BAHCESEHIR UNIVERSITY

Data Exploration and Stroke Prediction Using Logistic Regression on Patient Data

Authors:

Aleyna Benan Aydı

2003977

Contribution:

To the preparation of the report and the presentation

Burak Can Soyak

2100841

Contribution:

The Python Notebooks for EDA and Stroke prediction model, typesetting and first page of the report, Slides about the model.

Tarkan Özşen

2102276

Contribution:

EDA lead, pathfinding for data processing, EDA section of the report, reporting of references.

Content of the paper:

- What is Stroke
- What is our Objective
- What is the Data
- Data Analysis
- Why Logistic Regression and model parameters
- Data Splitting
- Regression and Scoring
- Results
- Conclusion
- Related Works

Stroke

A **stroke** is a condition that happens when the blood supply to a part of the brain is disrupted. When the brain doesn't receive enough blood and oxygen, it can lead to damage or death of brain cells.

Immediate medical attention is crucial for a stroke because early treatment can help minimize brain damage and improve the chances of recovery.

The Objective

- To explore the use of patient data and linear regression techniques to predict the probability of stroke occurrence.
- By utilizing readily accessible patient data, including metabolic risks and behavioral factors.
- Use a simplistic model to prove even with the most basic parameters, it is possible to get an OK score on the mostly artificial data.

The Data

METABOLIC RISK

- Hypertension
- BMI (Body Mass Index)
- Glucose Level
- Heart Disease

BEHAVIORAL FACTOR

- Smoking
- Ever Married
- Work Type
- Residence Type

1

Exploratory Data Analysis (EDA)

2

Logistic Regression

1

Exploratory Data Analysis (EDA)

Exploratory Data Analysis

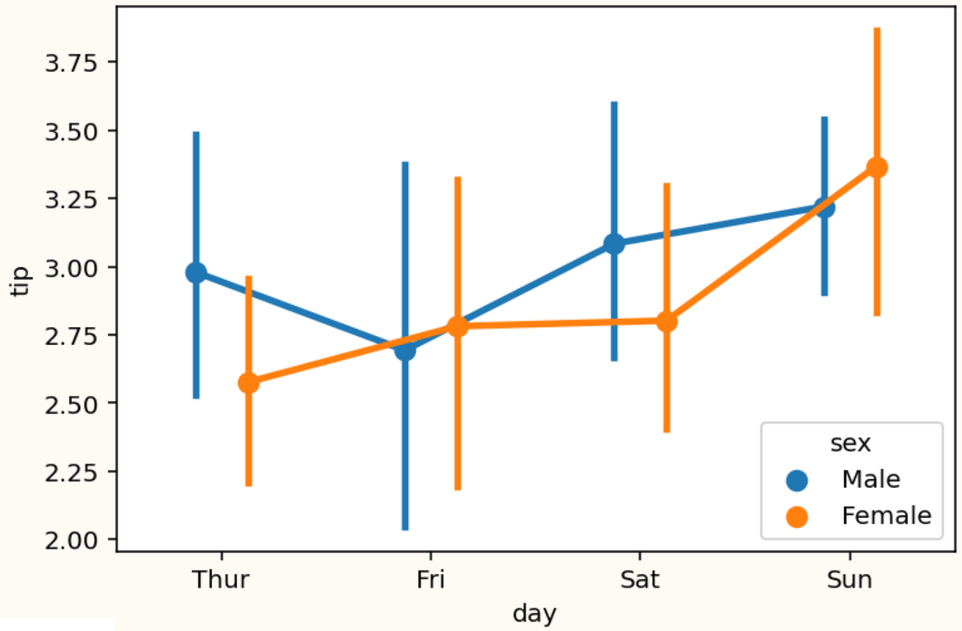
- 5111 Original Patient Data (**25.0%** of Combined Data)
- 15305 Train Data (AI generated)
- Visualization Techniques:

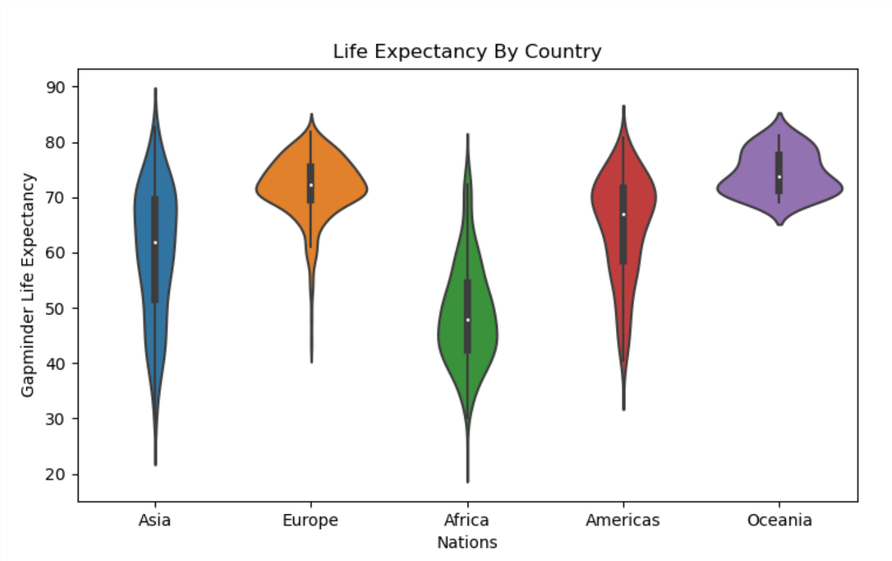
1. Pie Charts
2. Bar Plots
3. **Point Plots**
4. Kernel Density Estimate (KDE) Plots
5. **Violin Plots**
6. **Spearman Correlation Matrix**

Point Plot:

- Visuals for **discrete** values
- **Boolean** independent variables vs **probabilities**
- Confidence Intervals

SAMPLE GRAPH





SAMPLE GRAPH

Violin Plot:

- Visuals for **continuous** values
- **Boolean** dependent variables vs **weight distribution**
- Expected distribution
- Quartile Information

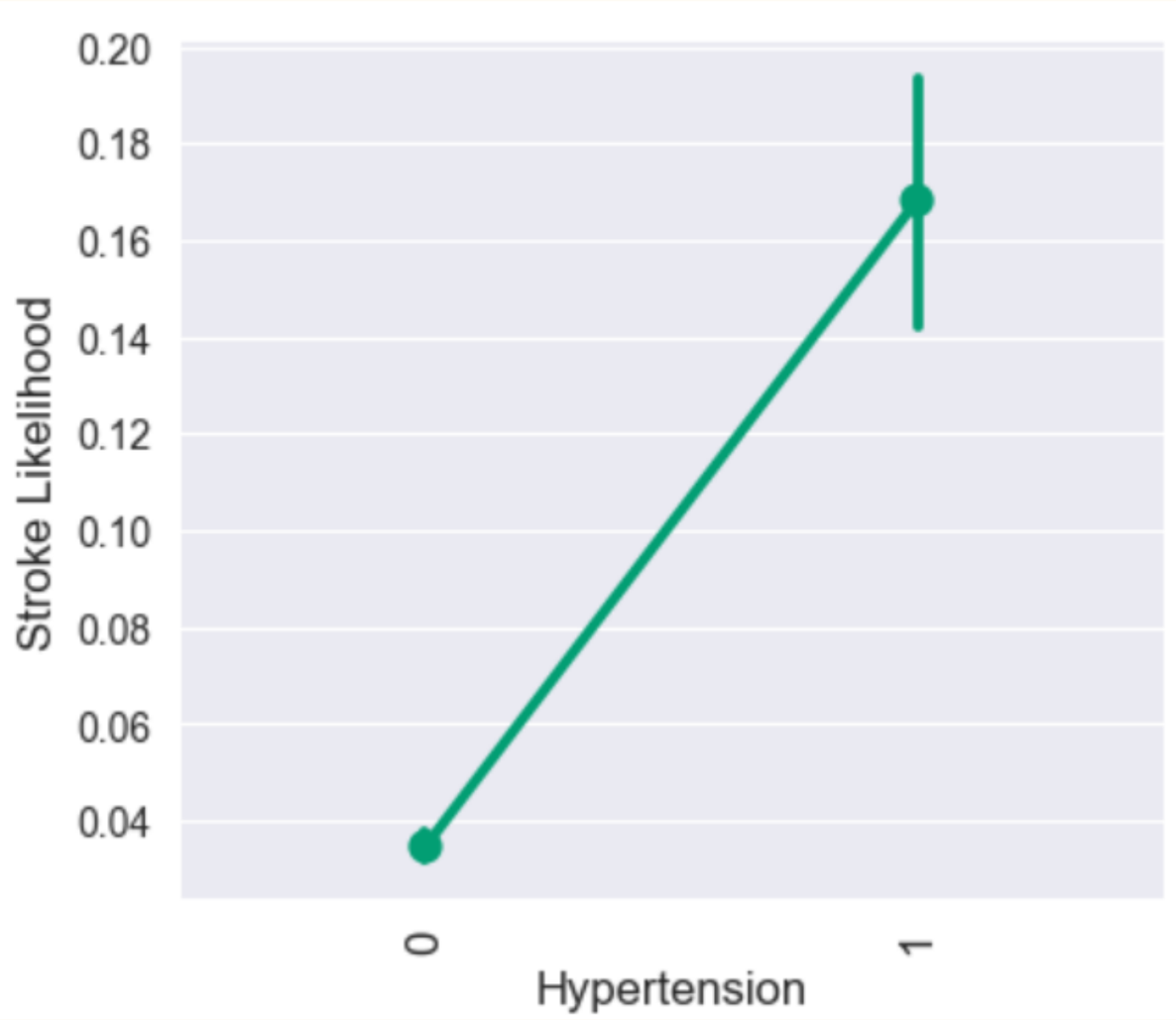
Exploratory Data Analysis

Point Plot: Hypertension vs Stroke Likelihood

- **~7%** of patients have hypertension

- **~3.5%** stroke likelihood for patients **without hypertension**
- Rate increases to **~17%** for patients who **have hypertension** (**4.85** times as likely)
- Similar results in the **Heart Disease** vs **Stroke Likelihood** Point Plot

Hypertension vs Stroke Likelihood



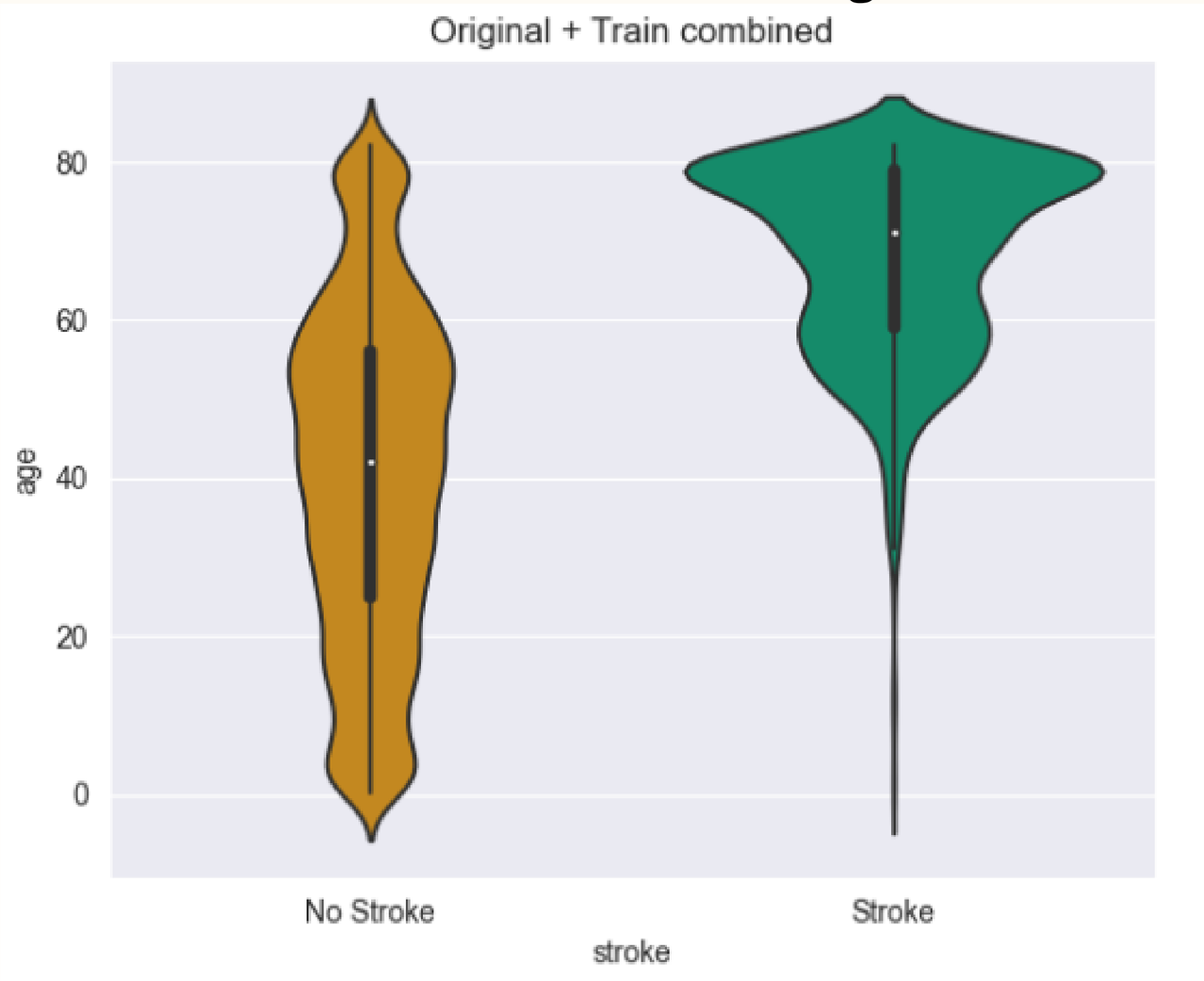
Indication of high correlation between variables

Exploratory Data Analysis

Violin Plot: Stroke Status vs Age

- Median for No Stroke vs Stroke: **41 vs 71** (combined data)
- Lower and Upper Quartiles: **24 - 58 vs 59 - 80**
- Lowest value on Stroke Violin Plot: **Age 30**

Stroke Status vs Age



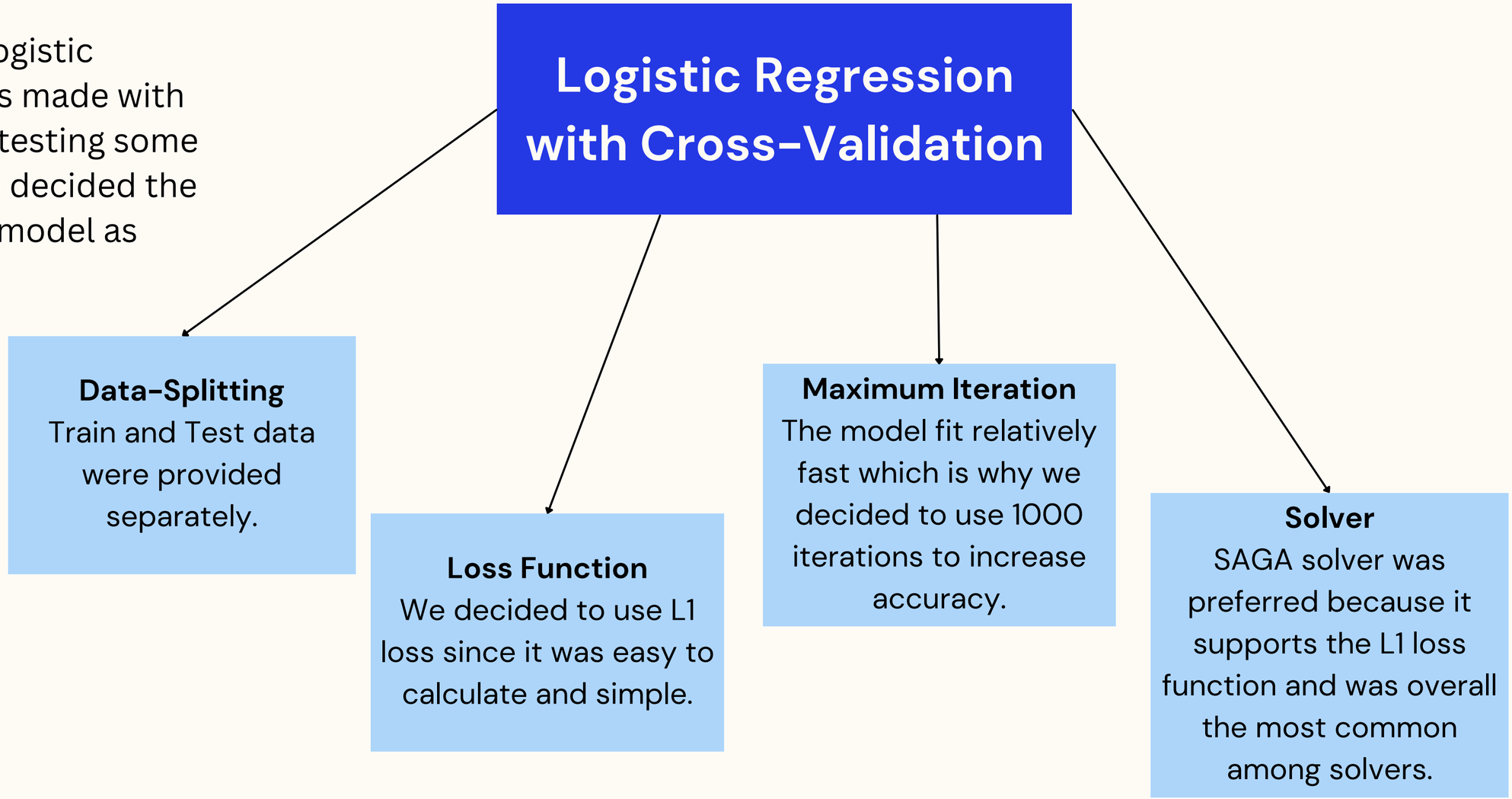
Increasing expectation of stroke for 50+ year old patients

2

Logistic Regression

Why Logistic Regression?

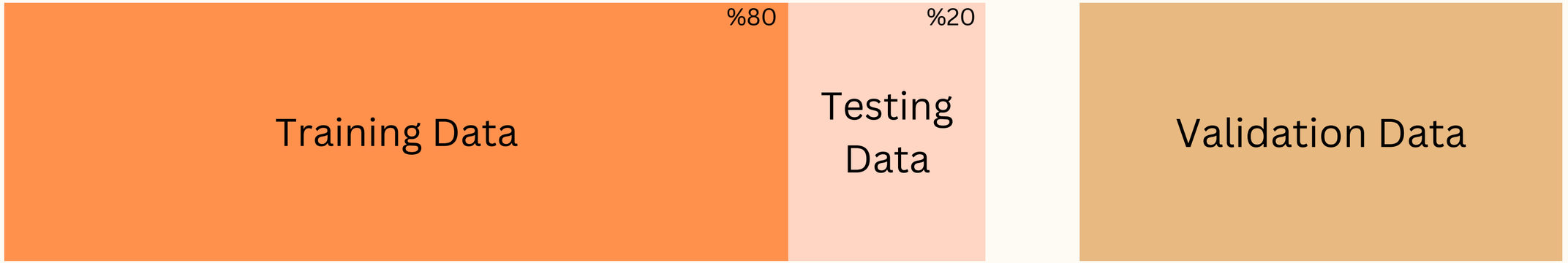
- The decision of the logistic regression model was made with trial and error. After testing some other parameters we decided the configuration of our model as shown.



Data Seperation

- **The provided data for the competition**

- **The data we prepared for usage**



- Yes Proportions are exact here!

Regression and Scoring

The Model iterated over the data 5 times.

- After each iteration score for that iteration was printed and saved.

Scoring with AUROC

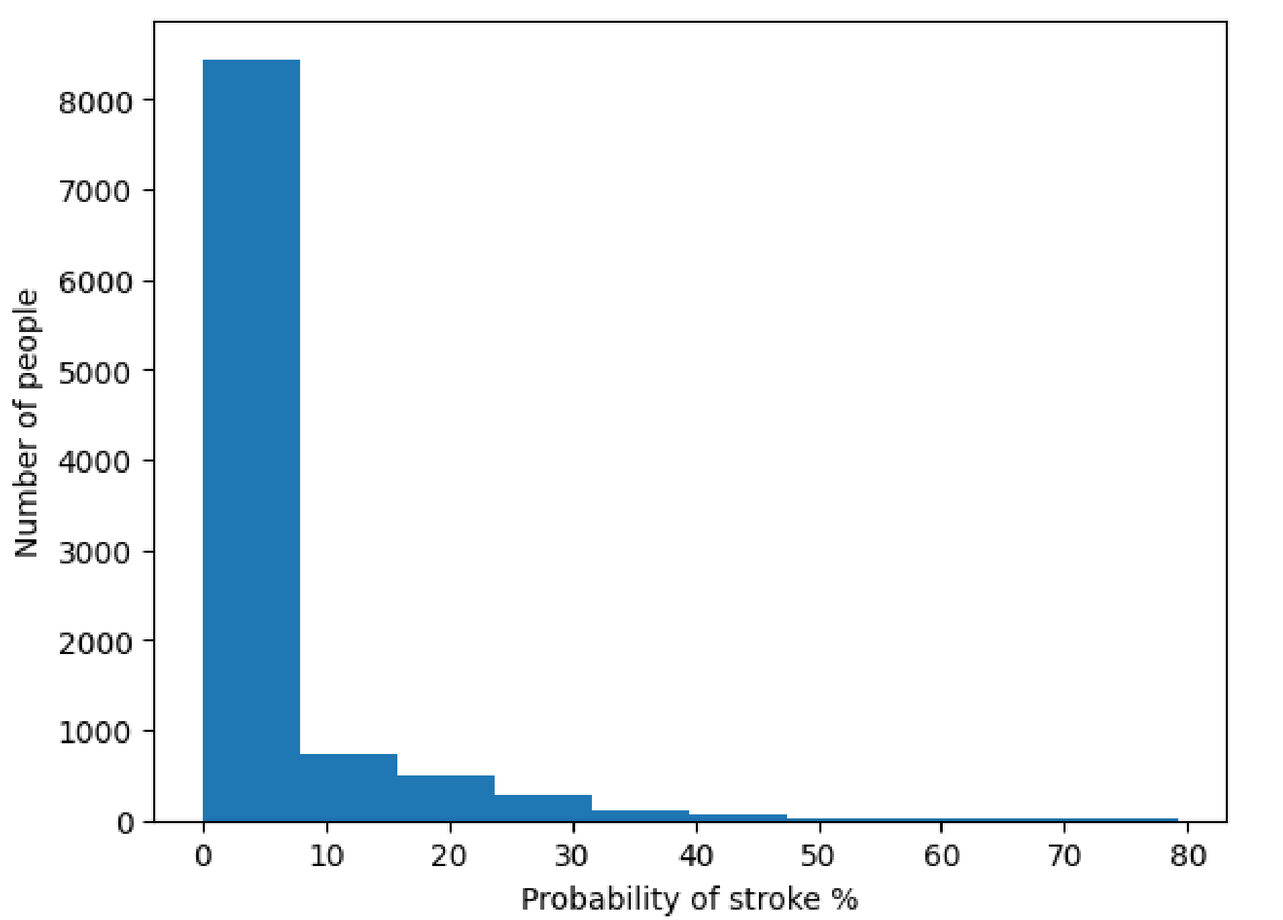
- For scoring AUROC (Area under receiver operating characteristic curve) was used.

We tested the Model on the test data

- The Model was tested over the test data and produced a remarkable score.

External Validation

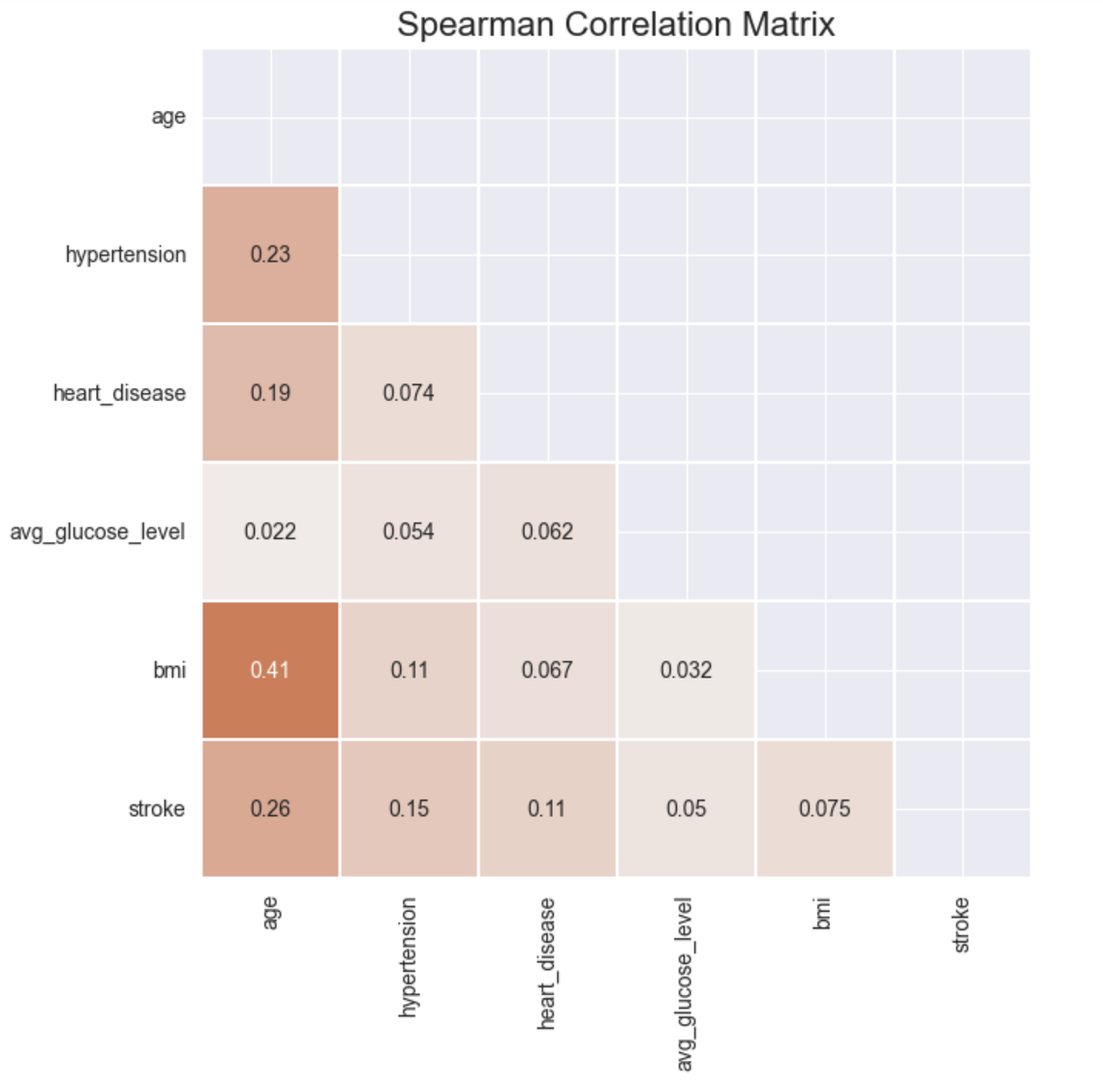
- Validation was done on the validation data prepared. Scoring was done online.



Results

Scores over the data

- Over the prepared test data, the model produced a score of 0.88.
- Predictions on the validation data were submitted to Kaggle for scoring and it scored a score of 0.89.



Conclusion

- Even with correlations of moderate and lesser significance, with the usage of logistic regression, the model performed a relatively OK score of 0.89.
- This concludes that even with small correlations, models are able to learn and predict the event of having a stroke in a notable performance.
- Or does it?

We Have Questions:

- Is this a real model?
- Are we interpreting the plots correctly?
- Are there any findings about the model we are hiding?
- Why are we asking you to do this?

Or the QR code provided here:

Discuss at:

- We encourage you to review the paper and the code using the **link here**:
- Please ask your questions and share your findings with others by creating issues!

Related Works

- Emon, Minhaz Uddin, et al. "Performance analysis of machine learning approaches in stroke prediction." 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2020.

- Liu, Yunfan, Baoying Ma, and Yan Wang. "Study on prediction model of stroke risk based on decision tree and regression model." 2021 IEEE International Conference on Big Data (Big Data). IEEE, 2021.