# STROKE RISK PREDICTION FOR HEALTHCARE

Jeet Faldu

33670021

# PROJECT AIM

- Stroke is a huge concern for society as it can lead to human loss
- Stroke can cause long-term impairment
- Stroke is fourth leading cause of death worldwide

# Project Description

## Objective

- Predict people's chance of having a stroke based on their lifestyle and demographics

## Methodology

- Predictive model based on patient information
- Predict Stroke for new patients
- Include models in existing applications in different domains

## Data Roles

- Data Scientist
- Business Analyst
- IT Support

# Business impact

## Insurance Providers

- This model can identify the new applicant and take decisions based on this model that particular customer will be given a high coverage plan or not

## Public Application

- People can get notified with integration of this model if they fall under the risk of stroke due to changes in their lifestyle or any features

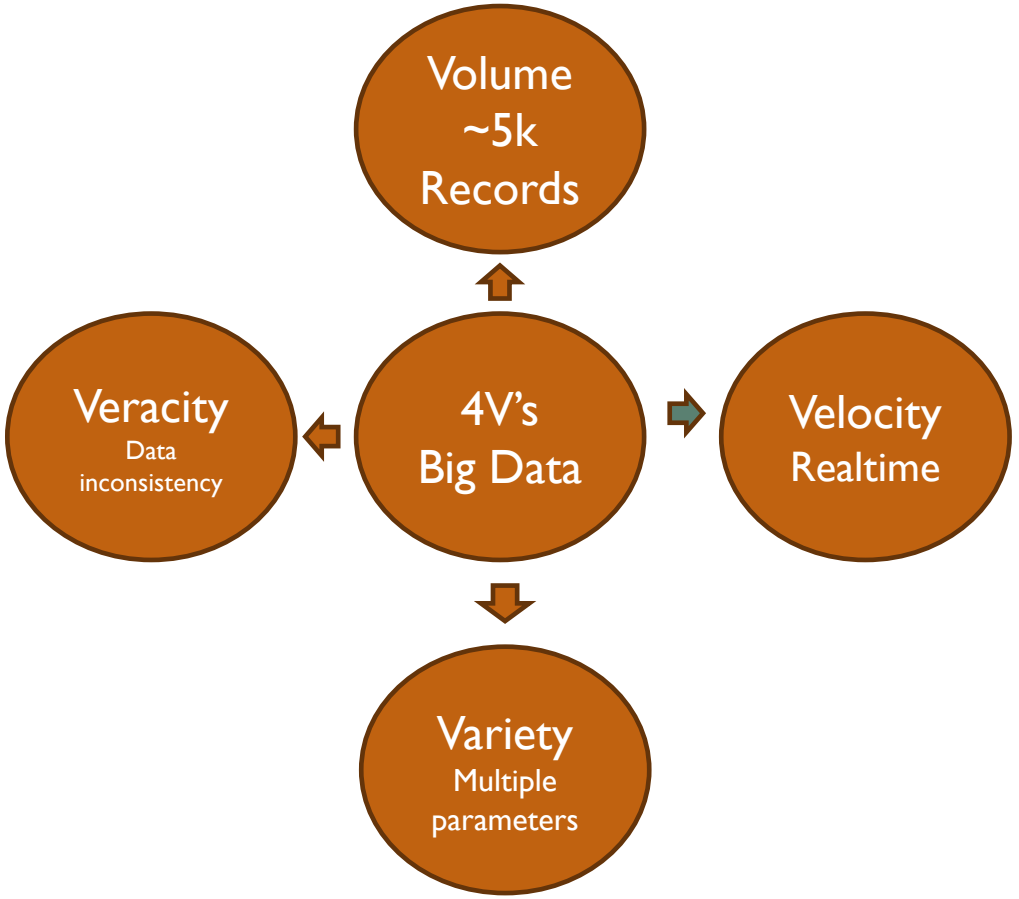
## Health Coaching Institutions

- Identify the target customers which are at risk of developing characteristics which can lead to Stroke in feature

# Data Characteristics

## Data Sources

- EHR
- Medical Image
- Wearable Devices
- Patient Surveys




| Storage | |
|---|---|
| Hadoop distributed File System | Cloud Based Solutions |

| Processing | |
|---|---|
| Hadoop Ecosystem | Apache Spark |

| Analysis | |
|---|---|
| Apache HBase | Python, R |

1

## Proposed data analysis

- Exploratory data analysis
- Predictive Modeling

2

## Exploratory data analysis

- Data checking and wrangling
- Data Exploration

3

## Predictive Modeling

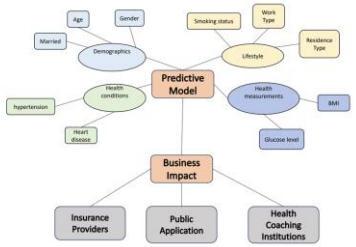
- Decision Tree
- Logistic Regression

# Demonstration

## Exploratory Data Analysis

### Data Checking and Wrangling

- Description and Data Type of variables




- Dropping missing value rows

- Handling Data Anomalies

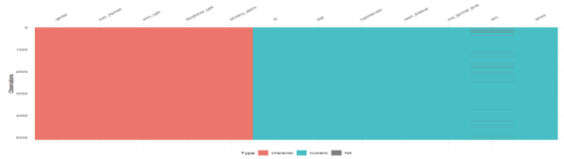


Figure 2 Null value check

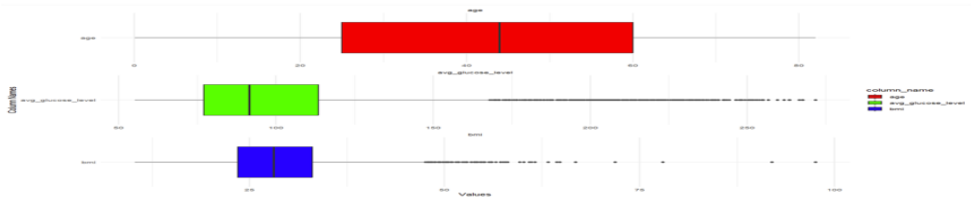


Figure 3 Box plot for numerical variables

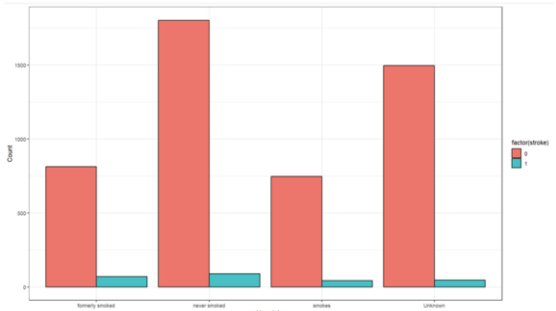


Figure 5 Smoking Status

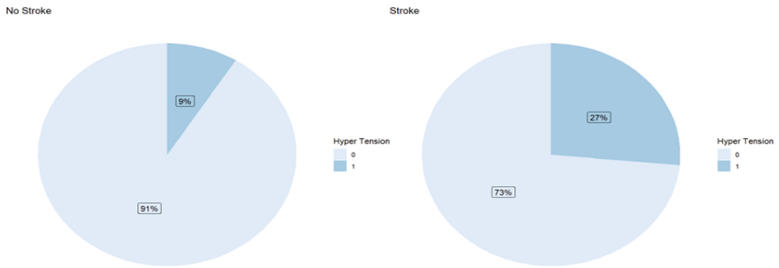



Figure 9 Hypertension

- Visualization of Numeric columns
- Distribution of Categorical Columns
- Pie charts for Boolean variables

# Demonstration (Continued)

## Feature Selection

- Below features are selected for Logistic Regression
  - Age
  - Hypertension
  - Heart disease
  - Avg glucose level

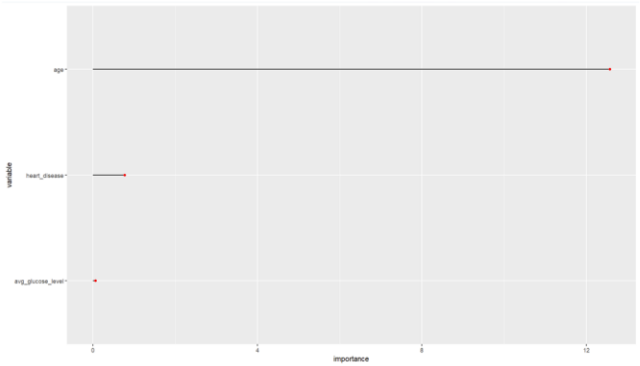



Figure 11 Importance matrix

## Predictive Modeling

### Decision Tree

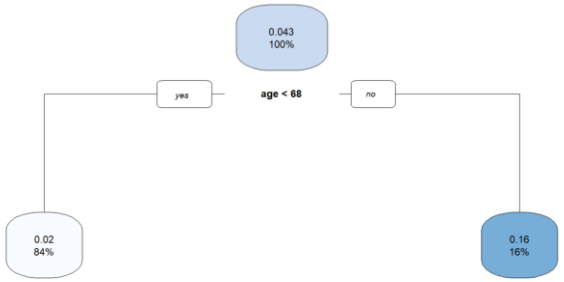
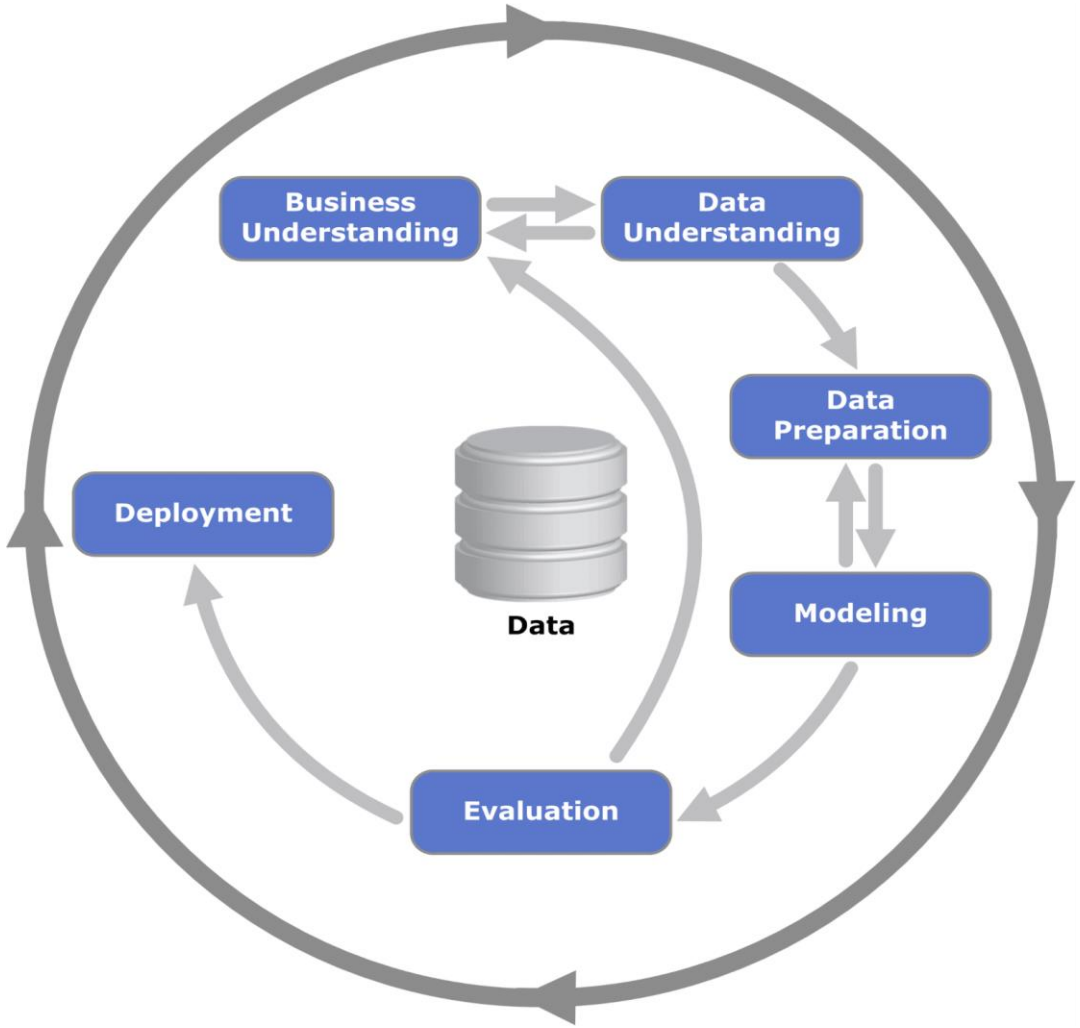



Figure 12 Decision Tree

### Logistic Regression

```
Call:
glm(formula = stroke ~ gender + age + hypertension + heart_disease +
    ever_married + work_type + Residence_type + avg_glucose_level +
    bmi + smoking_status, family = "binomial", data = df_hds3)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -7.359650   1.067322  -6.895 5.37e-12 ***
gender2          -0.014625   0.154382  -0.095 0.924525
age               0.073481   0.006347  11.578  < 2e-16 ***
hypertension      0.524857   0.175023   2.999 0.002711 **
heart_disease     0.348763   0.207231   1.683 0.092381 .
ever_married2    -0.115175   0.247289  -0.466 0.641394
work_type2       -0.681655   1.114151  -0.612 0.540660
work_type3       -9.823495 308.741641  -0.032 0.974617
work_type4       -0.520849   1.100279  -0.473 0.635943
work_type5       -0.945890   1.118910  -0.845 0.397906
Residence_type2   0.004514   0.149987   0.030 0.975990
avg_glucose_level 0.004652   0.001294   3.595 0.000324 ***
bmi               0.004062   0.011880   0.342 0.732387
smoking_status2  -0.067224   0.188630  -0.356 0.721556
smoking_status3   0.313918   0.229471   1.368 0.171310
smoking_status4  -0.275333   0.247112  -1.114 0.265193
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1728.3  on 4907  degrees of freedom
Residual deviance: 1363.2  on 4892  degrees of freedom
AIC: 1395.2

Number of Fisher Scoring iterations: 14
```

# Data Science Process Standard




1. Main objective of identifying the patients if they can have stroke has been defined
2. Stroke data has been obtained from secondary sources and exploration has been done on the collected data
3. Data checking, data wrangling and exploratory data analysis have been performed
4. Decision tree and Logistic Regression have been used to predict if patient can have stroke
5. Evaluated predicted model and obtained high accuracy
6. Proposing model for user applications from different domains

# Data Science Process Standard

| | |
|---|---|
| **Accessibility** | • Maintaining an extensive data catalogue that includes data sources, types, and definitions and is easily accessible to important stakeholders is necessary for accessibility |
| **Security** | • Sensitive health information must be protected from unauthorized access, which requires strict security measures including encryption, access control, and regular audits |
| **Confidentiality** | • Due to the sensitive nature of the material, confidentiality is of the utmost importance, and personal data is protected through techniques of masking |
| **Data Retention Policy** | • By routinely evaluating and archiving data, a well-defined data retention policy must be designed to reduce overhead costs and guarantee compliance |
| **Ethical Concerns** | • Transparency, explicability, informed permission, and adherence to ethical and legal requirements, such as data protection legislation, are all required by ethical considerations |


Thank you…