PLAYGROUND SERIES SEASON 3 EPISODE 2

TABULAR CLASSIFICATION WITH STROKE PREDICTION DATASET

PREPARED BY: Dwaipayan Mukherjee (Emp. Id. 2211569)

Prepared as part of Kaggle projects given by Mr. Mrityunjoy Pandey.

**About the Project:**

The dataset for this project was generated from a deep learning model trained on the Stroke Prediction Dataset. Feature distributions are close to, but not exactly the same, as the original.

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

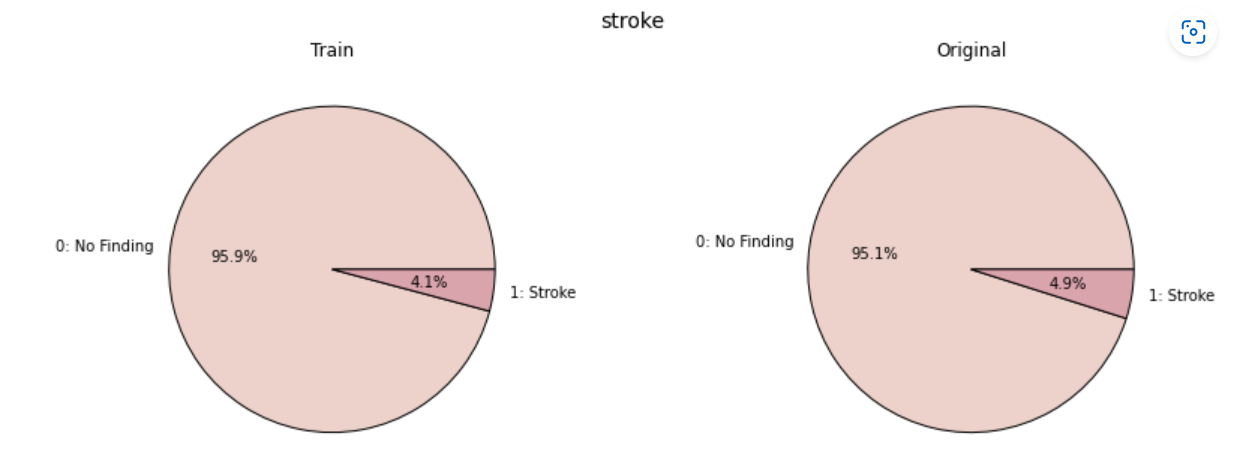
***Attribute Information:***

1. **id**: unique identifier
2. **gender**: "Male", "Female" or "Other"
3. **age**: age of the patient
4. **hypertension**: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5. **heart\_disease**: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6. **ever\_married**: "No" or "Yes"
7. **work\_type**: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"
8. **Residence\_type**: "Rural" or "Urban"
9. **avg\_glucose\_level**: average glucose level in blood
10. **bmi**: body mass index
11. **smoking\_status**: "formerly smoked", "never smoked", "smokes" or "Unknown"\*
12. **stroke**: 1 if the patient had a stroke or 0 if not

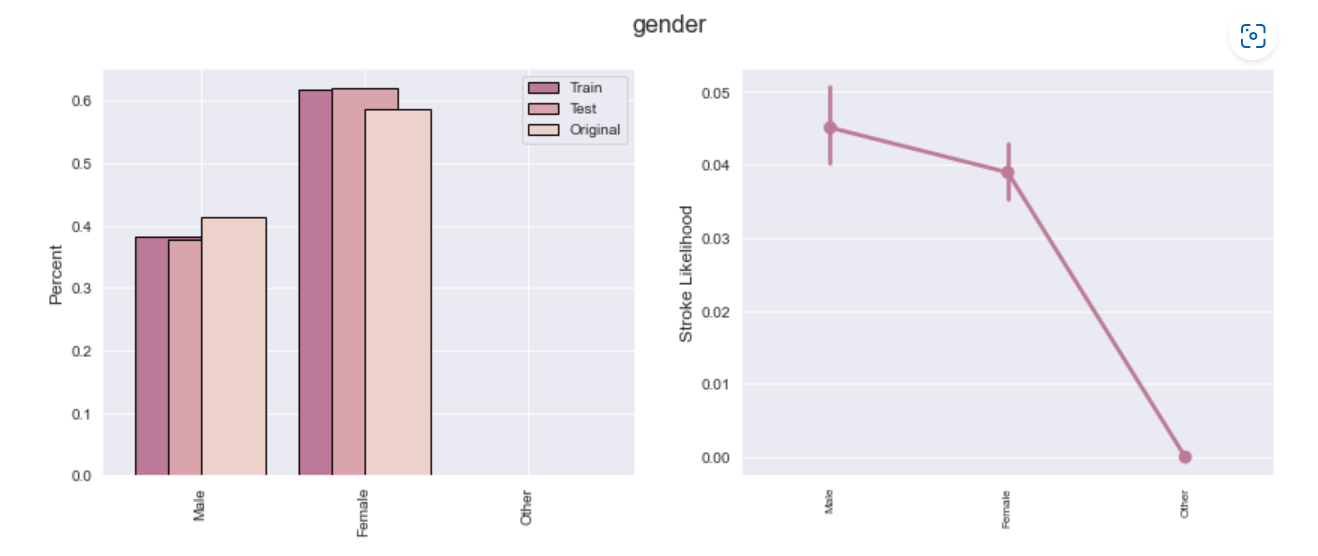
\*Note: "Unknown" in smoking\_status means that the information is unavailable for this patient

**Steps that I have followed throughout the project:**

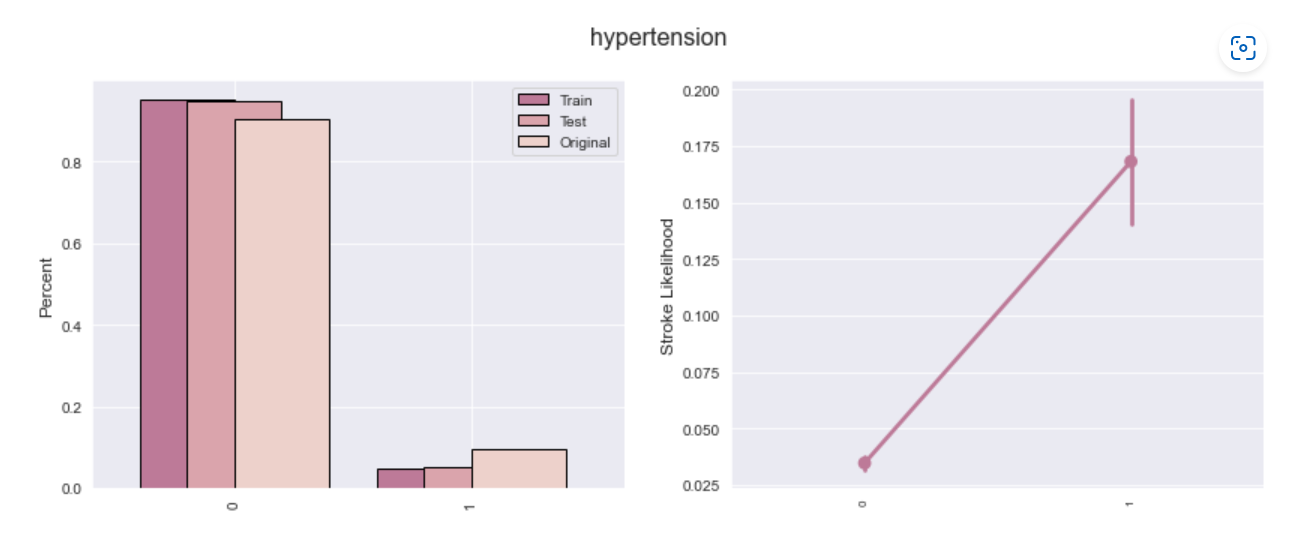
1. I have loaded the training, test, original Stroke Prediction data and the sample submission csv files as Pandas dataframe.
2. Next I have checked for the missing values in the training, test and original dataset. The training and test dataset had no missing values. The original Stroke Prediction dataset had 201 missing **bmi** values.
3. Next, I have tried to understand the dataset. The first job of understanding the dataset was to find the summary statistics (or, descriptive statistics) of the train, test and the original dataset. Next, I have enumerated the unique value counts as well as the unique values of each column in the three datasets to see if they are matching with the attribute descriptions.
4. Next, I have performed the Exploratory Data Analysis. I give my findings as follows:
5. The Train dataset represents the Original dataset well.



1. Males have a higher likelihood of a stroke than females. Others have zero likelihood since they are not in the dataset.



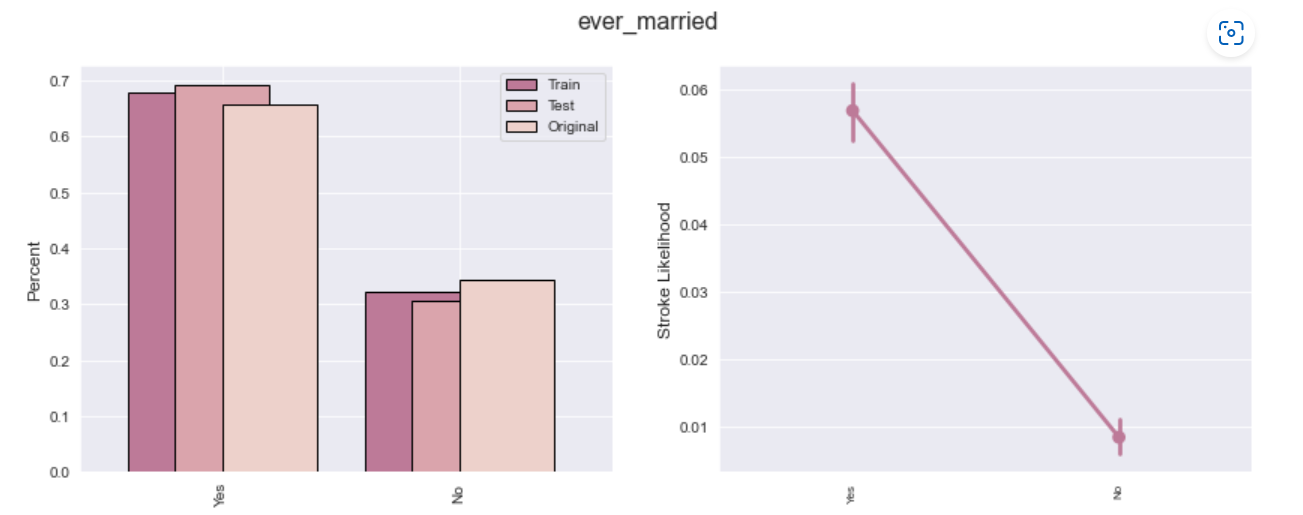
1. Only a small fraction of the population has hypertension. People having hypertension have a higher likelihood of having a stroke.



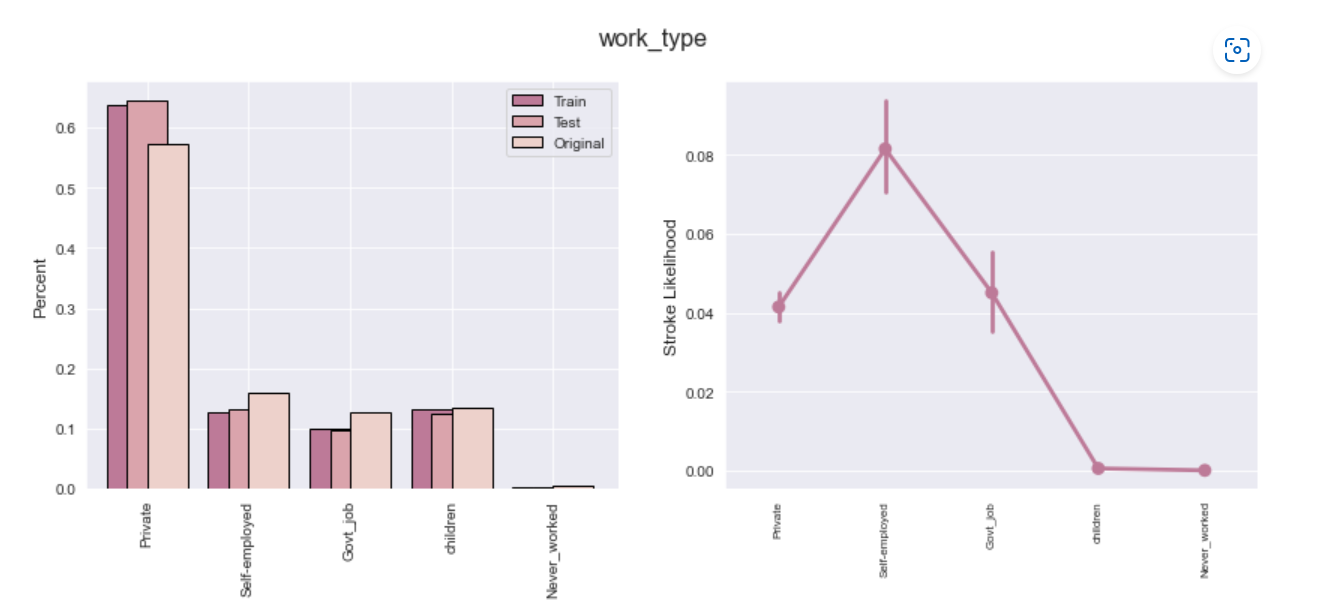
1. Only a very small portion of the population has heart disease. People having heart disease have a higher risk of stroke.



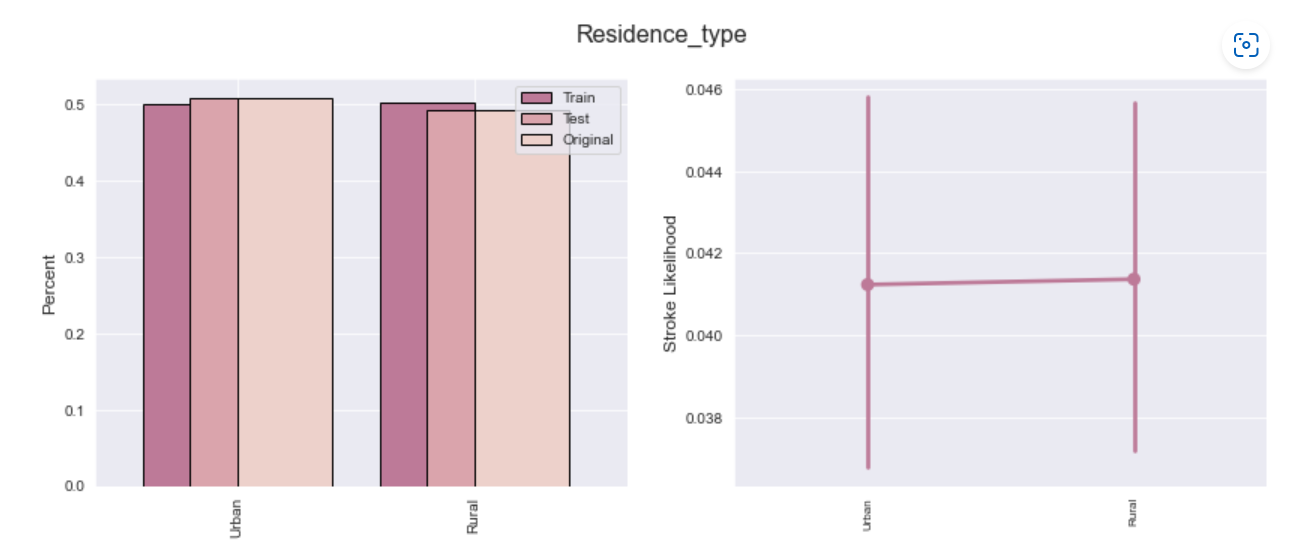
1. People who are ever married have a higher likelihood of experiencing a stroke.



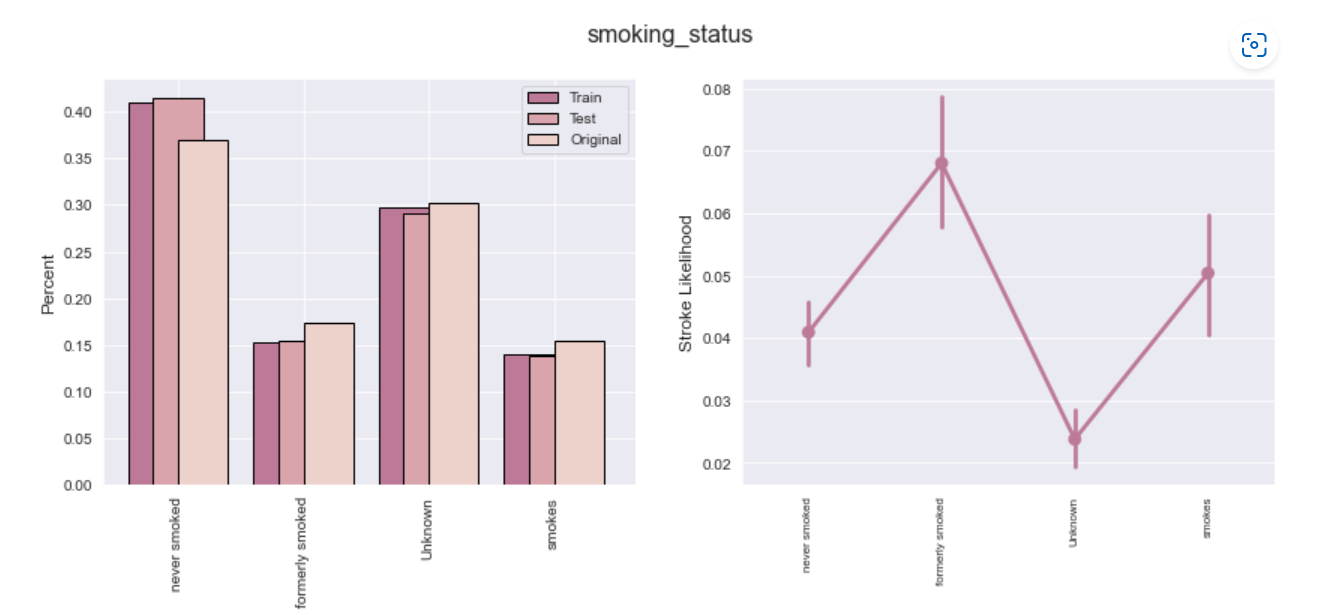
1. I find that fraction of people working in Private is the most in the population, while the fraction who have never worked is the least. People who are self-employed have the highest likelihood of stroke, followed by people working in government jobs and private, while children and those who never worked have zero likelihood of stroke.



1. Fraction of people living in rural and urban areas are more or less same and the likelihood of the rural people experiencing a stroke is more or less same as those living in urban areas. This variable does not give us much information regarding likelihood of stroke. Stroke is more or less equally likely between people living in rural and urban residence.



1. People who formerly smoked have the highest likelihood of stroke, followed by the people who smoke, followed by people who never smoked and the least likelihood is those of the unknown category.



1. The median age of people who have experienced strokes is higher than the median age of those who have not.
2. There is very insignificant difference between the median average glucose level of the people who have experienced stroke and those who have not. So, the variable **avg\_glucose\_level** does not significantly explain who will have stroke and who will not.
3. There is again very insignificant difference between the median values of **bmi** between people who had stroke and who did not. So, again, the variable **bmi** does not significantly explain who will have stroke and who will not.
4. Next I have performed a bit of Feature Engineering:
5. I have replaced the **Other** category in the **gender** column by the **Female** category, since the dataset had no observations in the **Other** category.
6. Since there was no visible explanatory relationship between **bmi** and **stroke**, I have created two new columns named **morbidity** (which equals 1 if **bmi** greater than 40, and 0 otherwise), and **obesity** (which equals 1 if **bmi** greater than 30 and 0 otherwise).
7. Again, since **avg\_glucose\_level**  and **bmi** show no direct effect on **stroke** but from a medicinal perspective they must be important, I created a new column named **risk\_factors** that uses a relationship between **avg\_glucose\_level**, **bmi**, **age**, **hypertension**, **heart\_disease** and **smoking\_status** to predict stroke likelihood. The relationship is as follows:

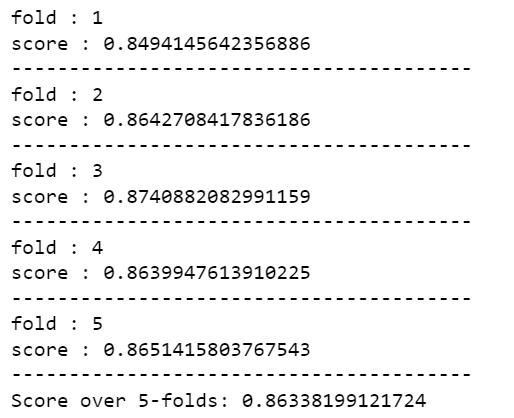
* 1 is added to **risk\_factors** if the **avg\_glucose\_level** is greater than 99
* 1 is added to **risk\_factors** if the **age**  is greater than 45
* 1 is added to **risk\_factors** if the **bmi** is greater than 24.99
* 1 is added to **risk\_factors** if respondent has **hypertension**
* 1 is added to **risk\_factors** if the respondent has **heart\_disease**
* 1 is added to **risk\_factors** if the respondent formerly smoked or smokes.

1. Next I have prepared the data to be fit into the models for training. I have followed the following steps:
2. I have encoded the categorical variables as follows:

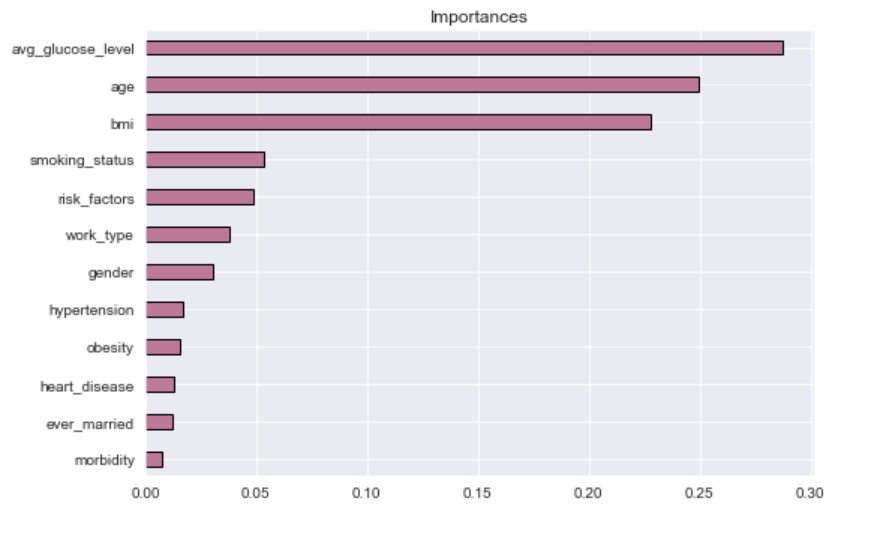
* **gender**: 0 if Female and 1 if Male.
* **ever\_married**: 0 if No and 1 if Yes.
* **work\_type**: 0 if never worked, 1 if children, 2 if Private, 3 if Government job, and 4 if Self-employed.
* **smoking\_status:** 0 if Unknown, 1 if never smoked, 2 if smokes and 3 if formerly smoked.

1. I have got rid of the **Residence\_type** variable based on the EDA result that this variable has no influence on the likelihood of occurrence of stroke.
2. Having prepared the data, I have gone for four models for this classification task, a Random Forest Classifier, an XGBoost Classifier, an LGBM Classifier and lastly a combination of the XGBoost and LGBM Classifiers. I have followed a Stratified 5-fold Cross Validation process while training each of the models. For each model, **roc\_auc\_score** has been used as the metric:
3. **Random Forest Classifier:**

I have used a Random Forest Classifier with 1000 as the number of estimators. The average **roc\_auc\_score** on the 5-fold cross-validation model came out as 0.8634.



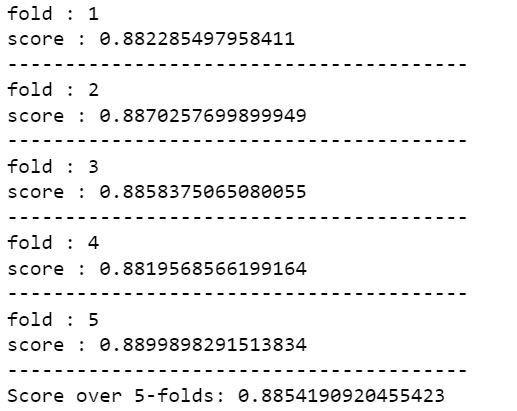
The importance of the variables came as:



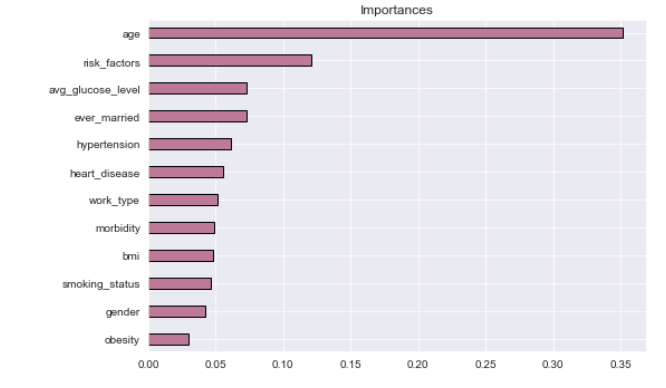
In the above figure, we can see that the variables with most importance are **avg\_glucose\_level**, **age** and **bmi**. However, in the EDA, we found that **avg\_glucose\_level** and **bmi** had little influence on stroke likelihood.

1. **XGBoost Classifier:**

Next I have used an XGBoost Classifier model along with stratified 5-fold cross-validation. The model parameters I used are *max\_depth* of 7, a *learning\_rate* of 0.01, *n\_estimators* as 1000, *subsample* as 0.7, *colsample\_bytree* as 0.7, *reg\_alpha* as 3 and *reg\_lambda* as 3. The average 5-fold CV **roc\_auc\_score** came out to be 0.8854.

****

The importance of the variables came out as:

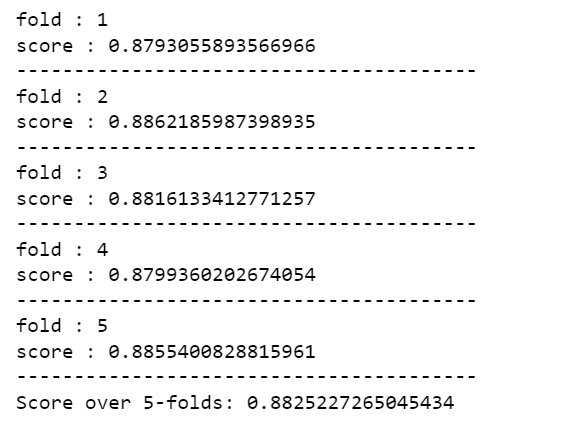
****

This shows that the most important feature is **age**. This is exactly what the EDA had conveyed. The next important feature is **risk\_factors**, which is the engineered feature including a various number of features which most probably affects **stroke**.

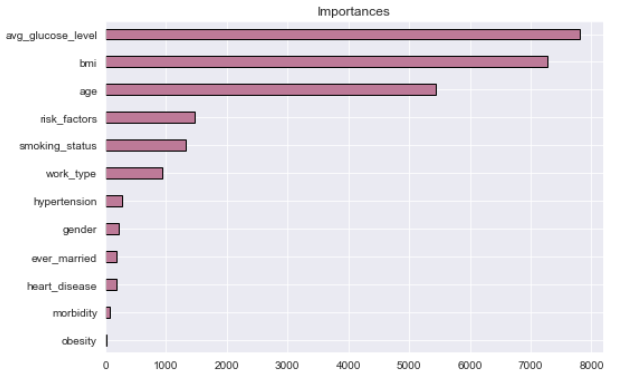
This model performs really well.

1. **LightGBM Classifier:**

Next I try a LightGBM Classifier algorithm with stratified 5-fold cross validation. The model is trained with the parameters: *max\_depth* of 7, a *learning\_rate* of 0.01, *n\_estimators* as 1000, *subsample* as 0.7, *colsample\_bytree* as 0.7, *reg\_alpha* as 3 and *reg\_lambda* as 3. The average 5-fold CV **roc\_auc\_score** came out to be 0.8825.



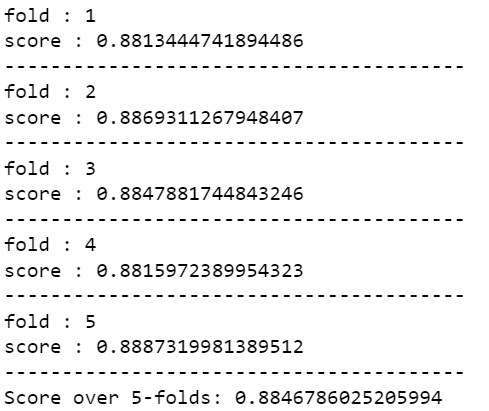
The importance of variables come out as:



This model also gives a good score yet the variables of importance are quite off what had been observed in the EDA.

1. **Combination of XGBoost and LightGBM Classifier:**

Lastly, I try a combination of the XGBoost Classifier and the LightGBM Classifier along with the stratified 5-fold cross-validation. While making predictions, I have given XGBoost a higher weightage of 0.6 and a lower weightage of 0.4 to LGBM Classifier since the XGBoost classifier performs better and the importance of the variables conform to the EDA exposition. The hyperparameters remain the same as in the previous models. The average **roc\_auc\_score** over the 5-folds come out to be 0.8847, which is better than the Random Forest model as well as the LightGBM model, but worse than the XGBoost model.



1. Since the XGBoost model gave the best results, I chose it for the test set predictions. I had already prepared the test set while preparing the training set and so I made the prediction using the XGBoost model with the same hyperparameters as used in the model for training. I have saved the resultant predictions in a csv format.