Prediction of Stroke

A Project Towards Partial Fulfillment for
BADM 590 - Social Media Analytics Course, Spring-2018.

By

Amit Darekar

Index

1. Introduction:

Application of analytics in healthcare is the new buzz in the industry. It's transforming the way healthcare industry problems are solved. Leading consulting organizations such as McKinsey and Co., Deloitte, EY, KPMG and PwC are offering services in this domain due to its high effectiveness.

Healthcare analytics have the potential to reduce costs of treatment, predict outbreaks of epidemics, avoid preventable diseases and improve the quality of life in general. Average human lifespan is increasing along world population, which poses new challenges to today's treatment delivery methods. Healthcare professionals, just like business entrepreneurs, are capable of collecting massive amounts of data and look for best strategies to use these numbers. There's a huge need for big data in healthcare as well, due to rising costs in nations like the United States. In other words, healthcare costs are much higher than they should be, and they have been rising for the past 20 years. Clearly, we are in need of some smart, data-driven thinking in this area.

Analytics tools are helping with better patient outcomes while staying within the cost curves with data driven transformations. Big data and advanced analytics are changing the way healthcare companies do business. Pharmaceutical and Life science companies are increasingly relying on data driven decisions to influence their target market. For example, they make use of market research data on doctor's prescription and come up with target market for their products. Similarly, physician decisions are becoming more and more evidence-based, meaning that they rely on large swathes of research and clinical data as opposed to solely their schooling and professional opinion. As in many other industries, data gathering and management is getting bigger, and professionals need help in the matter. This new treatment attitude means there is a greater demand for big data analytics in healthcare facilities than ever before.

Electronic Health Records (EHR), the most widespread application of big data in healthcare, is filling this gap. Every patient has his own digital record which includes demographics, medical history, allergies, laboratory test results etc. Records are shared via secure information systems and are available for healthcare providers from both public and private sector. Every record is comprised of one modifiable file, which means that doctors can implement changes over time with no paperwork and no danger of data replication. EHRs can also trigger warnings and reminders when a patient should get a new lab test or track prescriptions to see if a patient has been following doctors' orders.

Our project is based on similar EHR collected over couple of years. The objective is to predict whether a patient will be affected by stroke or not based on this data. We have applied several machine learning algorithms to find best model to predict the outcome.

2. Project overview:

We have used data collected from “McKinsey Analytics Online Hackathon – Healthcare Analytics” conducted jointly by Analytics Vidhya and McKinsey and Company.

In the subsequent sections, we will see the problem statement, dataset used, and variables of interest. Next, we present preprocessing of data that includes data inspection, data cleaning, data imputation, and data visualization. After that we start processing data. We identify and transform variables of interest so that we can apply machine learning algorithms to dataset. We define performance matric to evaluate how good our recommended algorithms are. Finally, we conclude our project report by discussing how this work can be extended and ways to improve performance of algorithms recommended.

3. The task at hand:

3.1 The problem statement:

Our client wants to have study around one of the critical disease "Stroke". Stroke is a disease that affects the 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 bursts (or ruptures). When that happens, part of the brain cannot get the blood (and oxygen) it needs, so it and brain cells die.

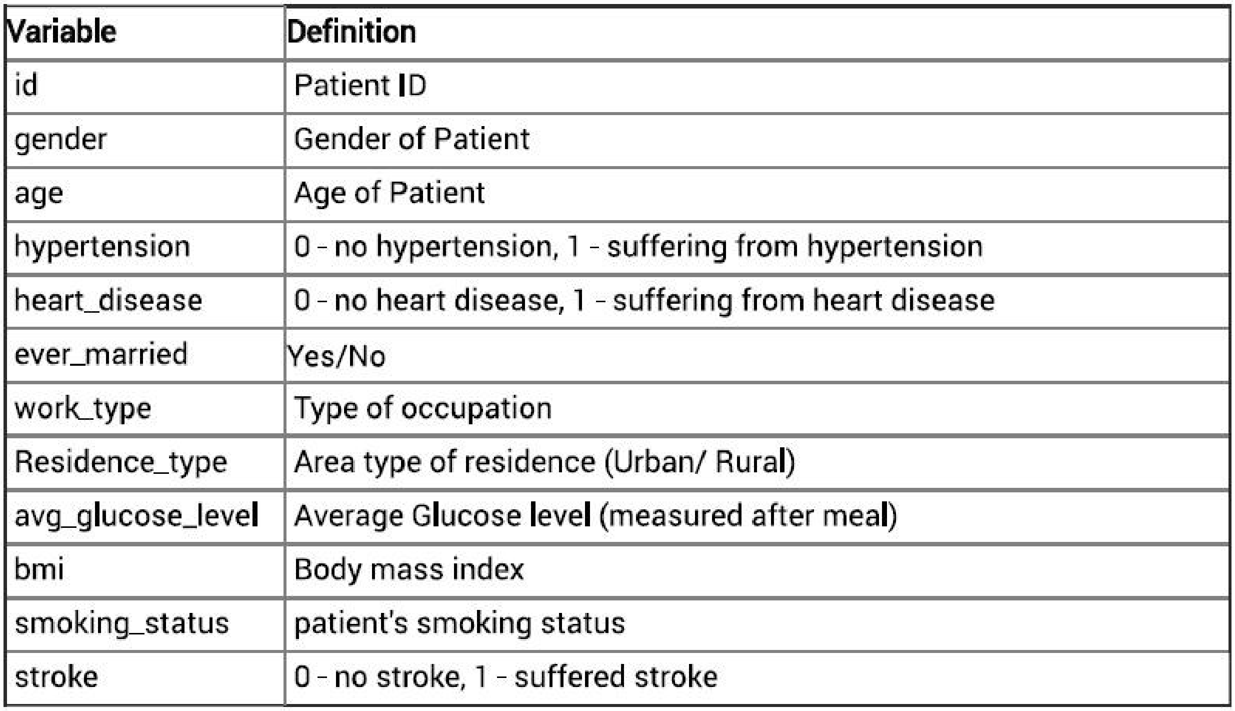
Over the last few years, the Client has captured several health, demographic and lifestyle details about its patients. This includes details such as age and gender, along with several health parameters (e.g. hypertension, body mass index) and lifestyle related variables (e.g. smoking status, occupation type).

The Client wants us to predict the probability of stroke happening to their patients. This will help doctors take proactive health measures for these patients.

3.2 Data used for project:

Our client has provided us EHR for 43,400 patients. To protect the identity of patients, the data is masked with dummy IDs. Our outcome variable is ‘Stroke’, which is binary variable. The data set contains several variables which are tabulated in below Data Dictionary.
ID is unique ID for each of the patient. Therefore, ID is just a label. The data dictionary is self-explanatory, clearly indicating whether a variable is categorical or continuous variable.

Data Dictionary



Variable	Definition
id	Patient ID
gender	Gender of Patient
age	Age of Patient
hypertension	0 - no hypertension, 1 - suffering from hypertension
heart_disease	0 - no heart disease, 1 - suffering from heart disease
ever_married	Yes/No
work_type	Type of occupation
Residence_type	Area type of residence (Urban/ Rural)
avg_glucose_level	Average Glucose level (measured after meal)
bmi	Body mass index
smoking_status	patient's smoking status
stroke	0 - no stroke, 1 - suffered stroke

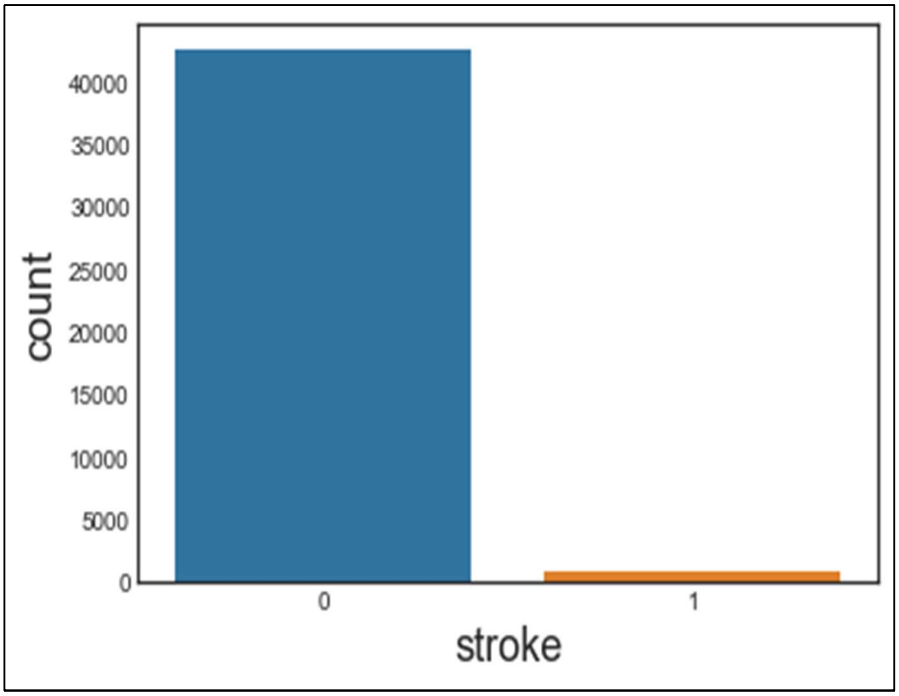
4. Data Preprocessing

4.1 Data Inspection

We begin our data preprocessing by inspecting dataset. We noticed that although our outcome variable 'stroke' has 43400 records, two variables namely 'bmi', and 'smoking_status' have missing values. Therefore, we need to impute missing values for these two variables.
Further, we notice that our outcome variable is un-balanced (Refer Image-1) It is quite understandable that people who don't suffer a stroke far outnumber people who actually suffer a stroke. This suggests that we must use a balanced weight approach for any of the classification algorithms we will use.

Apparently, our data is neat and structured. Therefore, we don't need to clean the data. However, we need to impute the data for missing value. Further, for the task at hand, we will not need patient ID. So, we will drop this variable from further analysis.

Image 1 – Count of people who are affected with stroke Vs those who have not



4.2 Data Imputation

We observe that we only have 30,108 values for variable 'smoking status'. To understand the distribution of data for each category, we considered count of patients. We observed that most of the values fall under category 'formerly smoked'. Therefore, to preserve the data distribution, we imputed missing values with 'formerly smoked' category.

Further, we visualized distribution of BMI values by plotting histogram (Refer Image-2). The literature suggests that Mode and Median are quite robust estimators for most of the distributions, and therefore least affected by outliers. Therefore, we decided to use Mode for variable BMI using rest all non-zero values as imputation for BMI. The data distribution for imputed values of BMI closely follows original data distribution (Refer Image 3).

Image 2 - Distribution of BMI pre-imputation

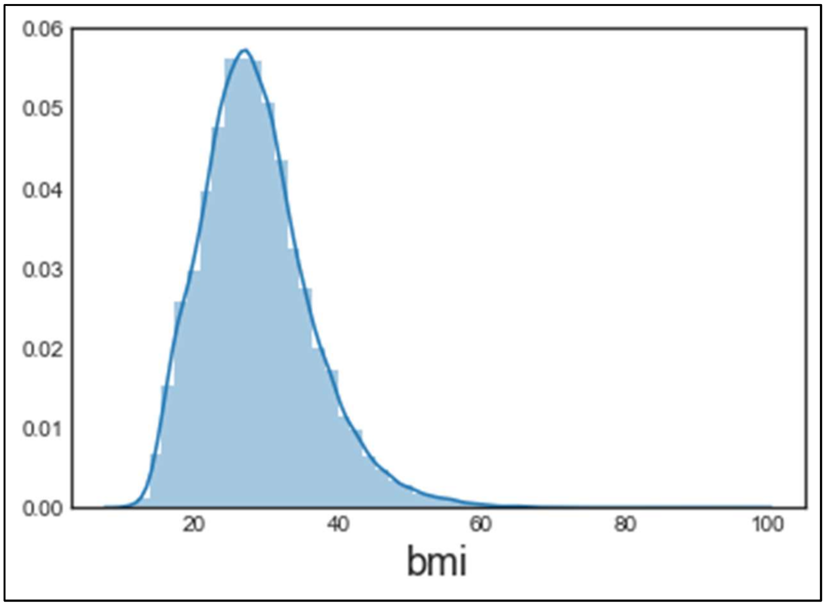
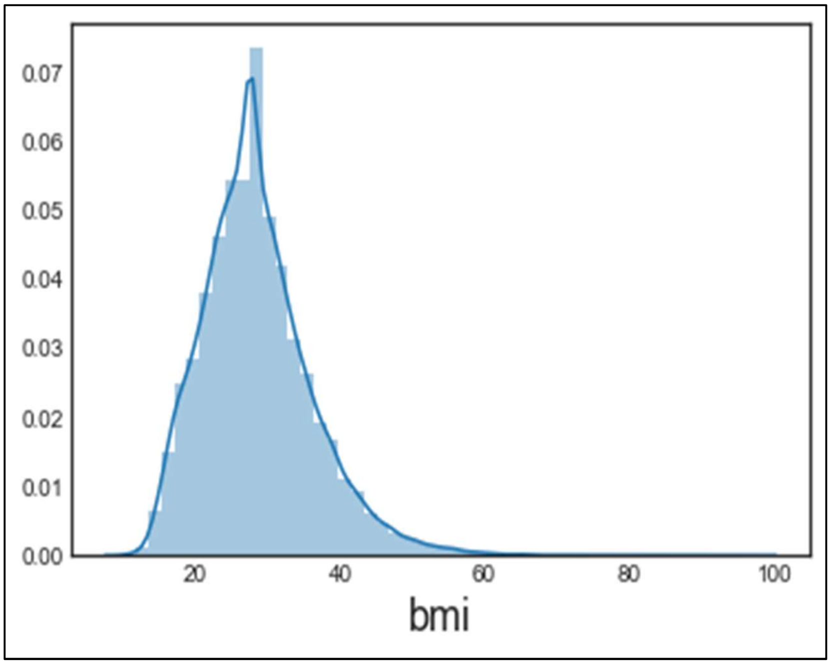


Image 3 - Distribution of BMI after imputation



4.3 Data Visualization:

To make our interpretation more intuitive, we tried to visualize variables of interest as follows:

When we viewed total counts by gender, we see that observations for female dominate over that of male (Image 4). Further, we plotted occurrence of stroke by gender, we see that male have higher occurrences of stroke (Image 5). Both this put together, we see that overall male have higher chances of suffering from stroke than that of female.

Image 4 – Distribution of data by Gender

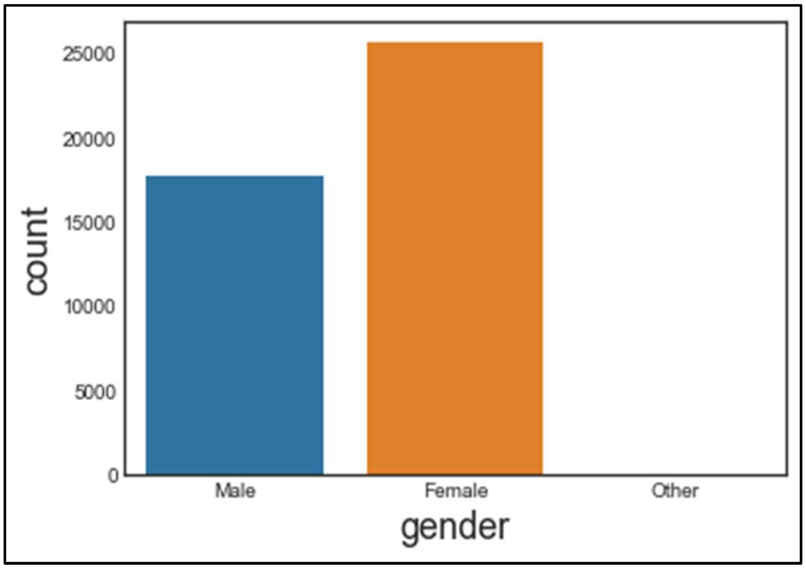
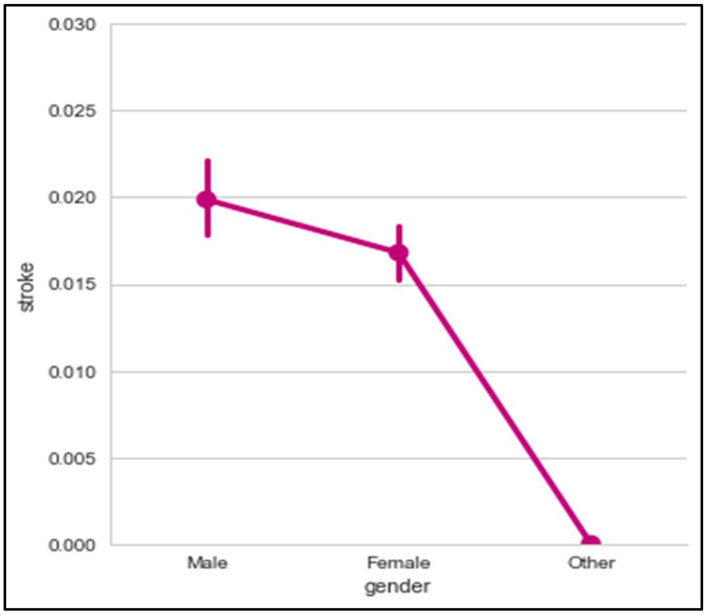


Image 5 – Occurrence of Stroke by Gender



Similarly, we plotted total count by people having prior history of hypertension (Image 6). Clearly, very few people in the sample have history of hypertension problem. When we also plotted stoke against hypertension, we clearly see that people who have prior history have hypertension have higher occurrences of stroke, which is almost 5 times likelihood of stroke than people who don't have hypertension.

Image 6 - Hypertension

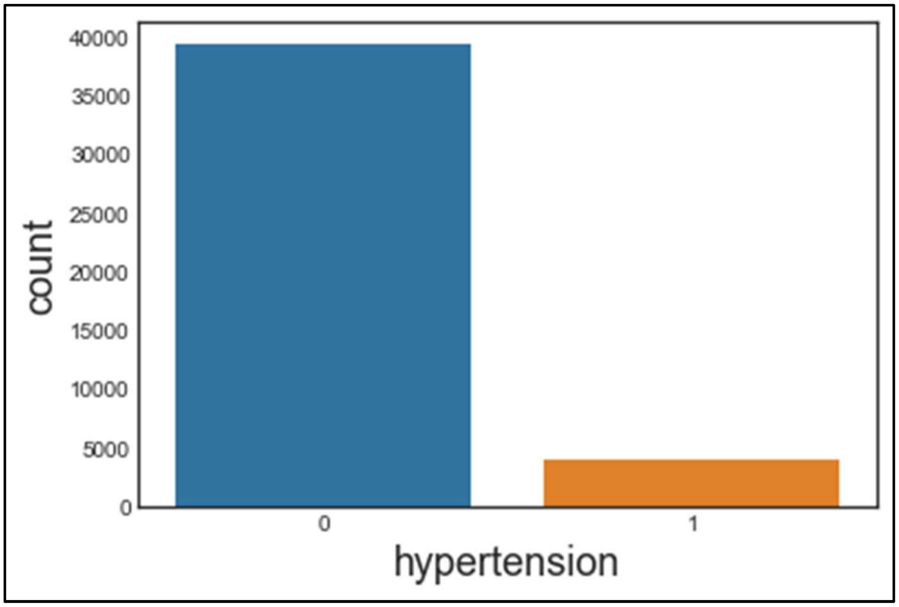
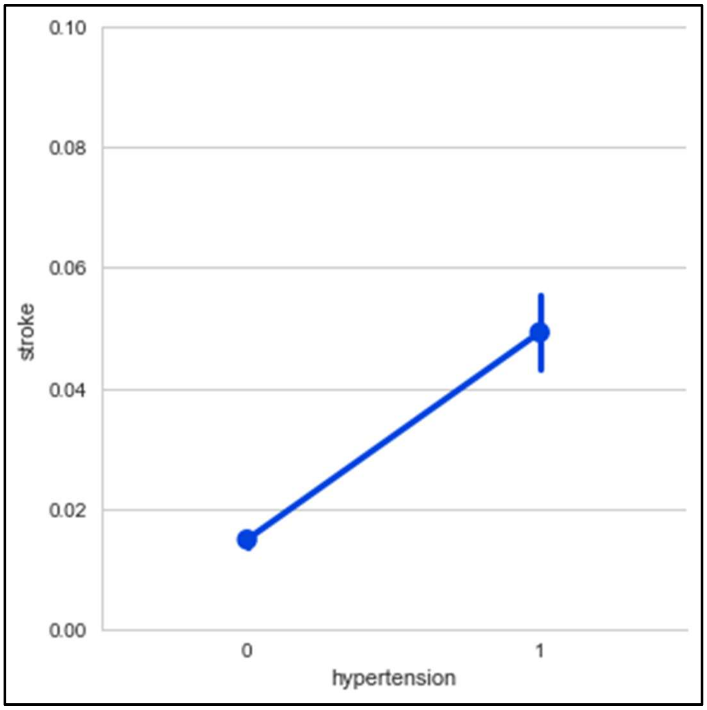


Image 7 – Stroke Vs Hypertension



Interestingly, we wanted to do similar comparison for people who are married. Our data contains most of the observations for married people (Image 8). We also plotted this variable against stroke (Image 9), and see that there are high chances that married people have almost 5 times higher chances of getting a stroke than the ones who never married.

Image 8 - Married vs not married

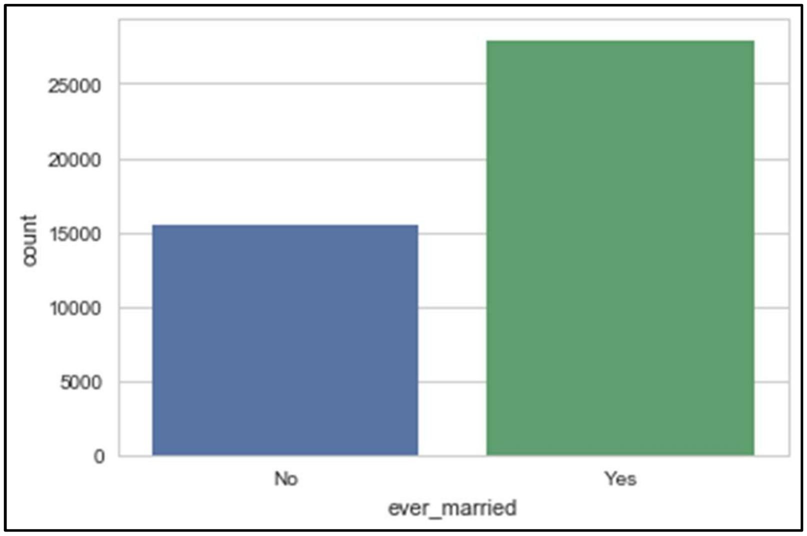
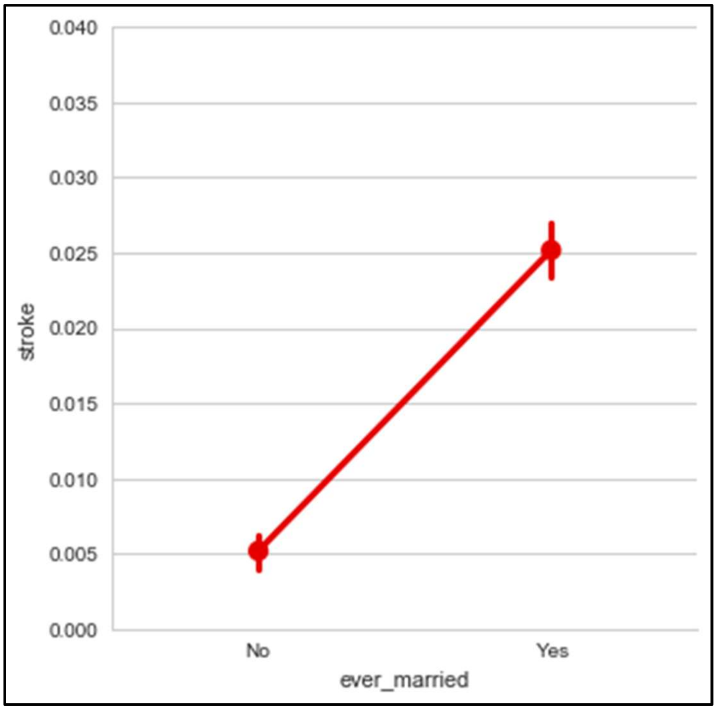


Image 9 – Stroke Vs Ever_Married



Next, we compared distribution of people by their work type (Image 10). When we combined this visualization with plot of Stroke Vs work type, we clearly see that despite self-employed people are being far few, the incidence of stroke among them is highest among all work-type classes. Possible explanation could be, self-employed people face a lot of stress in their work, and therefore have large number of stroke occurrences.

Image 10 – Work Type

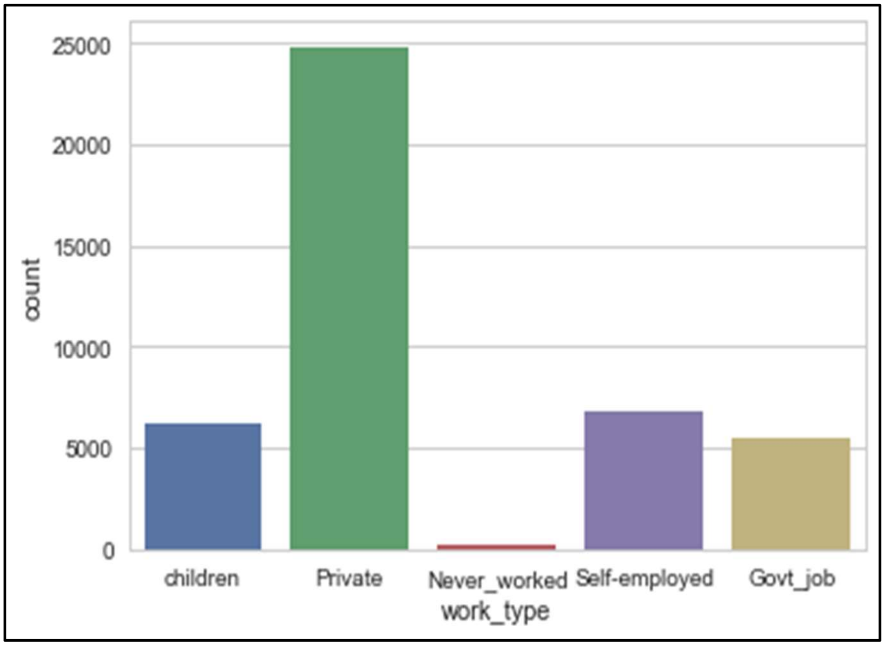
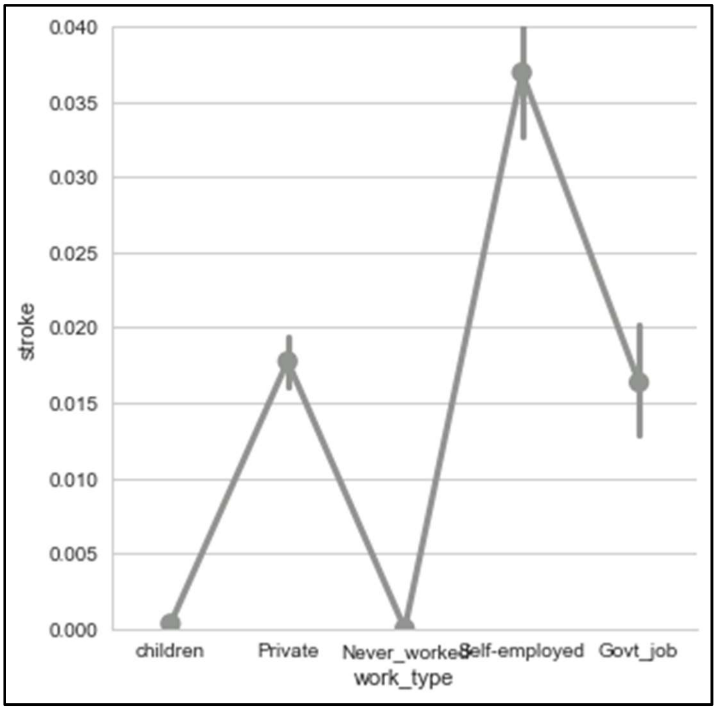


Image 11 – Stroke Vs Work-Type



We also wanted to visualize and understand whether ones' place of residence plays any role in getting a stroke (Image 12). We combine this with a plot of Stroke Vs Residence place, and found that there is no much difference in occurring stroke for people living in rural and urban areas (Image 13).

Image 12 – Place of Residence

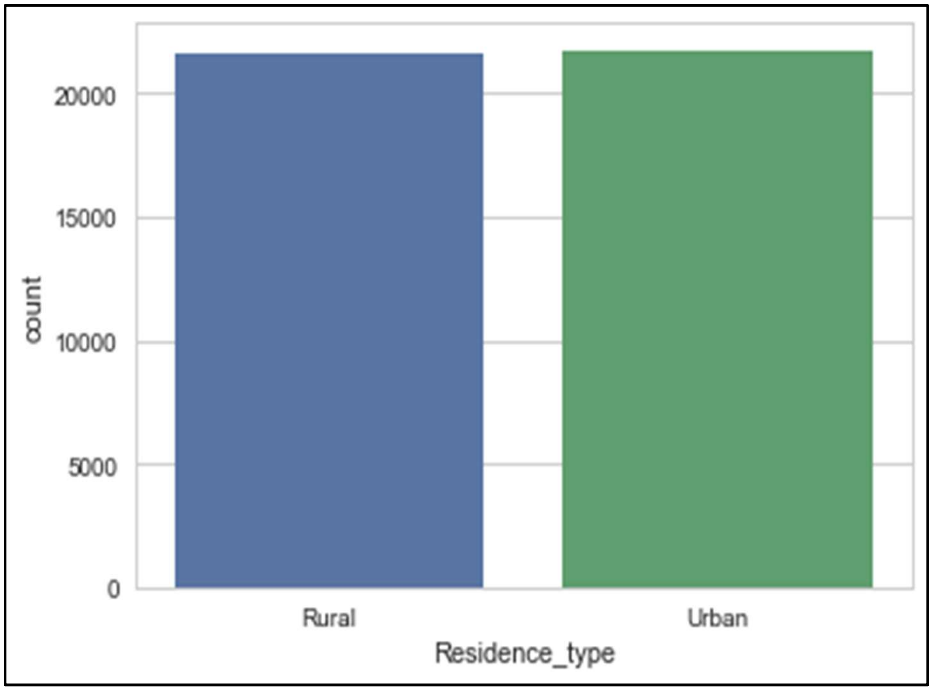
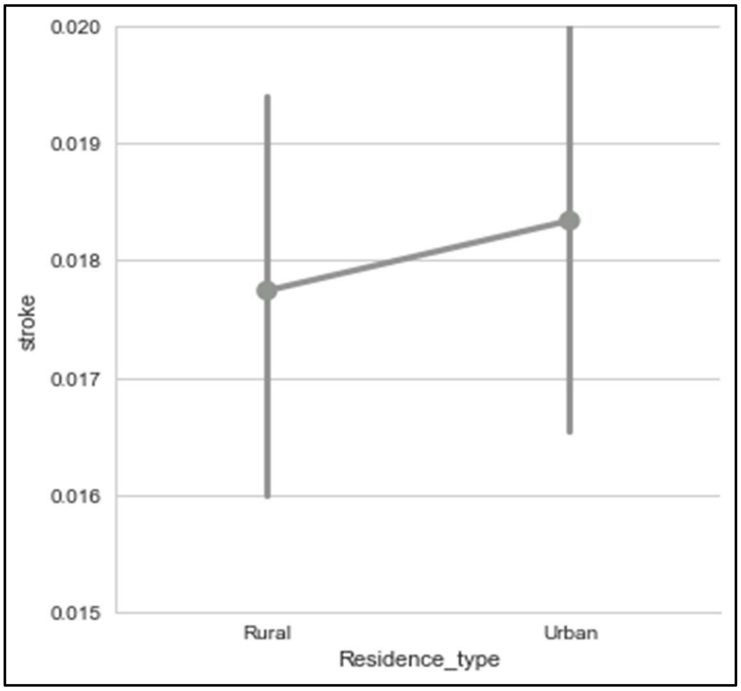


Image 13 – Stroke Vs place of residence



Finally, we wanted to see whether having smoking habits influence getting stroke(Image 14). As anticipated, the distribution is unbalanced with people ‘never smoked’ dominating dataset. We combined this with plot of Stroke Vs smoking status, people who formerly smoked and those who smoke have substantially higher occruences of stroke than people who never smoked (Image 15).

Image 14 – Smoking Status

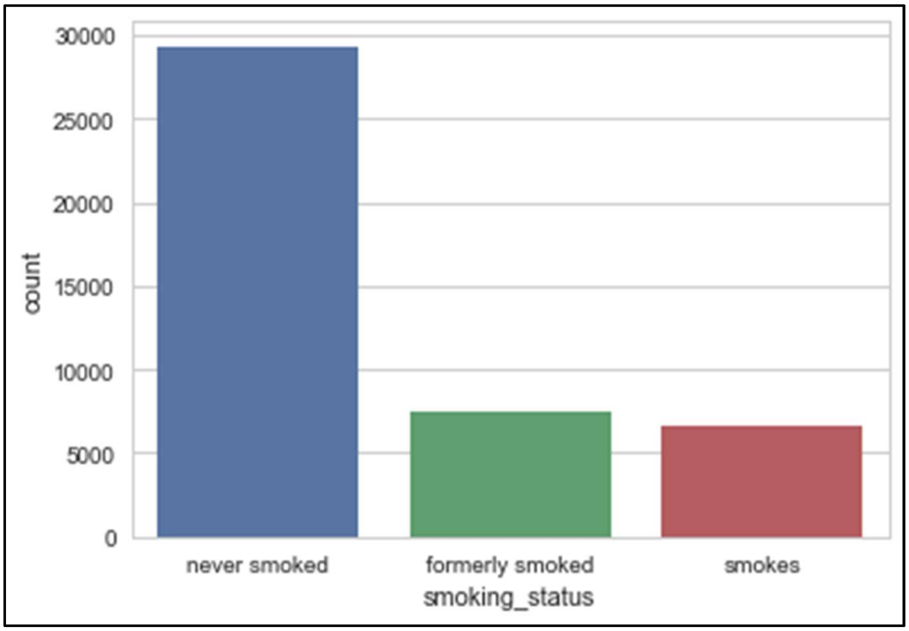
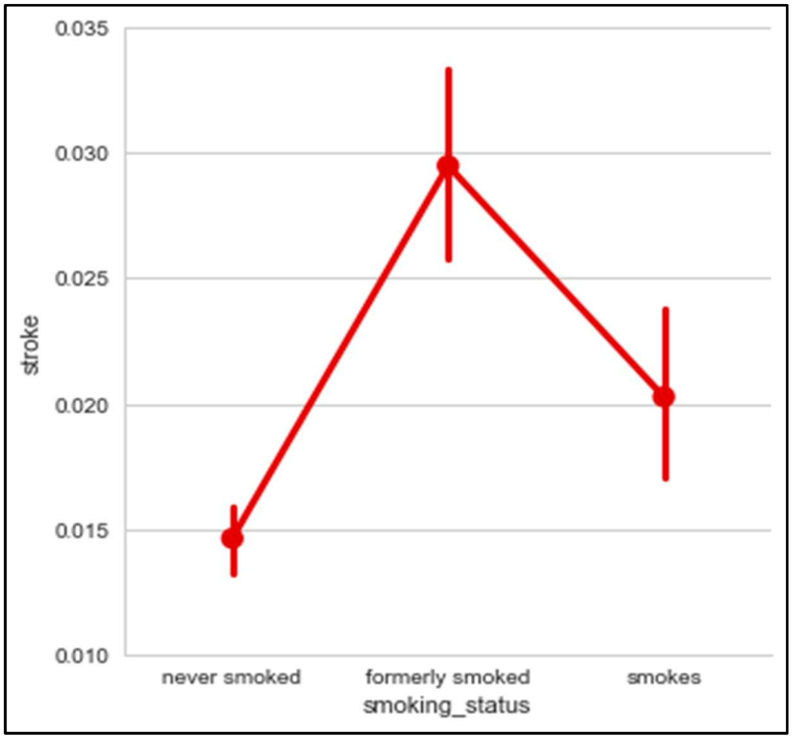


Image 15 – Stroke Vs Smoking Status



5. Data Processing:

We listed our variables into categorical and continuous variables. We have to create dummy variables for all categorical variables. So correct identification of type of a variable is important step in processing the data.

Our categorical variables are: 'gender', 'hypertension', 'heart_disease', 'ever_married', 'work_type', 'Residence_type', 'smoking_status'. Our continuous variables are: 'age', 'avg_glucose_level', 'bmi'.
Further our outcome variable is ‘stroke’.

In order to reduce overfitting concern, we split our dataset into training set and testing set. We kept 30% of total data as testing dataset, and rest all data is used to train our model/s.

6. Data Analysis

6.1 Performance Metrics

On the outset, we define three performance metrics for our algorithms. The first is accuracy of predicting outcome variables in the test set. This is computed as percentage of predictions that match with value of outcome variable in test set.

However, while the standard performance metrics are useful for a standard classification task, many algorithms now generate a probabilistic classification. As a result, we need a method to not only compare different estimators, but determine the optimal threshold for an estimator. To support this decision, we employ the receiver operating characteristic (ROC) curve. Originally developed during World War Two to predict the performance of an individual using a radar system, the ROC curve displays the relationship between the number of false positives (along the x-axis) and true positives (along the y-axis) as a function of probability threshold.

The ROC curve starts at the lower left, where nothing has been classified. From here the estimator is used to determine the true and false positives for very high probability thresholds. At this point, the curve should shoot upward from the lower left, wince we expect a good classifier (and what other type of classifier would we build) performs well at high threshold. In general, as the probability threshold is lowered, we will begin to predict more false positives, and thus the curve will shift to the right.

Given an ROC curve, another performance metric that can be measured is the area under the curve or AUC. In an ideal case this metric has the value of one, or perfect classification, and a random classification has the value of 0.5. This metric can provide a useful comparison between different estimators on the same data.

We will use these metrics to assess performance of different algorithms.

6.2 Algorithms

We have applied nine different classification algorithms, each of which having different strengths and weaknesses, as listed below:

1. Logistic regression
2. K Nearest Neighbor Classifier
3. Decision Tree Classifier
4. Gaussian Naive Bayes Classifier
5. Perceotron
6. SGD Classifier
7. Gradient Boosting Classifier
8. Support Vector Machines
9. Random Forest

Using these models, we calculated - Receiver Operating Characteristic Curve - Area Under Curve (AUC) with the following steps:

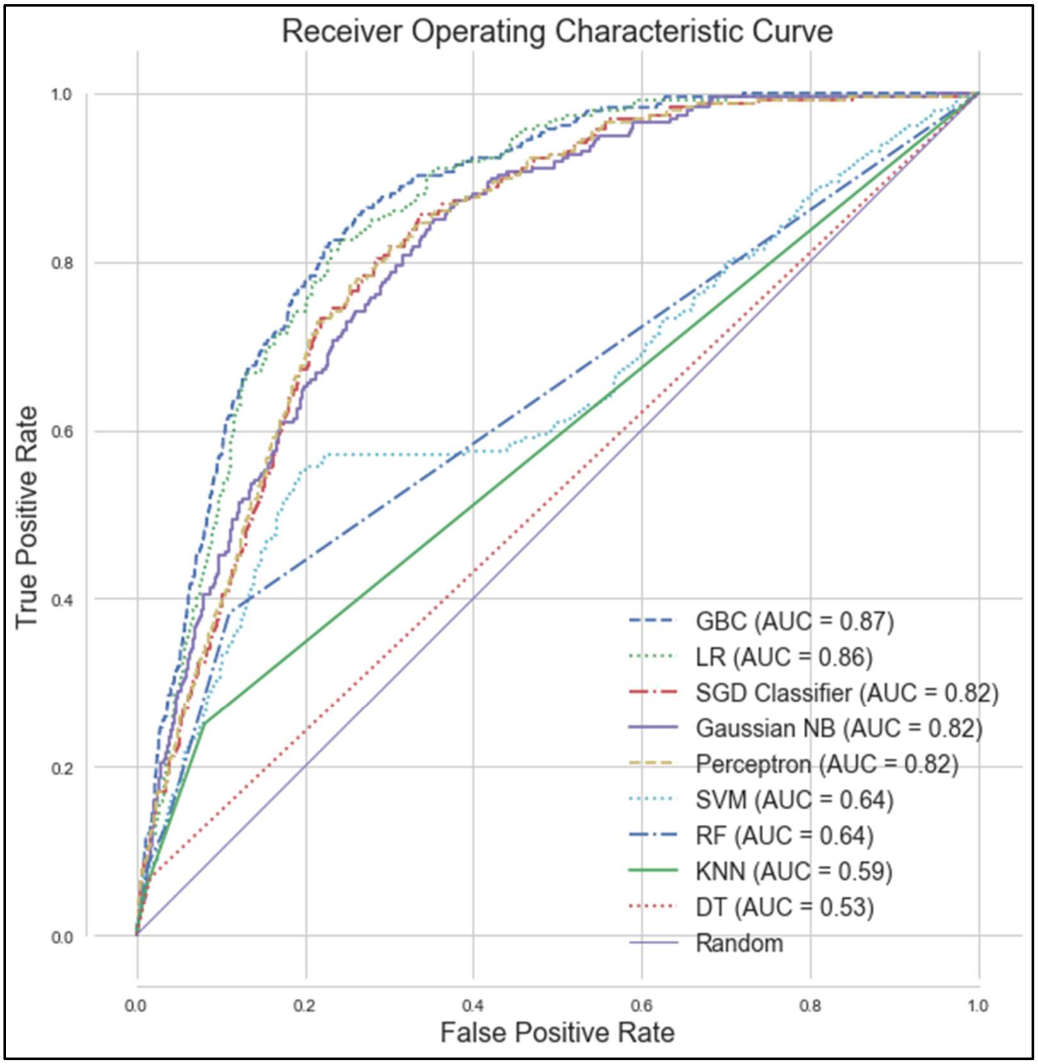
- Collect positive class probability for each model
- Compute ROC curve and ROC area for each model
- Plot the ROC curves

6.3 Performance

We summarize the results for various algorithms based on performance metrics as below:

Accuracy of Prediction

Sr	Algorithm	Accuracy %
1	K-Nearest Neighbors	98.17
2	Random Forest Classifier	98.14
3	Gradient Boosting Classification	98.13
4	Decision Tree	96.69
5	Support Vector Classification	93.72
6	Logistic Regression	73.80
7	Gaussian Naive Bayes	42.60
8	SGD Classification	11.77
9	Perceptron	10.67



7. Interpretation of Results

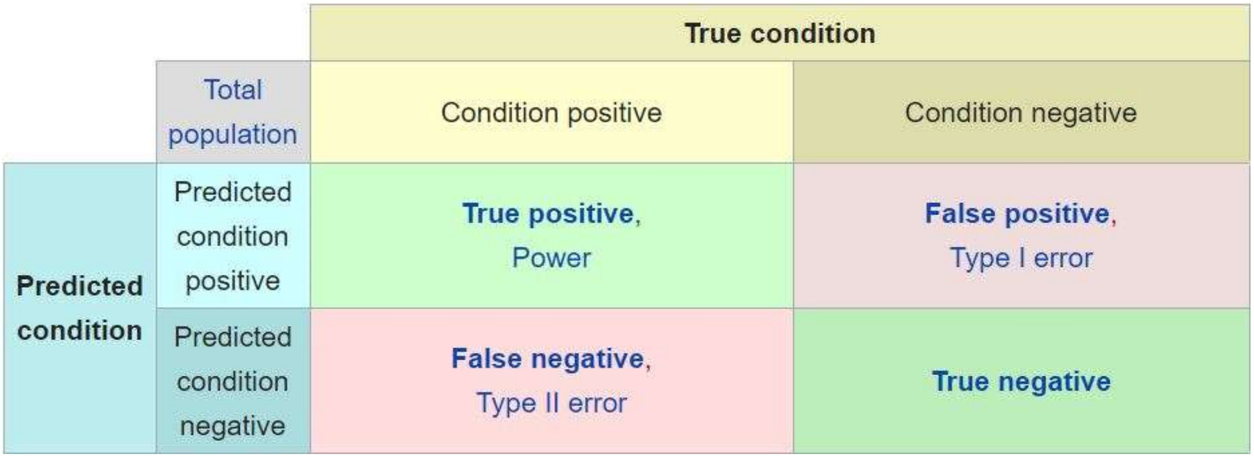
Based on accuracy score, we recommend KNN algorithm as best algorithm to solve problem at hand. 98.17% of cases are accurately predicted using this algorithm. However, Random Forest and Gradient Boosting Classifier also come very close to accuracy achieved by KNN, and therefore are close contenders.

Interpreting ROC Curve:

ROC analysis is part of a field called Signal Detection Theory, which is said to have been developed during the second World War for analysing radar images. Radar operators had to interpret whether the blip on the screen was an enemy target or a friendly ship. Signal detection theory measures this ability of the operators to classify the blip, and their ability to do so was referred to as Receiver Operating Characteristic. ROC curve is typically used to understand the performance of a binary classification, where there are only 2 possible outcomes. Here are some important characteristics to understand:

- Predicted probabilities are unlikely have smooth distribution
- ROC curves are useful in spite of the prediction not being properly calibrated
- We have to maximize the True Positive

The area under the curve measures the ability of the test to correctly classify the patient (in this case). The curve is essentially a plot of true positive rate against the false positive rate at different thresholds. True positive rate can also be called sensitivity and the ROC plot is simply the sensitivity as a function of the false positive rate. Here's an illustration (source - wikipedia.org):



		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

The best possible prediction result here would be on the top left corner, which would yield 100% sensitivity and 100% specificity. This can be called the perfect classification having the co-ordinates (0,1). The closer your curve gets to it, the better is the prediction.

Based on ROC and AUC criterion, Gradient Boosting Classifier performs a good job in predicting true positive. When we consider all three criteria together, Gradient Boosting Classifier algorithm bags the top position, and therefore we recommend using it.

8. Future extension of the project

All of the algorithms that we have used here are in their basic form. We simply used default hyperparameters. The only change we made was the use of 'balanced weight' approach since our outcome variable contains observations that are not balanced.

However, this is not the best performance that these algorithms could achieve. Infact algorithm such as Random Forest are much more powerful and capable of achieving higher accuracy.

Methods such as cross-validation can be used for fine-tuning hyperparameters for each of those algorithms. Model selection approach such as grid-search can help us select the best model out of a set of different models.

Therefore, we accept the limitations in our model, and propose further studies in future to enhance our recommended models.