**Machine learning approach for predicting the stroke**

**Abstract**

Within the scientific discipline of machine learning the primary objective is to assemble models that will provide sound predictions.

**Data Analysis and Preprocessing Approach**

In order to build an accurate model, several data steps were taken to reach our end goal. Jackson Heart Study data (N=2653) were initially merged on an identification variable to create an all-encompassing dataset used to conduct exploratory data analysis. During the data preprocessing phase, descriptive statistics highlighted several variables with missing values. Although, there are several methods to work around this hinderance, such as replacing values with the mean or median, variables where 90% of the values were missing were removed.

To improve on the merged JHS dataset steps to determine whether or not variables correlated were taken. Measures that didn't possess predictive power were dropped which condensed the number of columns (N=160). Highly correlated variables (> 0.95) were identified and removed from the dataset. In certain models, several issues can arise from variables that are correlated. When independent variables are highly correlated with one or more other independent variables, problems can occur with the interpretation of results. In addition, algorithms such as Random Forests, which is known to detect interaction between features can be misinterpreted when variables are highly correlated.

Imbalanced data occurs when there is an imbalnced ratio of observations which can lead to classification issues. Contingent on use of data, imbalanced data has proven to be problematic when applied to machine learning. In the case of the JHS dataset, several variables dealt with dispproportinate data. For example, the ECG determined Anterior Major Scar variable had 2,634 cases where this was absent and only 19 cases were it was present. In terms of categorical data, the Insurance Type variable also dealt with imbalanced data, (Private Only 1520, Uninsured 332, Private & Medicare 226, Medicare Only 173, Medicare & Medicaid 105). Lacking predictive power, all variables that were classified as imbalanced were removed.

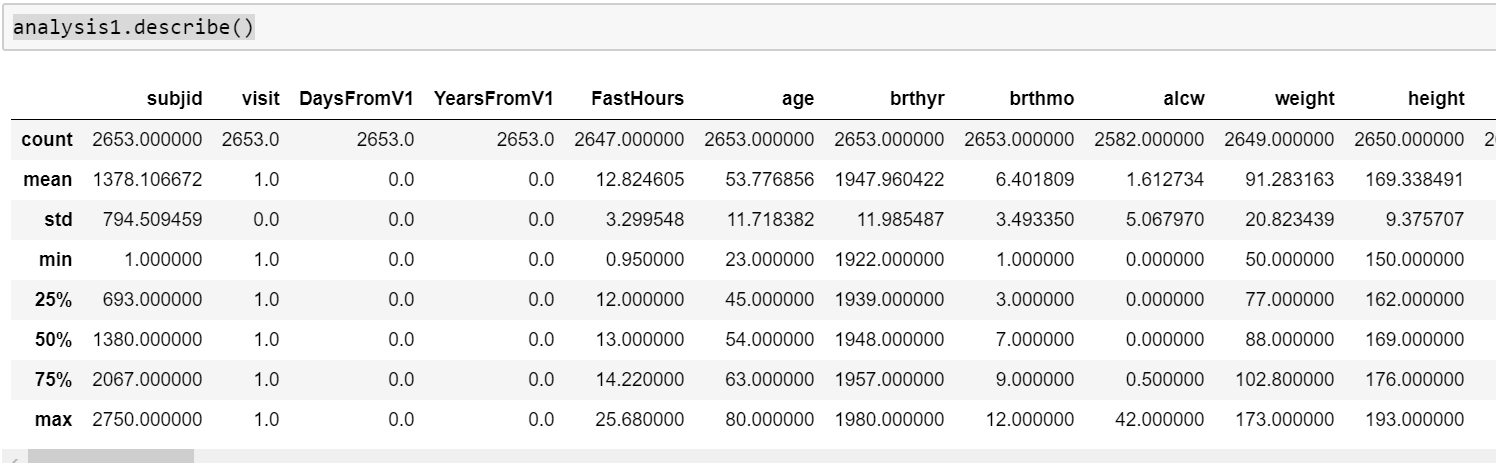
In this work, histograms were applied to the remaining variables to classify the shape, span, and outliers. Graphical summaries of the JHS dataset identified that the majority of the data were uniformly distributed. In addition to histograms, heatmaps were utilized to provide a visual account of missing data along with frequency tables. While variables such as Age and Sex have complete data, other measures require modifications due to it being incomplete. Several strategies can be implemented in the abscence of complete data. Imputation, a deep learning technique can be achieved in a variety of ways. Ranging from simple to overcomplex, each approach has pros and cons. The imputation method of calculating the median of the non-missing values was accomplished. Incomplete categorical variables were not altered. Lastly, records where the stroke value were missing were removed from the dataset.

**Summary of Data Analysis and preprocessing:**

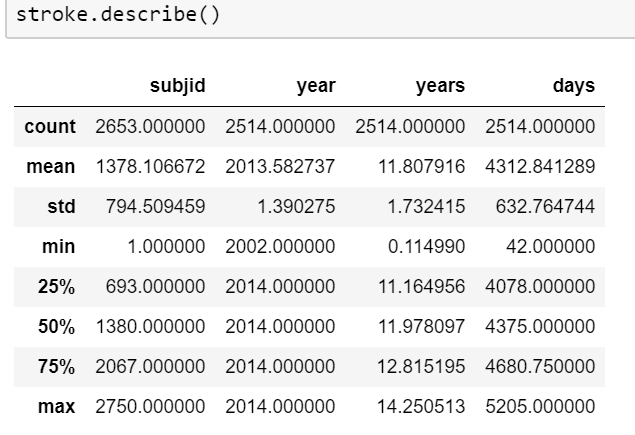
Data files: We requested the data from JHS through submitting a form and they send us the link to download the files. Below are the 2 files that we used for our research.

1. analysis1.csv – File has all the predictors.
2. incevtstroke.csv – File has the Target variable.

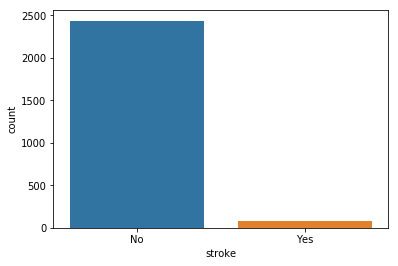
Initial level of summary for analysis1: We are having 2653 records and 211 variables.



Initial level of summary for incevtstroke: We are having 2653 records and 8 variables.



Below is the distribution of the target variable “stroke”. Looks highly imbalanced.

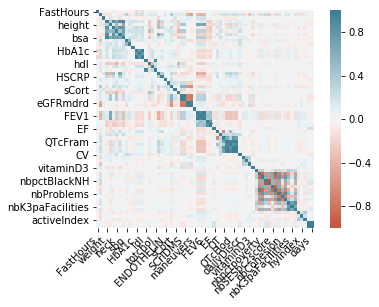


The above two dataframes were then merged to get out final full file for data analysis and model development.

We then dropped the variables with more than 90% of missing value. Also we dropped the variable that we thought that does not have predictive power. After these steps we were left with 138 variables.

Correlation Matrix:

We then worked on Correlation Matrix and heatmap. Depending on the correlations, we dropped the variables with high correlations.

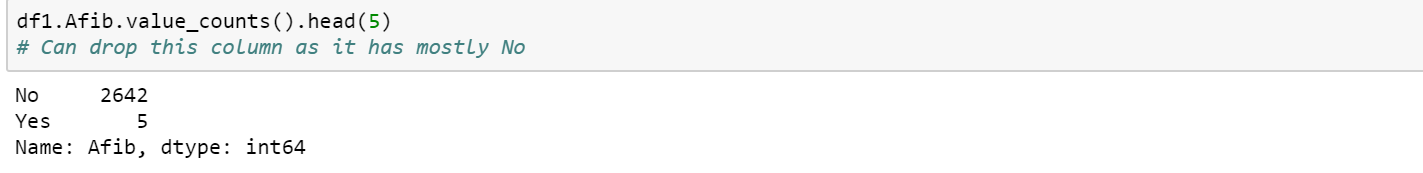


We then started to analyze each and every variable for their predictive power and see if the variables contains values with high count and other values contributing very little and dropped all those variables.

Example below:

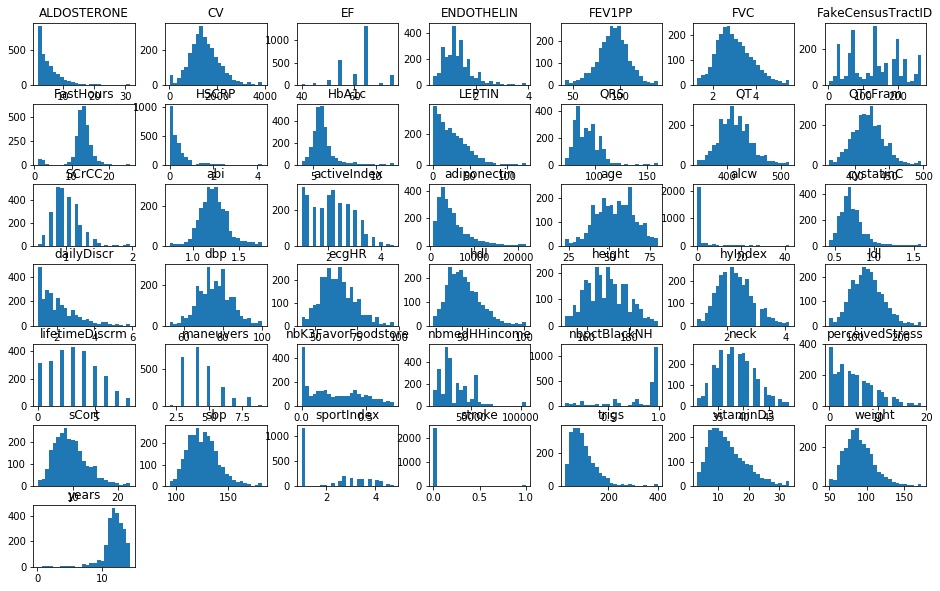


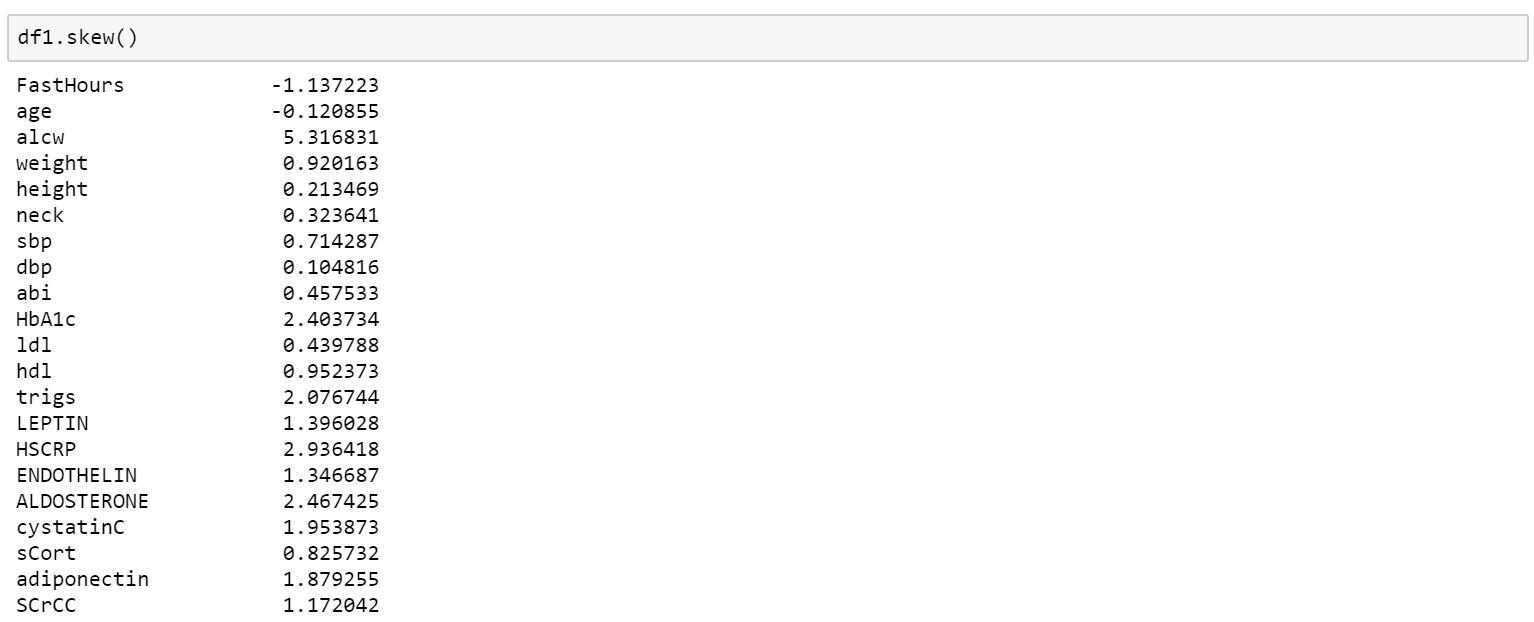




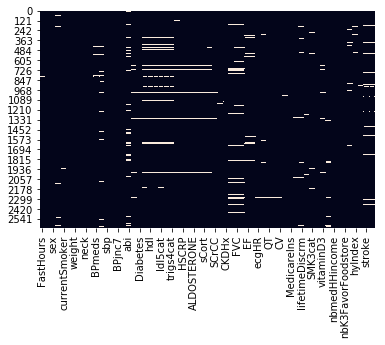


For the distribution of variables, we plotted Histograms: For most of the variables the distribution looked normal. But we still had few variables with left or right skewness.

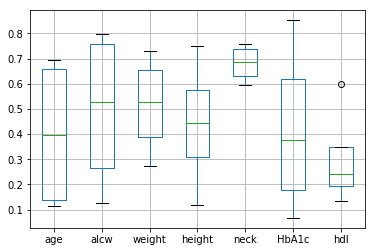


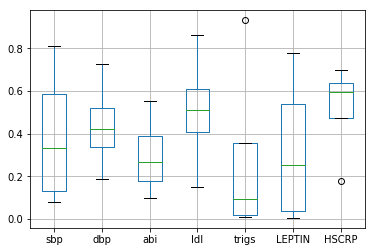


Missing value: We plotted heat map and looked for missing values. We have quite a few missing values. Replaced then with median values.



Outlier Detection: We used Box plot to detect the out liers. Box plot uses IQR method to detect the outliers.





After Data analysis and preprocessing we were left with 2653 rows and 62 variables.

Now lets see our Models:

**Methods**

**Logistic regression:**

Used on target variables that are categorical, a logistic regression are useful in the prediction of probability. The logistic regression model predicts P(Y=1) as a function of X. In order to keep outcomes in the range of 0 and 1 the logistic function is applied.



Data in this model were created only with the features and dependent variables while dummy variables were converted for select features. A variety of techniques were employed in order to split the data and handle outliers. The target variable was dropped from the model while the coefficients were obtained and placed within a dataframe. Regression coefficents in this particular model symbolize the transformation in the logit for each unit change in the predictor.

**K-Folds cross-validation:**

The K-folds cross-validation procedure evaluates machine learning models based on parameter k. Once a specific value for k is selected, the data is split into as many samples. Used to estimate the skill of a model, this technique divides the data into folds and tested based on the number set for the k parameter. In each iteration, the data is trained on each iteration of the K-Fold process then each score is appended and a mean is obtained to determine the accuracy of the model.

**Logistic regression with SGD (stochastic gradient descent):**

Stochastic gradient descent is an iterative algorithm that identifies the minimum of a function. In machine learning, this technique is utilized to revise the paramters of models. Considered fast, stochastic gradient descent computes the derivative from training data occurrence and calculates the update. Samples are randomly selected. In this work, 30 fits which were comprised of 5 folds for each of 6 candidates built the model.

**Random forest:**

Built on the premise of decision trees, the Random forest algorithm operates as an collaborative. In lay terms, the prediction is based on the tree within the ensemble with the most votes. Within the algorithm, k is randomly selected from the total and then CART (Classification And Regression Trees) is computed. Random forest has processes for balancing errors in data sets where classes are imbalanced.

**K-NN:**

Known as a statistical method that doesn't require data to fit a normal distribution, K-NN (k-nearest neighbors algorithm) finds the distances between a query and features in the data. The algorithm sums the distance and the index of the sample to an systematic assembly. It sorts the collection and picks the first k-entries. In case of outcomes, if it is a regression K-NN returns the mean of the k-labels, if a classification, it returns the mode of the k-labels.

**Experimental Results and Analysis**

**Model Comparison using ROC**

Performance of a classification model is graphically illustrated by the AUC (Area Under Curve) - ROC (receiver operating characteristic) curve. AUC exemplifies degree or measure or separability while the ROC serves as a probability curve. Within this work, the ROC curve is plotted with true positive rate (y-axis) against false negative rate (x-axis).

True Positive Rate



False Positive Rate



Models considered poor exhibit an AUC near to the 0 while an excellend model is near to 1; considered a good measure of separability.

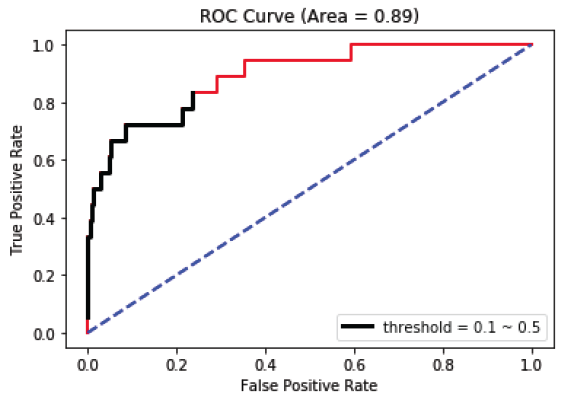
Based on the models output, the following table will highlight the comparison of the SGD Logistic Regression, Random Forest, and K-NN models; the model will display the Area Under Curve/Receiver Operating Characteristic scores with the output from the confusion matrix.

The confusion matrix showcases the predicted and actual class labels from the models. The 2-by-2 matrix from left-to-right are as follows: True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN).

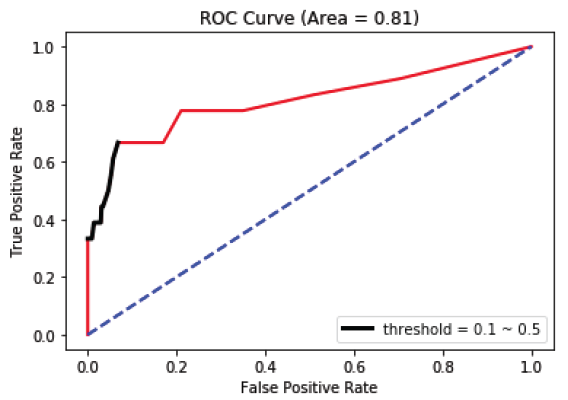
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | ROC AUC Training | ROC AUC Test | True Pos (Test) | False Neg (Test) | True Neg (Test) | False Pos (Test) | Accuracy Training | Accuracy  Test |
| Logistic Regression w/SGD | 0.920286034963608 | 0.8931028551771586 | 646 | 0 | 18 | 0 | 0.9693313222724987 | 0.9728915662650602 |
| Random Forest | 1.0 | 0.8126504987960095 | 646 | 0 | 17 | 1 | 1.0 | 0.9743975903614458 |
| K-NN | 0.9573540915583973 | 0.8279583763329894 | 646 | 0 | 18 | 0 | 0.9693313222724987 | 0.9728915662650602 |
| Logistic Regression | 0.7118180736004354 | 0.6612487100103198 | 639 | 6 | 12 | 6 | 0.9798893916540975 | 0.9713855421686747 |
| XGBoost | 0.9553571428571428 | 0.6298486161374178 | 852 | 6 | 17 | 1 | .992 | .9795 |

**ROC curves:**

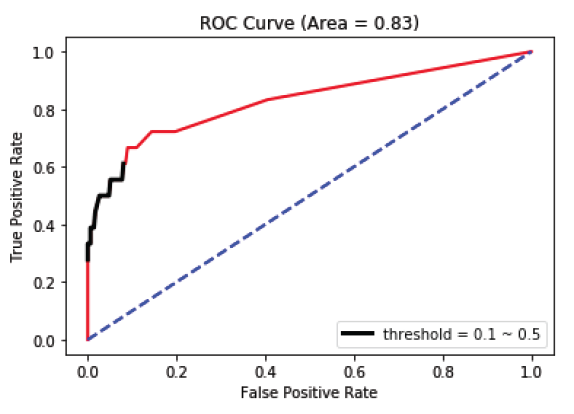
**Logistic Regression with SGD**



**Random Forest**



**K-NN**



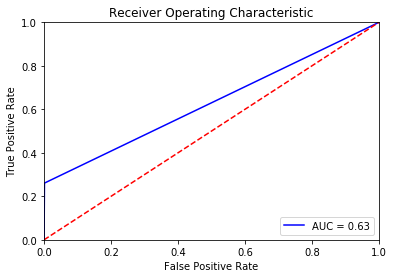
We then tried XGBoost method :

**XGBoost:**

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

Think of XGBoost as gradient boosting on ‘steroids’ (well it is called ‘Extreme Gradient Boosting’ for a reason!). It is a perfect combination of software and hardware optimization techniques to yield superior results using less computing resources in the shortest amount of time.

But we saw that even though the accuracy was 98%, the ROC AUC was just 62%. Please find below the ROC for XGBoost.



Conclusion

In order to perform machine learning on JHS data set, the most important steps are conducted during the data preprocessing phase. Its during this phase where data scientist can determine which variables will work best during the creation and implementation of the models. According to the findings from the models, the Random Forest algorithm performed best on the training set for the ROC AUC score and the accuracy scores for the test and train sets. On the test set for the ROC AUC test the Logistic Regression w/SGD performed best.

When something is concluded true and it is actually false, we have a false positive or type I error. On the other hand, when something is false and it is actually true, we have a false negative or type II error.

All medical tests can be resulted in false positive and false negative errors. Since medical tests can’t be absolutely true, false positive and false negative are two problems we have to deal with. A false positive can lead to unnecessary treatment and a false negative can lead to a false diagnostic, which is very serious since a disease has been ignored. So our team thinks we should choose the model with least false negative rate.

**Appendix:**

**Final list of variables considered for the model:**

'FastHours','age','sex','alcw','currentSmoker','everSmoker','weight','height','neck','OBESITY3cat',

'BPmeds','diureticMeds','sbp','dbp','BPjnc7','HTN','abi','HbA1c','Diabetes','ldl','hdl','trigs',

'ldl5cat','hdl3cat','trigs4cat','LEPTIN','HSCRP','ENDOTHELIN','ALDOSTERONE','cystatinC','sCort',

'adiponectin','SCrCC','DialysisEver','CKDHx','maneuvers','FVC','FEV1PP','EF','EF3cat','ecgHR','QRS',

'QT','QTcFram','CV','PrivateIns','MedicareIns','dailyDiscr','lifetimeDiscrm','perceivedStress',

'SMK3cat','BMI3cat','vitaminD3','FakeCensusTractID','nbmedHHincome','nbpctBlackNH',

'nbK3FavorFoodstore','sportIndex','hyIndex','activeIndex','years','stroke'

**Machine Learning Mode Code:**

