**Q&A on Stroke Prediction using Machine Learning**

1. Why do we need Machine Learning for predicting Stroke?

Stroke depends upon lot of risk factors such as hypertension, age, bmi, heart disease etc. Now it requires specialized medical experts to analyze all these factors manually for a patient and determine whether he/she will have stroke in future. This process is time consuming and human dependent. With the help of Data Science, we can automate this process and remove human dependency.

1. Does bmi have any impact on Stroke?

Before performing the EDA for this project, I had an assumption that bmi plays an important role in determine the stroke risk factor. But the EDA results shows that stroke risk is high between 28-30bmi, but after that its dropping. So, we can’t say that higher the bmi is, the chances of having a stroke is more.

1. How smoking affects stroke risk factor?

Persons who previously smoked or are current smokers, have more chances of getting stroke than non-smokers.

1. What are the most influencing risk factors for stroke?

Based on the EDA, age and blood glucose level are the two most influencing factors for stroke.

1. What are the other influencing factors for stroke?

smoking habit, hypertension and heart disease can increase the probability of future stroke.

1. How age impacts stroke risk factors?

Chances of stroke is pretty low in young ages, till 40yrs old – stroke risk factor is negligible. The risk of having a stroke starts increasing from 50yrs and is at highest at 80yrs old.

1. How blood glucose level influences stroke?

Low blood glucose level (below 90) patients are at highest risk of having stroke. Probability of having stroke is pretty low for patients with glucose level 100-120. Patients with high glucose level possess moderate risk of having a stroke.

1. Why you have selected Random Forest as the best model in stroke prediction?

Random Forest model average precision score is better than other two models. Also, Random Forest is a combination of multiple decision trees. So, it is easier to explain it to other stakeholders than another complex model such as ANN.

1. Why you have used SMOTE function?

This dataset is highly imbalanced. Training a model with imbalanced dataset can cause issues. SMOTE function is used to over sample no-stroke counts to create a balance between stroke and no-stroke patients counts in training dataset.

1. Why the test f1-score is low?

My model performed well in identifying the majority class, but it performed poorly in detecting the minority class. I will try to use other deep learning models to see if it increases the f1-score.