Analysis of Stroke Prediction Dataset

Data Set Description:

- Stroke Prediction dataset is obtained from Kaggle
  URL: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
- Total 5110 observations in the dataset.
- Each row in the data provides relevant information about the patient.

Goal:

To predict whether a patient is likely to get stroke based on the features such as gender, age, various diseases, and smoking status.

Link to Source Code

Features:

No	Clinical Features	Feature Values and Description	Feature Type
1.	**id**	Unique identifier	Numeric
2.	**gender**	Male, Female, Other	Categorical
3.	**age**	age of the patient	Numeric
4.	**hypertension**	0: if the patient doesn't have hypertension 1: if the patient has hypertension	Binary
5.	**heart_disease**	0: if the patient doesn't have any heart diseases 1: if the patient has a heart disease	Binary
6.	**ever_married**	No, Yes	Categorical
7.	**work_type**	children, Govt_job, Never_worked, Private, Self-employed	Categorical
8.	**Residence_type**	Rural, Urban	Categorical
9.	**avg_glucose_level**	Average glucose level in blood	Continuous
10.	**bmi**	body mass index	Continuous
11.	**smoking_status**	formerly smoked, never smoked, smokes, Unknown	Categorical
12.	**stroke**	1: if the patient had a stroke 0: if the patient did not have a stroke	Class Label

***There are 6 numeric features, 5 categorical features and one class label**

Data Preprocessing:

- Rows with missing values are removed from the dataset.
- Also, *gender = other* is removed since there is only 1 row in the dataset lacking enough representation in dataset.
- Categorical features (gender and smoking status) are converted into numerical features (dummy variables).
- **4908** observations in total after data cleaning.

Solving Imbalanced Dataset Problem

Problem:

- Stroke prediction dataset is highly imbalanced.
  - **209** observations with stroke = 1
  - 4699 observations with stroke = 0
- Class stroke =1 does not have enough representation as there are only 6% observations of that class in the dataset. Thus, the models created using different classifiers were underfitting the dataset with 0.00% accuracy of prediction for the class stroke =1 with TPR = 0.00 and TNR = 1.00

Solution:

- A balanced sample dataset is created by:
  - Combining all 209 observations (stroke = 1)
  - 10% of the observations (stroke = 0) were obtained by random sampling from the 4699 (stroke = 0) observations.
- The resulting sample dataset is then split into train and test set (70/30 split) respectively.
- Data is scaled and different classifiers are trained on the train set and applied on the modified test-set (whole data set excluding train set).

Classifiers used on the dataset:

Summary of performance measures:

Model	TP	FP	TN	FN	TPR	TNR	Accuracy%
K_NN	35	722	3657	19	0.6	0.8	83.3
Logistic Regression	41	764	3615	13	0.8	0.8	82.5
Decision Tree	41	1074	3305	13	0.8	0.8	75.5
Random Forest	30	518	3861	24	0.6	0.9	87.8

Although **K-NN**, **Logistic Regression** and **Random Forest** classifiers provide the highest overall accuracy of prediction **82.5%** and **85.7%** respectively, however, the **Logistic Regression** gives the highest **TPR** = **0.8** and **TNR = 0.8**. In this case TPR represents the accuracy of predicting patients who have had a stroke and TNR represents the accuracy of predicting patients who did not have a stroke. TPR is very important considering we have a highly imbalanced dataset with around 6% of the observations with patients who have had a stroke. That’s why performance of **Logistic Regression** classifier is the best among the four classifiers in the prediction of stroke.

Accuracy of prediction when *age* feature is removed from the dataset

The *age* feature, when removed, contributed the most to loss of *TPR*. Hence, age feature plays a significant role in the stroke prediction.

Summary of performance measures:

Model	**TP**	**FP**	**TN**	**FN**	**TPR**	**TNR**	**Accuracy%**
K_NN	27	613	3766	27	0.5	0.9	85.6
Logistic Regression	20	447	3932	34	0.4	0.9	89.1
Decision Tree	20	1275	3104	34	0.4	0.7	70.5
Random Forest	18	484	3895	36	0.3	0.9	88.3

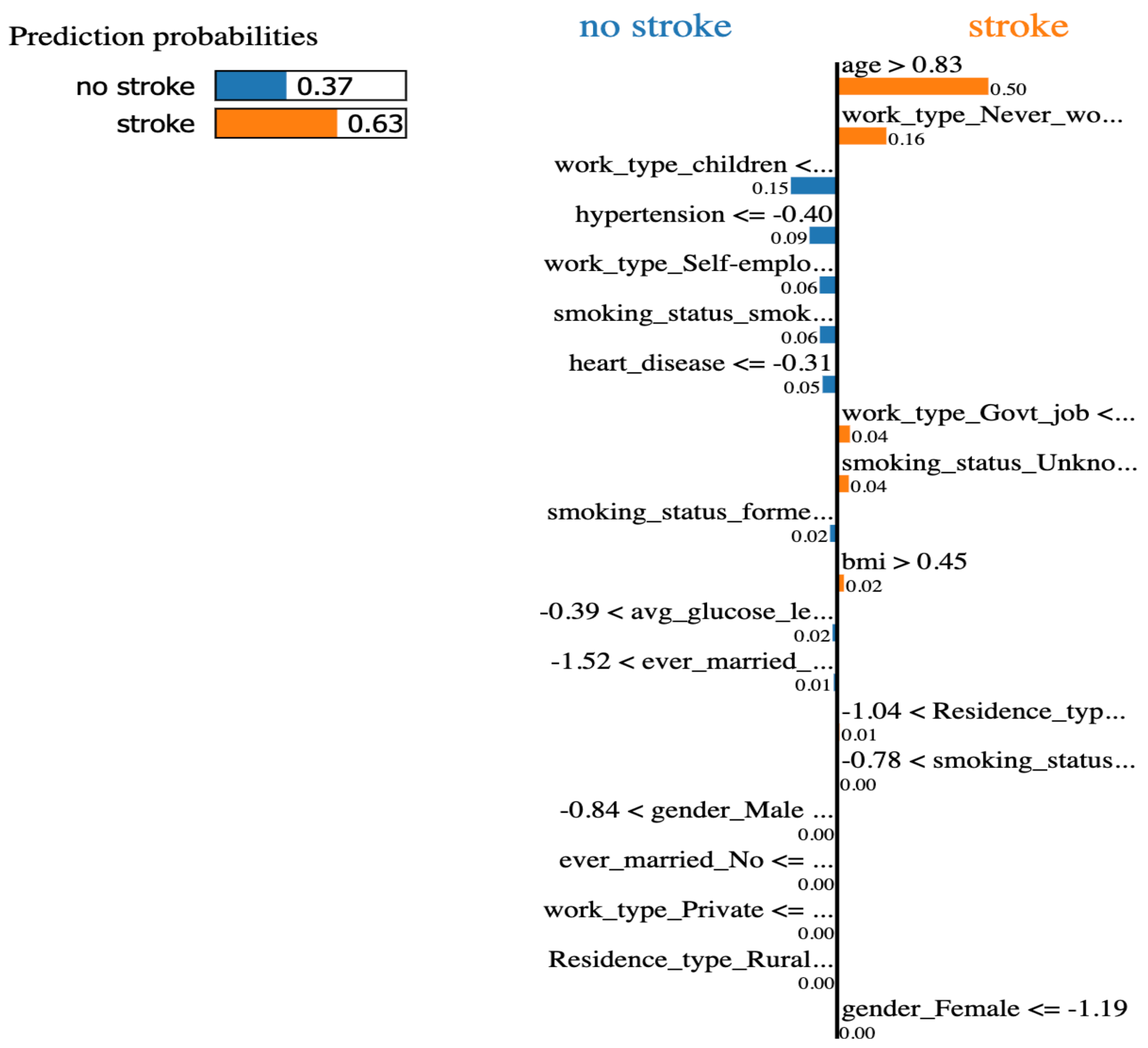
Significant drop in TPR from 0.8 to 0.4 (logistic regression)

Explanation of prediction using LIME (Locally Interpretable Model-Agnostic Explanations)

- LIME algorithm is used to explain the predictions of Logistic Regression classifier. Lime explains the prediction by approximating it locally with an interpretable model.
- The output of LIME provides an intuition into the inner workings of machine learning algorithms as to what features are being used to arrive at a prediction. It assigns weight to each feature based on its contribution to the prediction of a label within the local structure of the data.

Local explanation for class *stroke* = *Yes*

Output is shown for patient id = 11 in test set

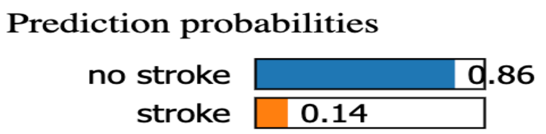


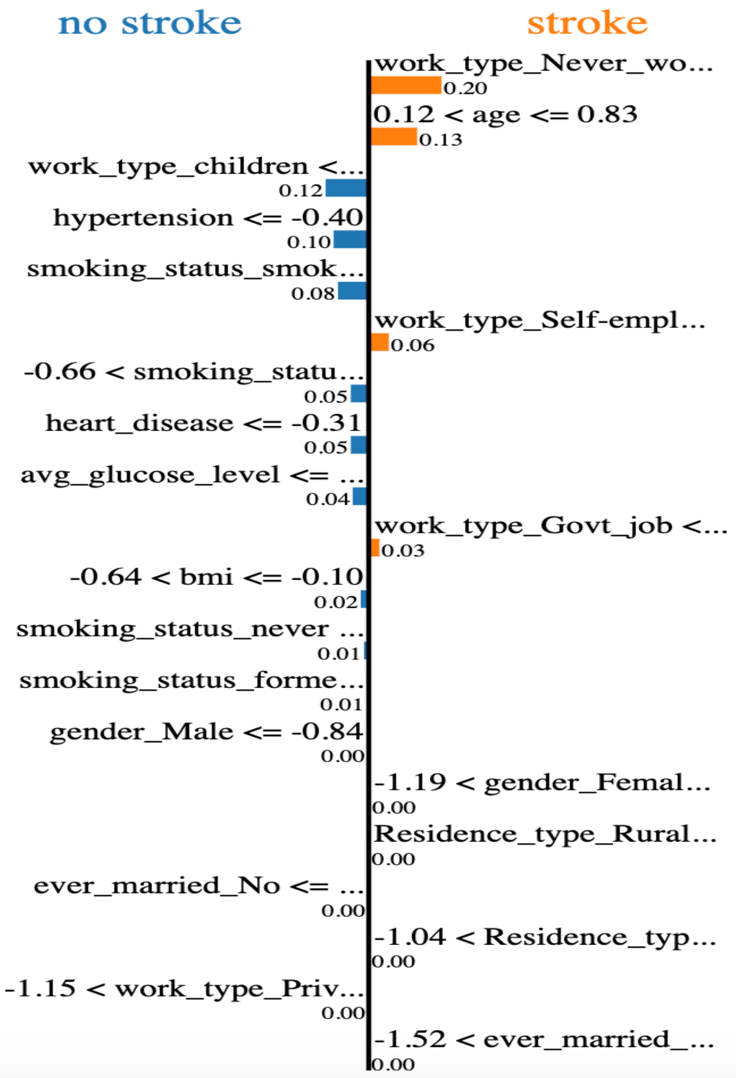
Feature	Value
age	1.38
work_type_Never_worked	-0.05
work_type_children	-0.37
hypertension	-0.40
work_type_Self-employed	2.10
smoking_status_smokes	-0.40
heart_disease	-0.31
work_type_Govt_job	-0.38
smoking_status_Unknown	-0.66
smoking_status_formerly smoked	-0.47
bmi	0.46
avg_glucose_level	-0.26
ever_married_Yes	0.66
Residence_type_Urban	0.96

- The model is 63% confident that this patient had a stroke.
- The top predictors are of *age, work_type_Never_worked, work_type_Govt_job, smoking_status_Unknown,* and *bmi*.
- The values of features *work_type_children, hypertension, work_type_Self-employed, smoking_status_smokes, heart_disease, smoking_status_ formerly smoked* decrease patient's chance to be classified as *stroke*.

Local explanation for class *stroke = No*

Output is shown for patient id = 100 in test set





Feature	Value
work_type_Never_worked	-0.05
age	0.28
work_type_children	-0.37
hypertension	-0.40
smoking_status_smokes	-0.40
work_type_Self-employed	-0.48
smoking_status_Unknown	1.52
heart_disease	-0.31
avg_glucose_level	-1.11
work_type_Govt_job	-0.38
bmi	-0.59
smoking_status_never smoked	-0.78
smoking_status_formerly smoked	-0.47
gender_Male	-0.84

- The model is 86% confident that this patient did not have a stroke.
- The top predictors are *work_type_children, hypertension,* and *smoking_status_smokes.*
- The values of *features work_type_Never_worked, age, work_type_Self-employed*, and *work_type_Govt_job* decrease patient's chance to be classified as *no stroke.*

Conclusion:

- Logistic Regression Classifier predicts with overall accuracy of 82.5%
- TPR for class stroke = 1 is 0.8
- TNR for class stroke = 0 is 0.8
- The *age* feature, when removed, contributed the most to loss of TPR. Hence, age feature plays a significant role in the stroke prediction.
- The output of LIME provides an intuition into the inner workings of machine learning algorithms as to what features are being used to arrive at a prediction. It assigns weight to each feature based on its contribution to the prediction of a label within the local structure of the data.