform distribution on both populations. Most of non-stroke population were non-diabetic as shown in below. On the other hand, distribution of mean glucose on stroke population showed bi-modal peaks.

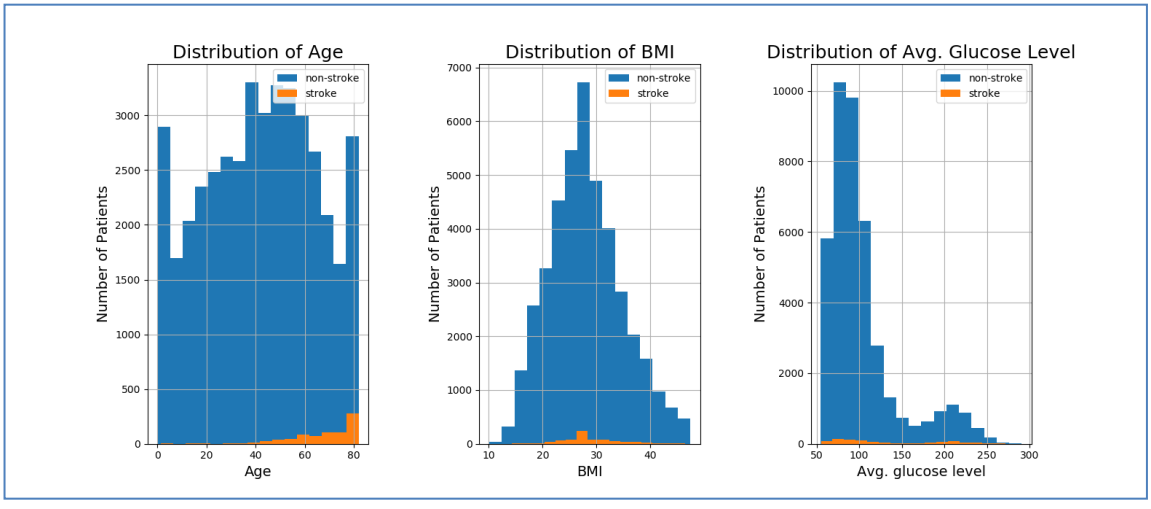

Figure 2. A histogram of age, bmi and average glucose on non-stroke and stroke population

Correlation matrix plot on entire training set

- There were some interesting correlations observed (see Figure 3). First, age and BMI showed a positive correlation (0.39). Second, age and glucose showed a positive weak correlation of 0.24. Although trend was not so clear but as patients get older, their average glucose level becomes higher. This was true for only within certain age ranges. In addition, there was a weak positive correlation on age and stroke condition (0.16) in comparison to other numerical features. Therefore, it was hard to make any generalization on relationships about health monitoring factors and stroke condition.

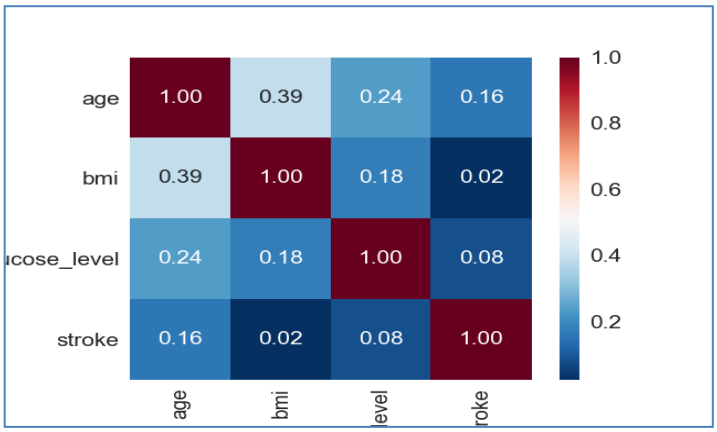

Figure 3. Correlation matrix plot of entire training set

3.2 Analysis on Stroke Patients Cases

Correlation matrix plot on stroke patients only

- As there weren't any strong trends among numerical features and stroke condition in the entire population and due to the highly imbalanced nature of this dataset, I decided to separate stroke patients from non-stroke patients and analyze the trends in individual cohorts. Another correlation matrix plot was created as shown in Figure 4. There was a weak negative correlation between age and BMI (i.e., value of -0.19). This is an exact opposite trend. Thus, for stroke patients as they become older, BMI is reduced. Interestingly, the correlation between BMI and avg. glucose level stayed same (i.e., value of 0.18) compared to the correlation matrix plot from earlier analysis. However, in comparison to the original correlation matrix plot, the correlation between age and avg. glucose level (i.e., value of 0.07) is reduced.

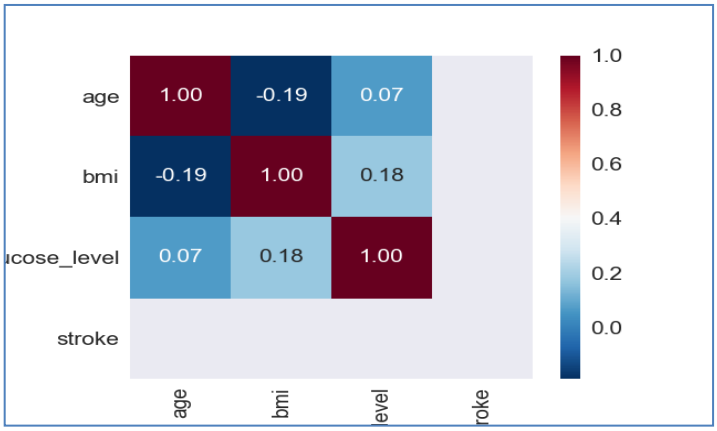

Figure 4. Correlation matrix plot of stroke patients only

Bar charts on mean age of stroke patients group by different factors

From earlier analysis, we found that age was the most important predictor for stroke condition (i.e., correlation value of 0.16). To explore potential relationships with age and other categorical factors, bar charts of stroke patient's group mean age were plotted. In most cases, mean age of stroke patient's group was in mid 60s (i.e., age 65). The difference in mean age of stroke patients by occupation type shows the expected trend: patients employed in government job have a lower mean compared to that of private and self-employed. The reason here may be the retirement age in these jobs. Thus, this indicated the presence of younger stroke patient population. Unfortunately, training set did not come with any features that might explain this variability (i.e., presence of a group of younger stroke patients). It would have been extremely useful if the hospitals originally collected information such as genetic pre-disposition. This can be a good indicator on what is probability of developing a stroke at such an early age (i.e., infants to children).

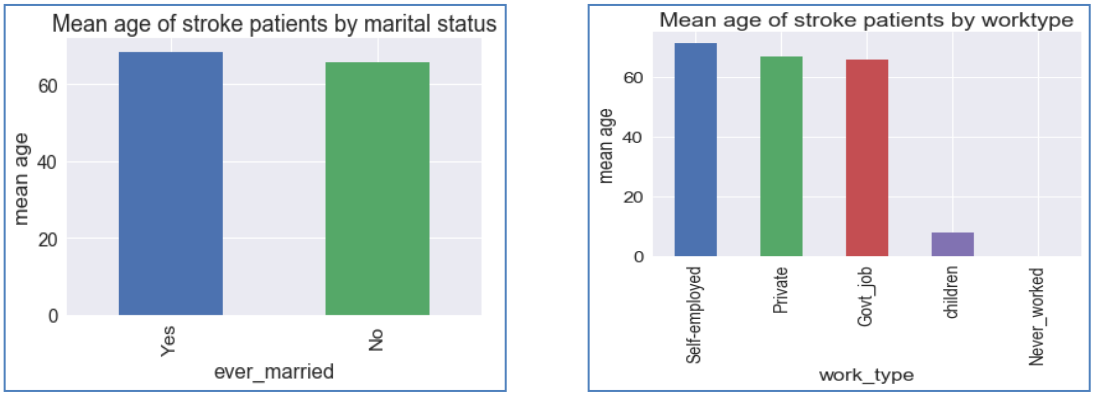

Figure 5. Bar charts mean age of stroke patients group by marital status and occupation types.

Pie chart of age stratification on stroke patients group

To have better understanding on age groups distribution within the cohort of stroke patients, the proportion of patients within each decade of age was plotted as a pie-chart. As it can be seen on a below pie chart, more than 90% of cases were older population (age > 50). On the other hand, very few cases of stroke patients were younger than 30-year-old.

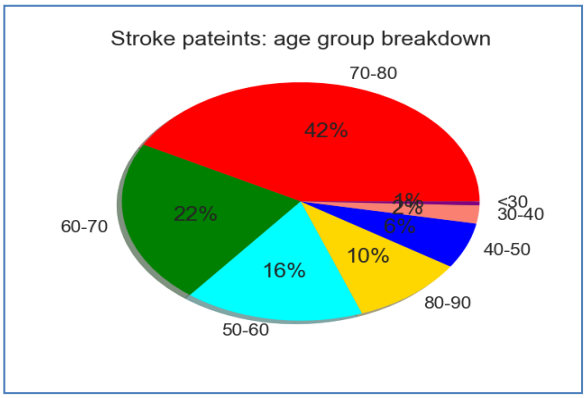

Figure 6. Pie chart of age group breakdown on stroke patient population

3.3 Analysis on a balanced dataset with matching age distribution

Correlation matrix plot on case control study:

To explain any hidden co-variates besides age, case control analysis was performed, where a sample of non-stroke patient population (n=783) was sampled by negative sub-sampling based on age distribution of stroke patient population counts (n=783). As expected, the correlation between age and stroke is significantly reduced to 0.03 when the proportion of age groups on non-stroke population is matched for age groups of stroke population. Conversely, a correlation between avg glucose level and stroke is increased to 0.12. Also considering the correlation between age and BMI, we ended up removing some correlated significance.

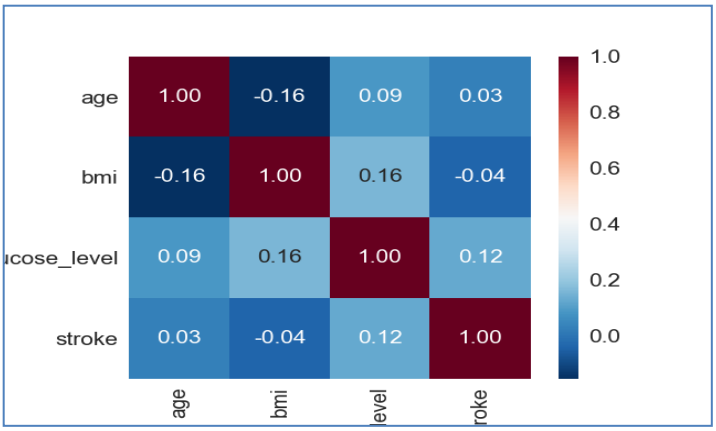

Figure 7. Correlation matrix plot of balanced data with matching age distribution

3.4 Analysis on Randomized Down-sampling for Non-stroke Patients Group

Correlation matrix plot on randomized down-sampling:

To account for the high imbalance in the distribution of stroke and non-stroke patients, the non-stroke patients were negatively subsampled through randomization to generate a balanced dataset. Noticed how correlation changed between age and a stroke condition with value of 0.59.

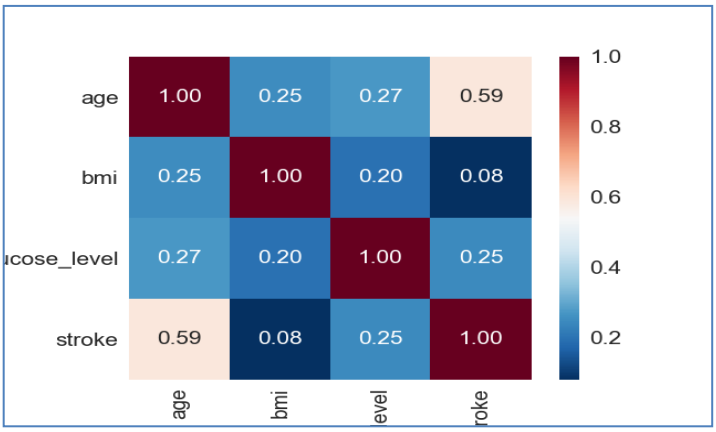

Figure 8. Correlation matrix plot of randomized down sample

Multi-faceted relationship of age vs. avg. glucose level by categorical factors:

To further explore the relationship of age and average glucose level, a linear regression analysis was performed on the balanced dataset. Relationship of age and avg. glucose level was explored by using a marital status and a stroke condition for condition factors. We selected a marital status and a stroke condition for conditional factors because they showed similar trends (i.e., distribution of scatter plots conditioned by stroke population and married population). Thus, we wanted to analyze this interesting insight by calculating Pearson correlation value. Here is a summary of linear regression fitted scatter plots:

- **Faceted by a stroke condition**: a scatter plot faceted by a stroke population showed very weak association (r value of 0.06) between age and average glucose level. As stroke patients got older, there was a slight increase in average glucose levels. Conversely, non-stroke patient population showed a bit improvement in correlation with value of 0.2.
- **Faceted by a marital status**: a scatter plot faceted by a marital status showed weak association (r value of 0.17) between age and average glucose level. As married patients got older, there was slight increase in average glucose over time. Conversely, single patient population showed a bit improvement in correlation with value of 0.3.

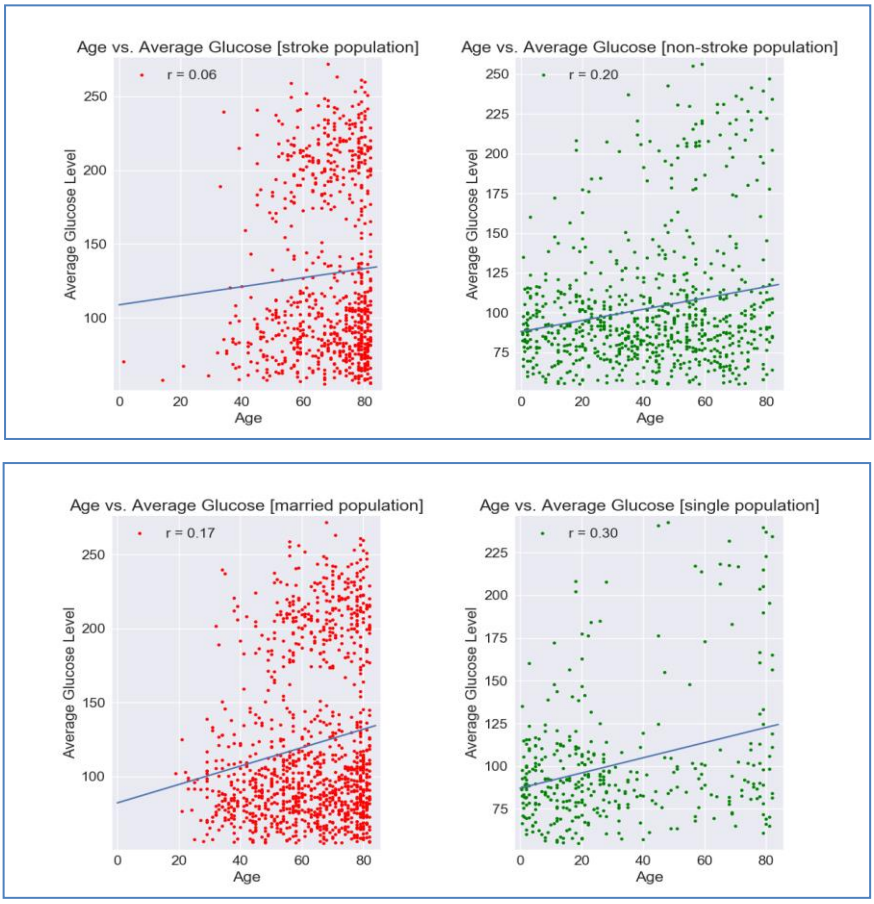

Figure 9. Faceted scatter plots of age vs. avg glucose by marital status and stroke label

4. MACHINE LEARNING CLASSIFICATIONS

4.1 Handling Class Imbalance by Resampling on Majority (non-stroke cases)

As mentioned earlier, original dataset posed a problem where 98% of patient cases were non-stroke and only 2% were stroke. This is a problem for where a classifier model will likely to predict non-stroke patient cases all the time. To recap, main goal of this project is to build a model which can screen patients with high risks for developing a stroke condition. Thus, resampling technique was required to adjust this imbalanced ratio (98% vs 2%). Therefore, the author performed down-sampling to reduce sample size from a group of non-stroke cases. Through randomization to match sample size on a group of stroke cases (n=783). Now, dataset is well-balanced with 50:50 ratio on both classes (i.e., non-stroke vs. stroke).

4.2 Data Partition and K-fold Cross Validation

After resampling, the dataset was split into train and test folds with partition size of 75% and 25% respectively. Due to the limited size of the balanced dataset, a k-fold cross validation approach was used for model training. Briefly, the training set was divided into multiple folds where in one iteration, one-fold was used as the validation set and the remaining folds were used for training. This procedure is repeated for K iterations where each fold serves as the validation set. Model performance over training is evaluated as the average performance across all the iterations. The hold-out test set was used for evaluating the final model performance.

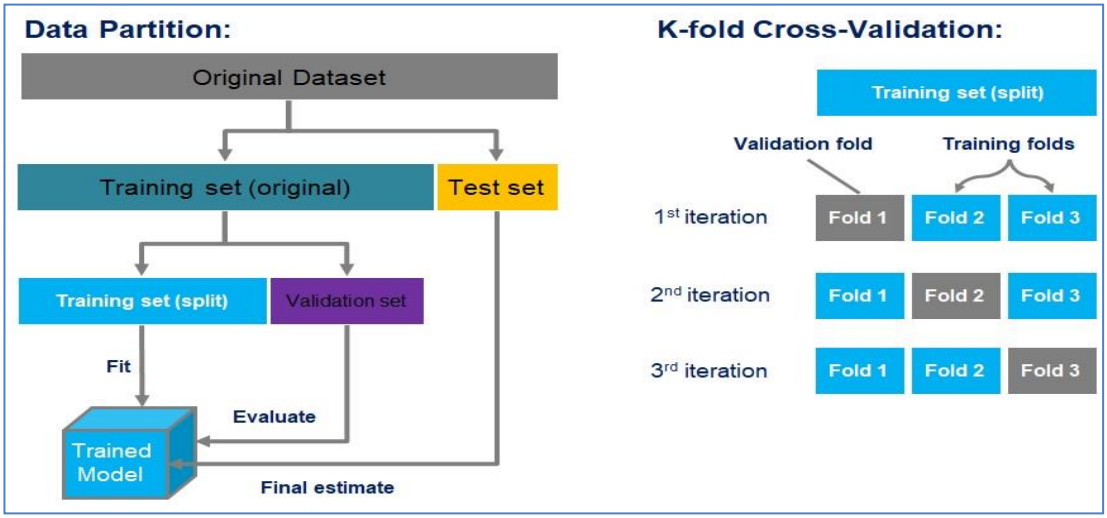

Figure 10. A diagram showing summarized view of data partition and k-fold cross validation

4.3 Model Selection and Feature Importance

Now, the pipeline of model training to validation was constructed, we wanted to evaluate and pick the best classification model. Four selected models were logistic regression, decision tree, random forest and XGBoost (i.e., Gradient Boosting) classifiers. In a nutshell, here is a summary of how each algorithm works. Logistic regression works by using a logit function to transform input value of features and calculate estimated probabilities of a label in range of [0,1]. For example, if $P(1=\text{stroke}) \geq 0.5$, an observation is predicted as a stroke. Whereas if $P(1=\text{stroke}) < 0.5$, an observation is predicted as a non-stroke. Decision tree is an algorithm where it predicts the value of a target variable (label) by learning simple decision rules inferred from selected features. Tree is generated and split data on features. It continues to split in repetitive process at each node until

leaves reached purity (i.e., remaining samples at each node belongs to same class either non-stroke or stroke cases only). Random forest is a typical ensemble learning model. It takes random sub-sample of data from each tree, so all constructed trees are different from each other. Thus, model makes classification based on predictions made from each tree with averaging (i.e., like picking a vote from majority). XGBoost is a type of gradient boosting model in which subsequent model learns from the mistakes (i.e., residual errors) of previous model in a step-wise forward manner. In Gradient Boosting, residual errors are identified gradients. These gradients help how XGBoost to improve model performances.

From exploratory data analysis, it was clear that there were irrelevant features for predicting patients with high risks of developing a stroke. Thus, feature selection process was implemented to rule out any irrelevant features and improved model's accuracy. The feature importance module can be computed by different criterion. Most common ones are Gini importance and permutation importance. Feature importance plots from Decision Tree to XGBoost models were plotted (Figure 11). Notice how all models picked age as the most important feature. Followed by 2nd and 3rd ranked features were different from model to model. For example, Random Forest model picked avg. glucose and BMI as 2nd and 3rd. Conversely, XGBoost model picked BMI then avg. glucose level. However, Decision Tree model picked heart disease_yes as the 3rd rank which was a bit unexpected. Based on importance score of each model, top three were selected to evaluate final model performances.

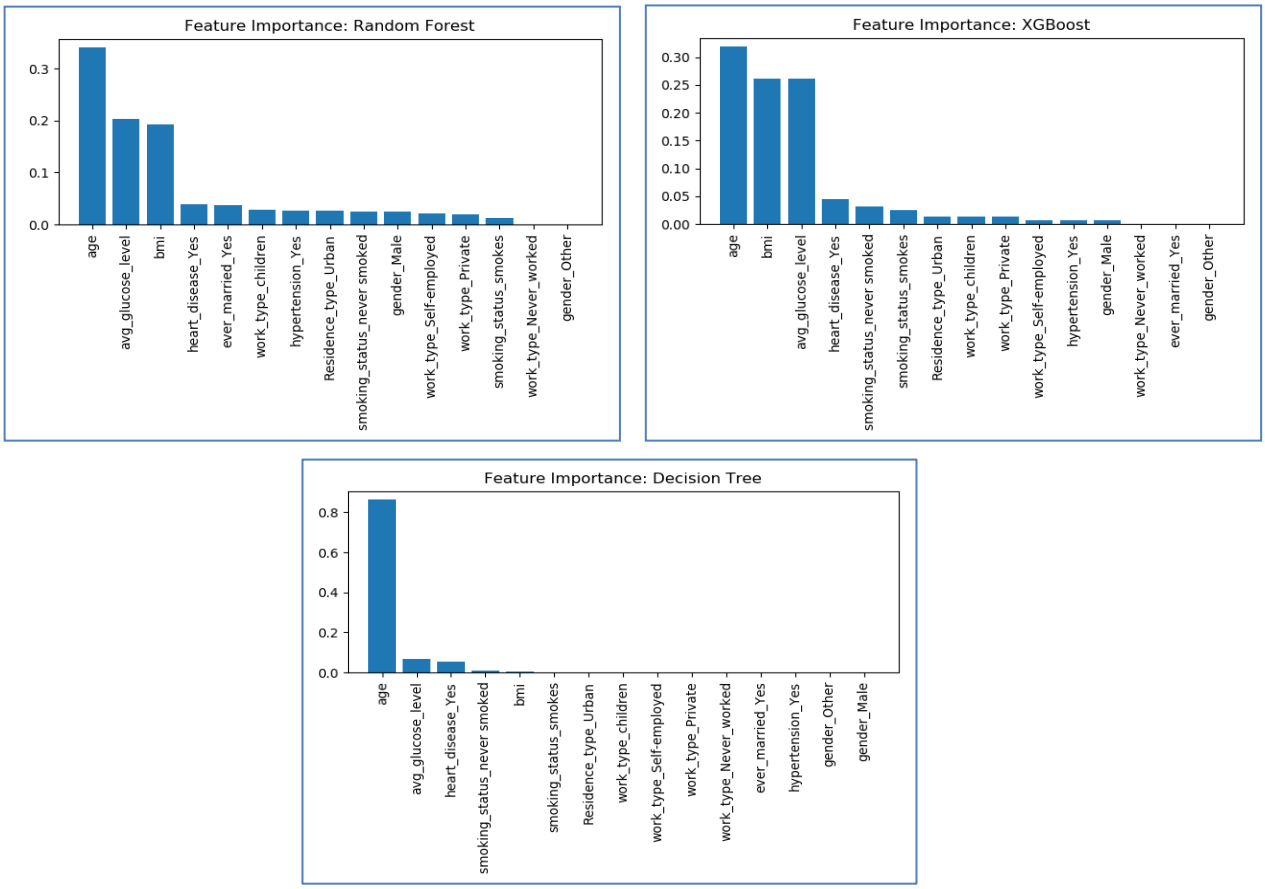

Figure 11. Feature importance plots of Random Forest, Decision Tree, XGBoost

4.4 Model Tuning with Hyper-parameters

After model selection, hyper-parameters tuning was performed using randomized hyper-parameter search to maximize our model performances. RandomizedSearchCV library from Python's sklearn package was used. Each classification model has a set of hyper-parameters where each parameter can be specified in a range. For example, logistic regression has the parameter "C" which is the inverse regularization strength and is used to prevent over-fitting. A parameter list which contained the range for a set of hyper-parameters for each model was specified and for any run the parameters were selected through randomization.

4.5 Model Evaluation and Results

Table 4 summarizes the performance metrics of each classification model after hyper-parameter tuning. Notice that accuracy and precision of all four classifiers were around 76.5 and 72.5%. However, the best classifier was identified as XGBoost. XGBoost had well-balanced performances on accuracy, ROC score and recall compared to other three models as shown in Table 4.

	Logistic Regression	Decision Tree	Random Forest	XGBoost
Accuracy	77%	75%	77%	77%
Precision	75%	68%	73%	73%
Recall	81%	93%	84%	86%
ROC Score	77%	75%	77%	77%

Table 4. Summary of classifier evaluation metrics on four models

Here is representation of similar information using a receiver operation characteristics (ROC) curve. Along the x-axis of the ROC curve is the false positive rate, which is how many of predictions are false from stroke predictions and along the y-axis is the true positive rate, which is how many of predictions are true from stroke predictions. In summary, mean area under the curve (AUC) of XGBoost model showed that AUC value of 0.84 (Figure 12).

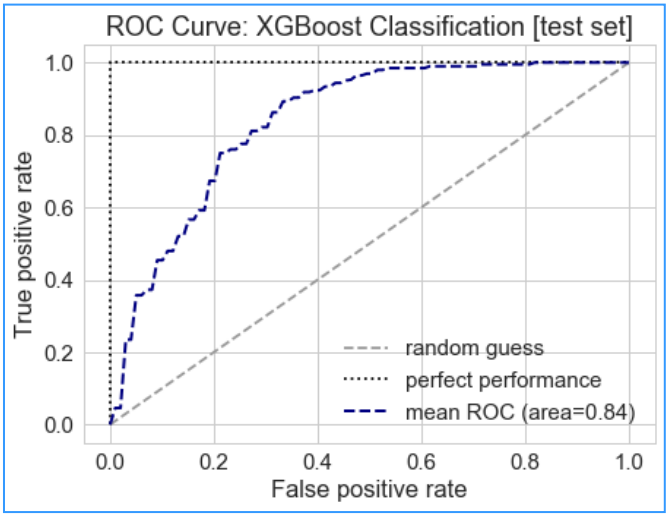

Figure 12. ROC curve of XGBoost model on test set with AUC = 0.84

Finally, a precision-recall curve was plotted to explore changes in precision and recall level of XGBoost classifier. On this plot, x-axis represents recall and y-axis represents precision. Recall is a metric in which if the model picks a random actual stroke case, what is the probability of making the right prediction (i.e., stroke prediction). On the other hand, precision is a metric if model makes a stroke prediction, what is the probability that it is indeed an actual positive case (i.e., stroke case). Overall, the average precision (AP) level of XGBoost model was around 0.81. AP summarized a precision-recall curve as the weighted mean of precisions achieved at each threshold, with increase in recall from the previous threshold used as the weight. AP is defined as $AP = (R_n - R_{n-1}) * P_n$. Where P_n and R_n are the precision and recall at the n_{th} threshold. Based on above weighted mean precision, 81% of time the stroke predictions were correctly retrieved from total number of actual stroke cases based on test set.

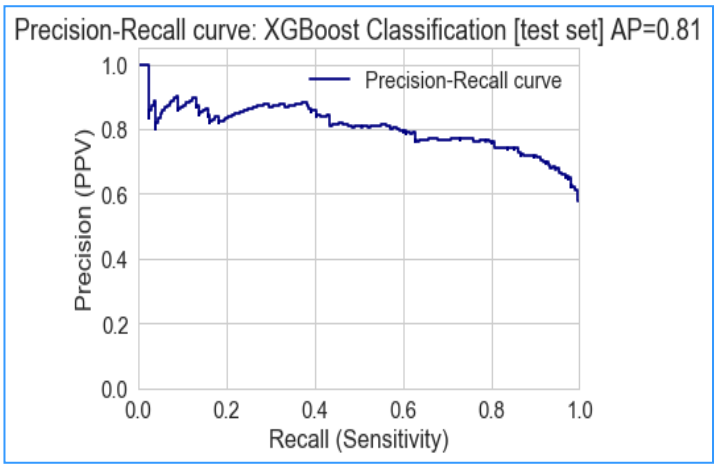

Figure 13. Precision-recall curve of XGBoost model on test set with average precision = 0.81

5. CONCLUSION AND FUTURE WORK

From model performance comparison and feature importance plots, we were able to find important findings on our stroke risk stratification problem. First, there were consistency among different models where age came out as the most significant predictor. This was proved in a correlation matrix plot of balanced dataset where a correlation between age and stroke with a value of 0.59. Second, average glucose level and BMI were considered 2nd and 3rd important features. However, the rank of importance was different from model to model. For example, XGBoost model picked BMI as 2nd rank whereas Random Forest model picked average glucose level as 2nd rank. Overall in terms of correlation power, age was predominant predictor with value of 0.59 for predicting patients with a stroke condition.

Third, there were few interesting trends observed from frequency counts on a group of stroke patients. Surprisingly, there were more stroke patients with no hypertension and history of heart diseases. This was extremely unexpected. Patients with high risks of a stroke condition were more likely to have hypertension (i.e., high blood pressure) and history of heart diseases. Fourth, it was interesting note that approximately 90% of patients were married and 10% were single in stroke patient population. It will be interesting to know whether this occurred by pure chance of sampling. Finally, upon dissecting a group of stroke patients, there were younger groups of patients with less than 30-year-old. However, the dataset did not have useful features which can explain this

phenomenon. It would have been extremely useful if the clients were able to collect feature like genetic pre-disposition (i.e., molecular genetic factors on causing stroke at younger ages).

In conclusion, this classification model can be further extended and improved based on a couple of approaches. First, a collection of more meaningful features that correlates better with a stroke condition (i.e., patient's stress level, genetic pre-disposition, physical exercise level etc.). Addition of these features will help to build more accurate and reliable classification model for targeting patients with high risks of developing a stroke condition. Second, testing the feasibility of implementing other types of an ensemble classifier model. It is a way to get the best predictions from all available models. For example, each individual classifier makes different prediction on every observation. Then the ensemble classifier predicts class label that has been predicted by majority of classifiers which receives more than 50% of votes (i.e., class being stroke or non-stroke). Thus, we may be able to improve model performance by a few margin (max. 2-3%). The Third, different resampling strategies can be tested. There are different kinds of resampling from down-sampling on majority class label, over-sampling on minority class label and SMOTE (i.e., synthetic sampling) with bootstrapping. It will be worth to test whether SMOTE or over-sampling improves model performances. Finally, from a modeling perspective, it will be worth to test building age group specific classifiers. One for younger and another one for senior population classifiers as highlighted from previous analyses and classification results.

6. RECOMMENDATIONS

In summary, here are three recommendations for hospital clients.

Recommendation 1: The finding suggested that there are significant percent changes in having more stroke conditions from age group 40-50 to 70-80. From each interval (i.e., age gap of 10-year-old), changes in having more stroke patients went up 10%, 6% and finally 20% from age group 40-50 to 70-80 (Figure 6). Thus, it is important to target prevention through an additional stroke screening at a recommended age. For example, an additional screening test for making stroke prognosis is recommended before at age 65 (mean age of stroke patients is 65-year-old).

Recommendation 2: Second recommendation is connected to an age variable. Given some younger patients who have stroke conditions, it is recommended to collect genetic pre-disposition for developing a stroke condition and other health monitoring indicators to be included on data collection. There is a higher chance to explain this variability on younger cohort of stroke patients. This will ultimately benefit hospitals understand the demographics and incorporated into building an accurate classification model.

Recommendation 3: Based on weak correlations of BMI vs. stroke and avg. glucose level vs. stroke, it is recommended to explore these relationships little bit deeper. For example, many research studies proved that people with diabetes and obesity are at high risks of developing a stroke condition than healthy people. However, there were no clear trends observed from scatter plots and a correlation matrix plot of balanced dataset. It will be worth to conduct cohort studies, using a selected features and separate observations into different patient cohorts. For example, if patients whose BMI are higher than 30, these patients are categorized under obese patient cohort. This will help stratify patient population into normal and abnormal (i.e., obese and diabetic) groups.

It may be helpful for building an accurate classifier which can predict patients with high risks of having a stroke.