Stroke Prediction in R

2022-12-03

# ————————————————————————

# Team -

Sreevani Basvaraj Patil; Pranjal Pathak; Filip Forejtek; Casey Cooper; Sanchana Sundarraj; Akhilesh P Patil

## CONTENTS

1. **Introduction**

* 1.1. Preface
* 1.2. The Project
* 1.3. What is Stroke Diagnosis ?
* 1.4. Causes of Stroke Diagnosis
* 1.5. Impact Analysis of this study

1. **Analysis**

* 2.1. Packages
* 2.2. Data Collection
* 2.3. Data Preparation
* 2.4. EDA
* 2.5. Logistic Regression

1. **Results**
2. **Conclusion**
3. **Personal Biases**

# ————————————————————————

# 1. Introduction

## **1.1. Preface**

The second biggest cause of death worldwide is stroke. The World Health Organization estimates that 5 million people get a stroke each year in the world. One third of them pass away, while the other third are permanently crippled. Every 40 seconds, someone in the US experiences a stroke, and every four minutes, someone passes away. As a result of the disruption of blood supply to the brain, sufferers might experience a wide range of incapacitating symptoms, such as abrupt paralysis, loss of speech, or blindness. In the US, the annual financial burden on the healthcare system is estimated to be $34 billion. Care for elderly stroke survivors costs an additional $40 per year.

Stroke carries a high risk of death. Survivors can experience loss of vision and/or speech, paralysis and confusion. Stroke is so called because of the way it strikes people down. The risk of further episodes are significantly increased for people having experienced a previous stroke. The risk of death depends on the type of stroke. Transient ischaemic attacks or TIA – where symptoms resolve in less than 24 hours, have the best outcome, followed by stroke caused by carotid stenosis (narrowing of the artery in the neck that supplies blood to the brain). Blockage of an artery is more dangerous, with rupture of a cerebral blood vessel the most dangerous of all.

## **1.2. The Project**

A localized lesion to the central nervous system brought on by decreased blood supply to the brain is what is known as a stroke, an acute neurological malfunction. Stroke is becoming a major global threat that can result in early death and severe economic effects. Therefore, it is essential to proactively predict how various risk factors affecting the likelihood of having a stroke, and machine learning models and AI seems to be the right instrument for the process. Machine learning methods for the analysis of neuro-imaging data are used to help diagnose a **stroke**.

*A stroke can be life-threatening, if you or someone has had a stroke or has the warning signs, it is extremely important to seek medical attention immediately.*

## **1.3. So, what is Stroke Diagnosis ?**

Stroke is one of the topics which will be taught as part of neurological disorders and covers the following information:

1. Types of stroke
2. Risks and Symptoms
3. Diagnosing Conditions
4. Advanced treatments
5. Detection and Prevention

## **1.3. How are strokes Diagnosed ?**

A stroke can be life-threatening, if you or someone has had a stroke or has the warning signs, it is extremely important to seek medical attention immediately. An expert will diagnose a stroke based on signs and symptoms, medical history, and physical examination.

1. Physical exam
2. Blood testing
3. Computerized tomography (CT) scan
4. Magnetic resonance imaging (MRI)
5. Medical history

## **1.4. Predicting Stroke Outcome with Data Analytics and Machine Learning**

In recent years, ML models have been successfully used to evaluate stroke risk variables and outcomes. They include building a deep neural network (DNN) model, using logistic regression and random forest to predict poststroke motor outcomes, as well as assessing a mixed-effect linear model to predict the likelihood of cognitive deterioration poststroke.

Machine learning methods for the analysis of neuroimaging data are used to help diagnose stroke. Since neuroimaging data covers a wide range of aspects relating to anatomy, function, and pathology, it is both highly extensive and complex. The application of machine learning techniques to the processing of complicated brain imaging data has garnered increasing attention in neuroimaging research. The existence of large-scale research that is supplemented with a broad and deep phenotyping is crucial to the success of machine learning methodologies. An unbalanced dataset containing data on thousands of people with known stroke outcomes was used in the research to develop a model that can predict stroke results with high accuracy and validity. A number of algorithms, including decision trees, naive bayes, and random forests, were evaluated; the random forest classifier was shown to be the most accurate at predicting the result of strokes (92%).

## **1.5. Why this study is required and how it can help ?**

The main-usecase for this study is to examine various contributing factors that lead to a potential life-threatning stroke attack.

# ————————————————————————

# Libraries

Before proceeding to our project, kindly download the listed packages as below:

Once you have installed the packages, you’re good to go:

# Kindly import these pacakges to run our project:  
  
library(httr) # for downloading data  
library(tidyverse) # generic functions and deplyr  
library(tidymodels) # models  
library(skimr) # descriptive stats  
library(stringr) # works with strings  
library(themis) # for SMOTE and other recipes for target balancing  
library(vip) # for variable importance  
library(probably) # for performance calibration  
library(glmnet)

# Let's set directory paths here...  
  
# Parent dir where our source code is located on your system  
dir\_root <- getwd()  
  
# Ideally this would be an empty dir location at first  
dir\_data <- paste0(dir\_root, "/data")  
dir.create(dir\_data)

## Warning in dir.create(dir\_data): 'C:\Users\Akhilesh\Documents\Data Science as a  
## Field\Project\_Report\source\source\data' already exists

# Output location  
dir\_output <- paste0(dir\_root, "/output")  
dir.create(dir\_output)

## Warning in dir.create(dir\_output): 'C:\Users\Akhilesh\Documents\Data Science as  
## a Field\Project\_Report\source\source\output' already exists

# sanity display  
cat("\* Your working dir is ---> \n", dir\_root)

## \* Your working dir is --->   
## C:/Users/Akhilesh/Documents/Data Science as a Field/Project\_Report/source/source

cat("\n\n\* Data is located in ---> \n", dir\_data)

##   
##   
## \* Data is located in --->   
## C:/Users/Akhilesh/Documents/Data Science as a Field/Project\_Report/source/source/data

cat("\n\n\* All the outputs will be generated in ---> \n", dir\_output)

##   
##   
## \* All the outputs will be generated in --->   
## C:/Users/Akhilesh/Documents/Data Science as a Field/Project\_Report/source/source/output

# ————————————————————————

# Data Collection

The source of the dataset in scope is Kaggle. The URL for the dataset is <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

To download this dataset, user can use the mirror link hosted on **GitHub.com** (this step does not require installation of additional packages, but this could reflect an older snapshot version of the data since the source dataset was hosted on Kaggle and was being updated on a frequent basis).

This directly downloads and loads data from a mirror link which may not reflect the latest snapshot of the source dataset uploaded on Kaggle.

# Utility Helper  
download\_from\_git <- function(url, download\_path) {  
 # Download file from url  
 download.file(url, download\_path, mode="wb")  
 # Read csv  
 # upon quick look at the data set  
 # if you set smoking\_status to factor in col\_types, na() won't work  
 # remove ID, Sex == Other  
 # output to a factor  
 df <-   
 readr::read\_csv(download\_path, col\_types = "cfdfffffddcf", na = c("Unknown", "N/A")) %>%   
 mutate(smoking\_status = factor(smoking\_status),  
 stroke = factor(ifelse(stroke == 1, "yes", "no"),   
 levels = c("yes", "no"))) %>%   
 select(-id)  
 return(df)  
}

# Mirror URL  
dataset\_url <- 'https://raw.githubusercontent.com/KOcasey/Stroke-Diagnosis-Analysis/main/Data/StrokeDiagnosis.csv'  
  
# download path and name  
download\_filename <- "healthcare-dataset-stroke-data.csv"  
download\_path <- paste0(dir\_data, "/", download\_filename)  
  
df <- download\_from\_git(dataset\_url, download\_path)  
head(df)

## # A tibble: 6 × 11  
## gender age hypertens…¹ heart…² ever\_…³ work\_…⁴ Resid…⁵ avg\_g…⁶ bmi smoki…⁷  
## <fct> <dbl> <fct> <fct> <fct> <fct> <fct> <dbl> <dbl> <fct>   
## 1 Male 67 0 1 Yes Private Urban 229. 36.6 former…  
## 2 Female 61 0 0 Yes Self-e… Rural 202. NA never …  
## 3 Male 80 0 1 Yes Private Rural 106. 32.5 never …  
## 4 Female 49 0 0 Yes Private Urban 171. 34.4 smokes   
## 5 Female 79 1 0 Yes Self-e… Rural 174. 24 never …  
## 6 Male 81 0 0 Yes Private Urban 186. 29 former…  
## # … with 1 more variable: stroke <fct>, and abbreviated variable names  
## # ¹​hypertension, ²​heart\_disease, ³​ever\_married, ⁴​work\_type, ⁵​Residence\_type,  
## # ⁶​avg\_glucose\_level, ⁷​smoking\_status  
## # ℹ Use `colnames()` to see all variable names

# ————————————————————————

## Data Cleaning

In smoking\_status ‘Unknown’ should be changed to NA.

Also, it can be ordered: never < formerly < smokes

ever\_married can be re-coded as 0/1 in accordance with heart\_disease and hypertension

id can be removed

Other columns seem to be OK

## Descriptive statistics

skim(df) %>%  
 yank("factor")

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| gender | 0 | 1.0 | FALSE | 3 | Fem: 2994, Mal: 2115, Oth: 1 |
| hypertension | 0 | 1.0 | FALSE | 2 | 0: 4612, 1: 498 |
| heart\_disease | 0 | 1.0 | FALSE | 2 | 0: 4834, 1: 276 |
| ever\_married | 0 | 1.0 | FALSE | 2 | Yes: 3353, No: 1757 |
| work\_type | 0 | 1.0 | FALSE | 5 | Pri: 2925, Sel: 819, chi: 687, Gov: 657 |
| Residence\_type | 0 | 1.0 | FALSE | 2 | Urb: 2596, Rur: 2514 |
| smoking\_status | 1544 | 0.7 | FALSE | 3 | nev: 1892, for: 885, smo: 789 |
| stroke | 0 | 1.0 | FALSE | 2 | no: 4861, yes: 249 |

* Target ‘stroke’ is hugely imbalanced!
* ‘smoking\_status’ completeness rate is low
* One ‘Other’ gender can be removed

df <- df %>%   
 filter(gender != "Other")  
  
skim(df) %>%  
 yank("numeric")

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | 0 | 1.00 | 43.23 | 22.61 | 0.08 | 25.00 | 45.00 | 61.00 | 82.00 | ▅▆▇▇▆ |
| avg\_glucose\_level | 0 | 1.00 | 106.14 | 45.29 | 55.12 | 77.24 | 91.88 | 114.09 | 271.74 | ▇▃▁▁▁ |
| bmi | 201 | 0.96 | 28.89 | 7.85 | 10.30 | 23.50 | 28.10 | 33.10 | 97.60 | ▇▇▁▁▁ |

* BMI completeness rate 0.96

### How many smoking\_status in each target class?

Keep in mind that if smoking NAs are mainly in “healthy” class, we can simply remove them

df %>%   
 group\_by(stroke, smoking\_status) %>%   
 count()

## # A tibble: 8 × 3  
## # Groups: stroke, smoking\_status [8]  
## stroke smoking\_status n  
## <fct> <fct> <int>  
## 1 yes formerly smoked 70  
## 2 yes never smoked 90  
## 3 yes smokes 42  
## 4 yes <NA> 47  
## 5 no formerly smoked 814  
## 6 no never smoked 1802  
## 7 no smokes 747  
## 8 no <NA> 1497

A lot of them are in “stroke” group, some imputation will be needed.

### How many skipped BMI values in each target class?

df %>%   
 filter(is.na(bmi)) %>%   
 group\_by(stroke) %>%   
 count()

## # A tibble: 2 × 2  
## # Groups: stroke [2]  
## stroke n  
## <fct> <int>  
## 1 yes 40  
## 2 no 161

* We have too many NAs in BMI among ‘stroke-yes’ cases to simply remove them. Some imputation is needed.

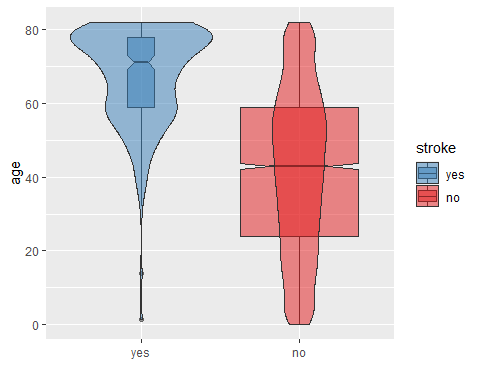
# ————————————————————————

# Exploratory Data Analysis(EDA)

## In details

### Stroke vs Age

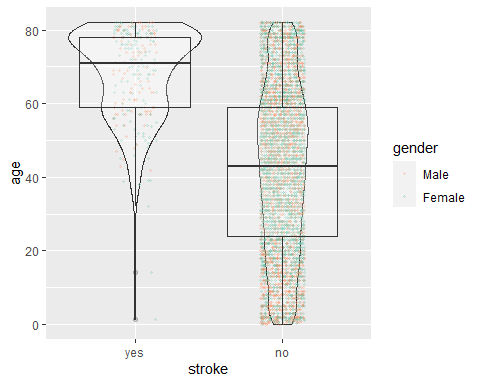
ggplot(df, aes(stroke, age)) +  
 geom\_boxplot(aes(fill = stroke), alpha = 0.5, varwidth = T, notch = T) +  
 geom\_violin(aes(fill = stroke), alpha = 0.5) +  
 scale\_fill\_brewer(palette = "Set1", direction = -1) +  
 xlab("")



* No surprises here: the older you get the higher the chance of getting stroke.
* There are observation with age much below 20 y.o., even close to 0! These are very young kids or babies - should we even include them in the analysis? If yes, the rest will be prediction only for adults.
* Stroke in kids probably has very different causes compared to stroke in adults/older folk.

### Stroke vs Age + Gender

ggplot(df, aes(stroke, age)) +   
 geom\_violin(alpha=0.3) +  
 geom\_jitter(alpha=0.2, size=0.8, width = 0.15, height = 0.1, aes(color = gender)) +   
 geom\_boxplot(alpha = 0.2) +  
 scale\_color\_brewer(palette = "Set2", direction = -1)



* No gender imbalance with respect to age and stroke

### Stroke vs Glucose

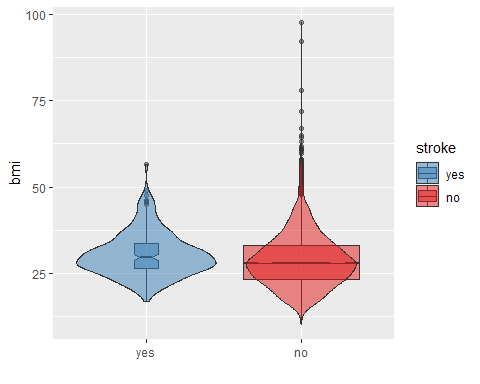
* Observations with stroke tend to have higher glucose level

This average glucose level is probably the results of fasting blood sugar test

If I correctly understand this [CDC information](https://www.cdc.gov/diabetes/basics/getting-tested.html#:~:text=Fasting%20Blood%20Sugar%20Test&text=A%20fasting%20blood%20sugar%20level,higher%20indicates%20you%20have%20diabetes.) on diabetes, values greater than 126 is evidence of diabetes. But >250? Is it realistic?

### Stroke vs BMI

ggplot(df, aes(stroke, bmi)) +  
 geom\_boxplot(aes(fill = stroke), alpha = 0.5, varwidth = T, notch = T) +  
 geom\_violin(aes(fill = stroke), alpha = 0.5) +  
 scale\_fill\_brewer(palette = "Set1", direction = -1) +  
 xlab("")

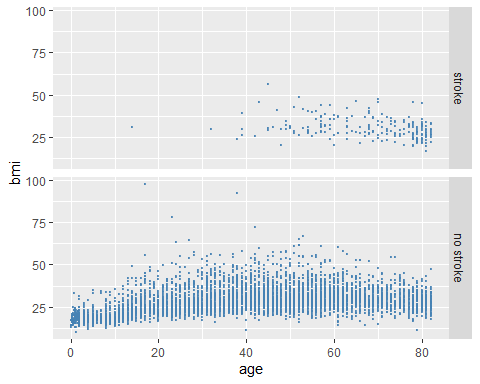


* BMI over 40 is the 3rd class of obesity - BMI over 75 should not exist at all, I think.

Let’s look’ at this weird points

### Age vs BMI

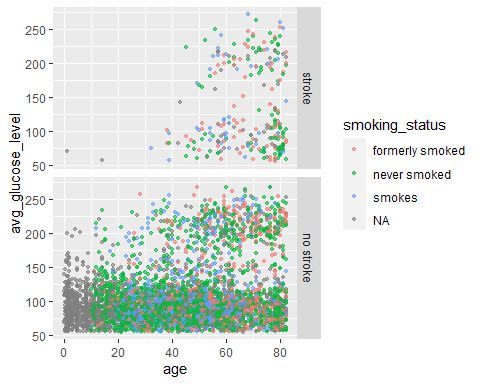
facet\_names <- c("no" = "no stroke", "yes" = "stroke")  
  
ggplot(df, aes(age, bmi)) +  
 geom\_point(color = "steelblue", alpha = 0.8, size = 0.5) +  
 facet\_grid(rows = vars(stroke), labeller = as\_labeller(facet\_names)) +  
 guides()



Patients with BMI over 75 are also very young. Suspicious.

### Glucose vs Age + smoking

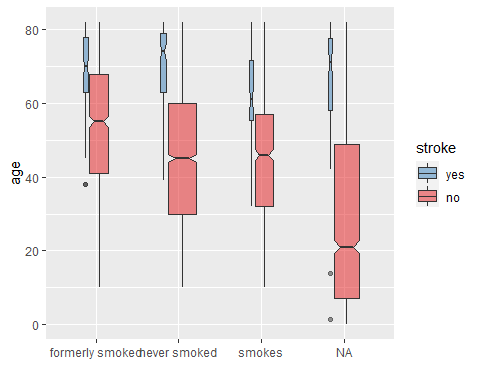
ggplot(df, aes(age, avg\_glucose\_level)) +  
 geom\_point(aes(color = smoking\_status), alpha = 0.6, size = 1) +  
 scale\_fill\_brewer(palette = "Set1", direction = -1) +  
 facet\_grid(rows = vars(stroke), labeller = as\_labeller(facet\_names)) +  
 guides()



* Children are mainly of ‘Unknown’ smoking status; both target groups are divided into two clusters – I am curious why. It is not gender, nor heart disease or any other factor we have in the data set.

### Age vs Smoking

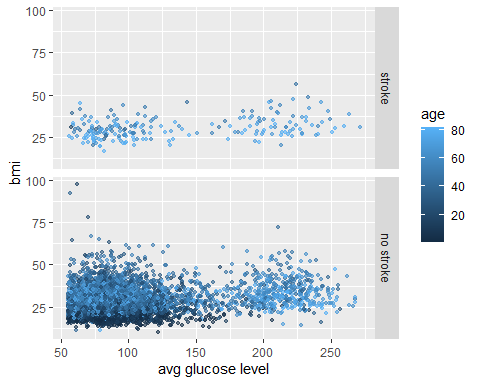
ggplot(df, aes(smoking\_status, age)) +  
 geom\_boxplot(aes(fill = stroke), alpha = 0.5, varwidth = T, notch = T) +  
 scale\_fill\_brewer(palette = "Set1", direction = -1) +  
 xlab("")



* Kids are main non-smokers of course. Note the same two outliers of age below 20 in ‘stroke-yes’

### Glucose vs BMI

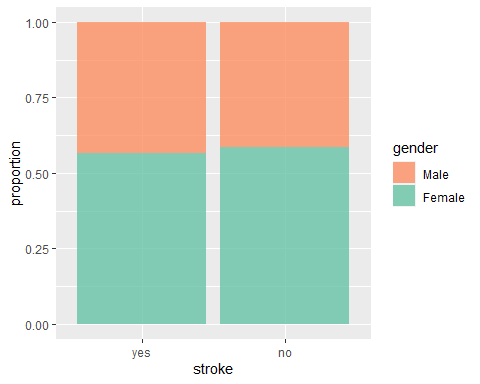
ggplot(df, aes(avg\_glucose\_level, bmi)) +  
 geom\_point(aes(color = age), alpha = 0.6, size = 1) +  
 scale\_fill\_brewer(palette = "Set1", direction = -1) +  
 facet\_grid(rows = vars(stroke), labeller = as\_labeller(facet\_names)) +  
 guides() +  
 xlab("avg glucose level")



* BMI outliers: super high BMI but super low glucose levels? How is that possible?
* Can it be a bias introduced by testing protocol misuse? Like instead of measuring glucose *before* eating, in some samples it was done *after* eating?
* Again, all the observations in both target classes form two distinct clusters.

### Stroke vs Gender

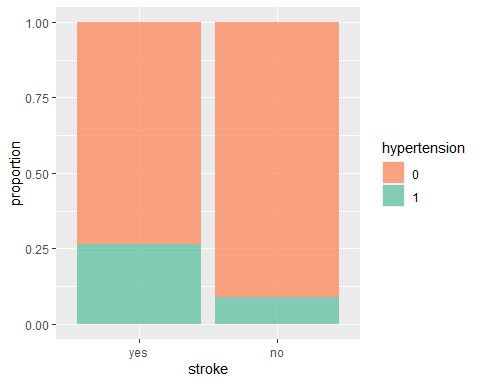
gender <- df %>%   
 group\_by(stroke, gender) %>%   
 summarize(N=n())  
  
ggplot(gender, aes(stroke, N)) +  
 geom\_bar(aes(fill=gender), alpha = 0.8, stat = "identity", position = "fill") +  
 scale\_fill\_brewer(palette = "Set2", direction = -1) +  
 ylab("proportion")



* No disproportions here

### Stroke vs Hypertension

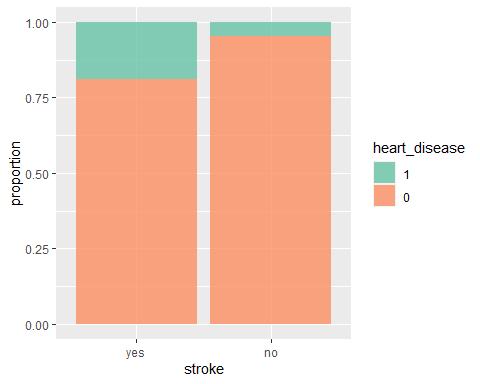
hyptens <- df %>%   
 group\_by(stroke, hypertension) %>%   
 summarize(N = n())  
  
ggplot(hyptens, aes(stroke, N)) +  
 geom\_bar(aes(fill = hypertension), alpha = 0.8, stat = "identity", position = "fill") +  
 scale\_fill\_brewer(palette = "Set2", direction = -1) +  
 ylab("proportion")



* Hypertension occurred more often in stroke-yes

### Stroke vs Heart Disease

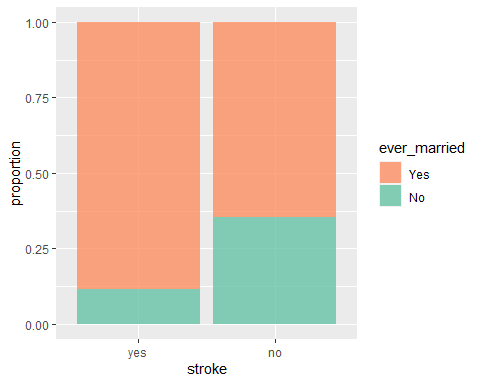
heart <- df %>%   
 group\_by(stroke, heart\_disease) %>%   
 summarize(N=n())  
  
ggplot(heart, aes(stroke, N)) +  
 geom\_bar(aes(fill = heart\_disease), alpha = 0.8, stat = "identity", position = "fill") +  
 scale\_fill\_brewer(palette = "Set2", direction = 1) +  
 ylab("proportion")



* Heart disease occurred more often in stroke-yes

### Stroke vs Ever Married

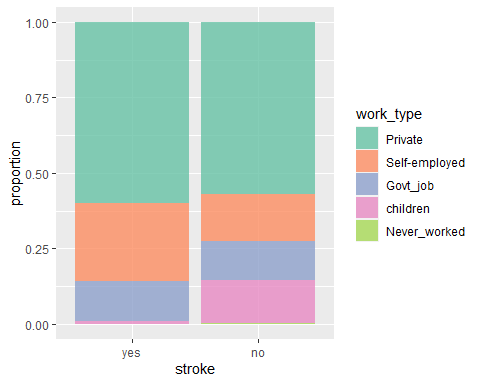
married <- df %>%   
 group\_by(stroke, ever\_married) %>%   
 summarize(N=n())  
  
ggplot(married, aes(stroke, N)) +  
 geom\_bar(aes(fill = ever\_married), alpha = 0.8, stat = "identity", position = "fill") +  
 scale\_fill\_brewer(palette = "Set2", direction = -1) +  
 ylab("proportion")



* Marriage is bad :)

### Stroke vs Work Type

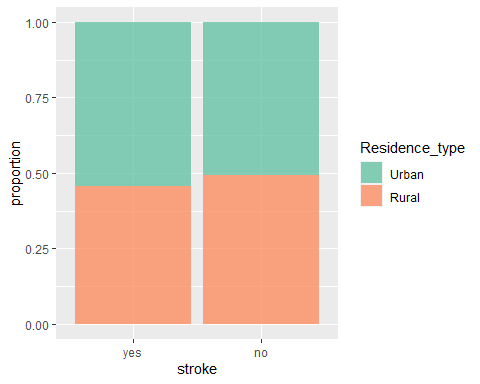
work <- df %>%   
 group\_by(stroke, work\_type) %>%   
 summarize(N=n())  
  
ggplot(work, aes(stroke, N)) +  
 geom\_bar(aes(fill = work\_type), alpha = 0.8, stat = "identity", position = "fill") +  
 scale\_fill\_brewer(palette = "Set2", direction = 1) +  
 ylab("proportion")



* It is good to be a kid :)
* It is bad to be self-employed :)

### Stroke vs Residence Type

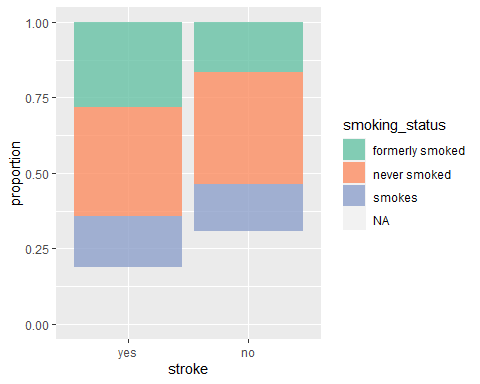
residence <- df %>%   
 group\_by(stroke, Residence\_type) %>%   
 summarize(N=n())  
  
ggplot(residence, aes(stroke, N)) +  
 geom\_bar(aes(fill = Residence\_type), alpha = 0.8, stat = "identity", position = "fill") +  
 scale\_fill\_brewer(palette = "Set2", direction = 1) +  
 ylab("proportion")



* No disproportions here

### Stroke vs Smoking

smoking <- df %>%   
 group\_by(stroke, smoking\_status) %>%   
 summarize(N=n())  
  
ggplot(smoking, aes(stroke, N)) +  
 geom\_bar(aes(fill = smoking\_status), alpha = 0.8, stat = "identity", position = "fill") +  
 scale\_fill\_brewer(palette = "Set2", direction = 1) +  
 ylab("proportion")



* There are more smokers of all types in stroke-yes

### Kids and Smoking

df %>%   
 filter(work\_type == "children") %>%   
 group\_by(smoking\_status) %>%   
 summarise(N = n(),   
 avg.age = mean(age),   
 max.age = max(age),   
 min.age = min(age))

## # A tibble: 4 × 5  
## smoking\_status N avg.age max.age min.age  
## <fct> <int> <dbl> <dbl> <dbl>  
## 1 formerly smoked 13 11.8 15 10   
## 2 never smoked 54 12.5 16 10   
## 3 smokes 2 11 12 10   
## 4 <NA> 618 6.23 16 0.08

A lot of NAs in smoking\_status comes from group ‘Children’ (see work\_type). I can replace them with ‘never smoked’ during imputation stage of the analysis.

### EDA Summary

* There are several suspicious outliers (like in BMI and glucose). So we can safely remove BMI > 75, maybe even BMI > 60 (Remember that BMI > 40 is the highest class of obesity).
* What is less safe - it is removing non-adults (younger than 20 y.o.). On one hand they provide valid information (age is very important predictor), on the other hand their stroke cases where a lot of predictors do not have sense (or are obvious NAs) for non-adults (like smoking, marriage status, employment type, residence type etc.); model-based imputation of smoking\_status in children does not have sense as well, I should rather replace with “never smoked”.
* Since, modelling using all predictors and observations has given me very moderate results (TPR = 1 comes with very high FPR and very low probability cutoff close to 0), I will try various trimming of the data.

# ————————————————————————

# Data preprocessing

## Stratified split

set.seed(124)  
  
data\_split <- initial\_split(df, prop = 3/4, strata = stroke)  
  
df\_train <- training(data\_split)  
df\_test <- testing(data\_split)

## 10-fold CV repeated 10 times

set.seed(345)  
  
# Stratified, repeated 10-fold cross-validation  
cv\_folds <- vfold\_cv(df\_train, strata = "stroke", v = 10, repeats = 1)  
  
# metrics  
cls\_metrics <- metric\_set(roc\_auc, j\_index)

## Recipe

prep\_recipe <- recipe(stroke ~ ., data = df\_train) %>%  
   
 step\_impute\_bag(bmi, smoking\_status) %>%  
 step\_dummy(all\_nominal\_predictors()) %>%   
 step\_zv(all\_predictors()) %>%   
 #step\_impute\_bag(all\_predictors()) %>%   
 step\_normalize(age, avg\_glucose\_level, bmi) %>%   
 step\_smote(stroke, over\_ratio = 1, seed = 100) %>%  
 check\_missing(all\_predictors()) %>%   
 step\_zv(all\_predictors())  
  
prep\_recipe

## Recipe  
##   
## Inputs:  
##   
## role #variables  
## outcome 1  
## predictor 10  
##   
## Operations:  
##   
## Bagged tree imputation for bmi, smoking\_status  
## Dummy variables from all\_nominal\_predictors()  
## Zero variance filter on all\_predictors()  
## Centering and scaling for age, avg\_glucose\_level, bmi  
## SMOTE based on stroke  
## Check missing values for all\_predictors()  
## Zero variance filter on all\_predictors()

# ————————————————————————

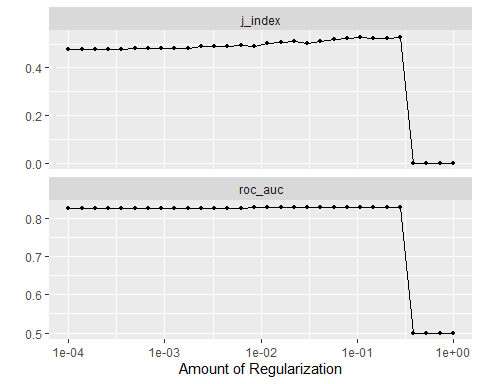
# Logistic Regression

I add one more step to the recipe - remove correlated predictors (threshold = 0.75)

**\*\* NOTE: The following module uses GridTuning for performing Cross Validation using a set of penalty terms iteratively, and is estimated to take approx. 15 mins of execution run-time on a decent system configuration.\*\***

# recipe for LR  
lr\_recipe <- prep\_recipe %>%   
 step\_corr(all\_predictors(), threshold = 0.75)  
  
# set model type/engine  
lr\_mod <- logistic\_reg(penalty = tune(), mixture = 1) %>%   
 set\_engine("glmnet")  
  
# define the workflow  
lr\_workflow <-   
 workflow() %>%   
 add\_model(lr\_mod) %>%   
 add\_recipe(lr\_recipe)  
  
# create a tune grid  
lr\_reg\_grid <- tibble(penalty = 10\*\*seq(-4, 0, length.out = 30))  
  
# train and tune the model  
# Approx. 15 mins on run-time  
lr\_res <- tune\_grid(lr\_workflow,  
 grid = lr\_reg\_grid,  
 resamples = cv\_folds,  
 control = control\_grid(save\_pred = TRUE),  
 metrics = cls\_metrics)  
  
# autoplot(lr\_res)

autoplot(lr\_res)



top\_models <-  
 lr\_res %>%   
 show\_best("roc\_auc", n = 5) %>%   
 arrange(penalty)   
  
top\_models %>% arrange(penalty)

## # A tibble: 5 × 7  
## penalty .metric .estimator mean n std\_err .config   
## <dbl> <chr> <chr> <dbl> <int> <dbl> <chr>   
## 1 0.0117 roc\_auc binary 0.828 10 0.0133 Preprocessor1\_Model16  
## 2 0.0161 roc\_auc binary 0.828 10 0.0127 Preprocessor1\_Model17  
## 3 0.0221 roc\_auc binary 0.829 10 0.0123 Preprocessor1\_Model18  
## 4 0.0304 roc\_auc binary 0.829 10 0.0118 Preprocessor1\_Model19  
## 5 0.0788 roc\_auc binary 0.828 10 0.0113 Preprocessor1\_Model22

lr\_best <- lr\_res %>%   
 select\_best(metric = "roc\_auc")  
  
lr\_best

## # A tibble: 1 × 2  
## penalty .config   
## <dbl> <chr>   
## 1 0.0221 Preprocessor1\_Model18

# ————————————————————————

# ————————————————————————

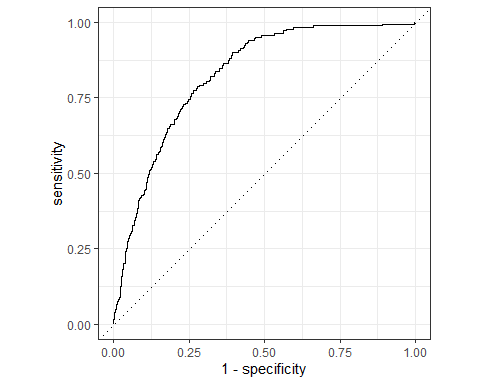
# Results

# ————————————————————————

# Final Fit

Fit the penalized Logistic Regression model with chosen hyper-parameters to the entire training data set and test it on the test data set.

lr\_auc <-   
 lr\_res %>%   
 collect\_predictions(parameters = lr\_best) %>%   
 roc\_curve(stroke, .pred\_yes) %>%   
 mutate(model = "Logistic Regression")  
  
autoplot(lr\_auc)



# ————————————————————————

### 4. Conclusion

Predictors: age, avg glucose level, hypertension, marriage, self-employed

Non-predictors: gender, bmi, smoking status

Based on our exploratory data analysis The variables that seem like they could be good stroke predictors are age, average glucose level, hypertension, marriage status, and employment status. There are variables you would expect such as, age, average glucose level, and hypertension. Older age, high glucose levels, and hypertension often lead to negative health effects. However, the fact that marriage status and employment status also up is somewhat surprising and definitely requires a more in depth analysis of why they showed up as possible stroke predictors. The variables that were non-predictors after our exploratory analysis were gender, bmi, and smoking status. It is somewhat surprising that bmi and smoking status appeared as non-predictors. Smoking and high bmi are often associated with negative health effects. It is possible these just don’t contribute to causing a stroke, so it is also worth looking at them more in depth to find out why this might be the case. This exploratory data analysis was informative, however we cannot make any definitive conclusions on good stroke predictors on that alone. To make a definitive conclusion modeling is required.

# ————————————————————————

### 5. Personal Biases

*Casey* Coming into this I already knew that age is one of the biggest factors contributing to a stroke. So, I had a pre-made assumption that when we looked at the data, age should be one of the best predictors of having a stroke or not. Will add more….

*Filip* Surprisingly, smoking status does not seem like a good predictor. However, I think it would be biased to say that smoking does not significantly contribute to stroke based on just this analysis. There are researches available which suggest the opposite. In our data, smoking is divided into three categories. The reason for smoking not being a good predictor might be the lack of information. If we knew more accurate amounts (# of cigarettes people smoke/day, etc.), the result might be different.

*Pranjal* I had a personal bias towards the societal factors such as employment and marriage status. I had an intuition that these would turn out to be the major factors primary because coming from India these societal factors play a significant role in individuals’ personal and mental life. Though we didn’t consider mental health factors but those could be important next steps. As for the scope of this project, I feel my consensus pushed me towards examining these factors in a much greater detail as compared to other relevant ones like smoking, etc.

*Sanchana* I had a leaning towards lifestyle being one of the major reasons for stroke Things such as smoking, drinking too much alcohol, being overweight and eating unhealthy foods can damage your blood vessels, increase your blood pressure and make your blood more likely to stroke. So during the analysis, I focused on the factors like BMI, and smoking to check if my hypothesis turns out to be true.

X - X