Data 621 Final Project

2024-04-30

Determinants of Recurrent Stroke Incidence in a Clinical Sample

Packages

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.2

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)
library(ggplot2)
library(summarytools)
```

```
## Warning: package 'summarytools' was built under R version 4.3.3
```

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3

## corrplot 0.92 loaded
```

```r
library(gt)
```

```
## Warning: package 'gt' was built under R version 4.3.3
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.2

## Loading required package: lattice
```

```r
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.3.2

## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 4.3.2

##
## Attaching package: 'Matrix'
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-8
```

```r
library(rpart)
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 4.3.3
```

```r
library(ggfortify)
```

```
## Warning: package 'ggfortify' was built under R version 4.3.2
```

```r
library(tibble)
```

```
##
## Attaching package: 'tibble'
```

```
## The following object is masked from 'package:summarytools':
##
##     view
```

```r
library(webshot)
```

```
## Warning: package 'webshot' was built under R version 4.3.2
```

Loading the Data

```r
stroke <- read.csv("https://raw.githubusercontent.com/Mattr5541/DATA-621-Final-Project/main/621clean_sho
```

Data Cleaning

First, I will code all "Unknown" observations as NA, as their presence may confound our analysis

```r
stroke <- stroke %>% mutate(across(where(is.character), ~na_if(., "Unknown")))
stroke <- stroke %>% mutate(across(where(is.character), ~na_if(., "Unknown or Not Reported")))
```

Recoding NA values with N where applicable

Certain variables were not coded with N where the presence of an outcome was false. This can be seen in variables with only one outcome

```r
#unique_values <- lapply(stroke, unique)
#print(unique_values)

stroke <- stroke %>%
  mutate(
    CovidAtVisitFlag = replace_na(CovidAtVisitFlag, 'N'),
    FamilyHistoryStrokeFlag = replace_na(FamilyHistoryStrokeFlag, 'N'),
    prior.COVID.19 = replace_na(prior.COVID.19, 'N'),
    hypertension = replace_na(hypertension, 'N'),
    diabetes.mellitus = replace_na(diabetes.mellitus, 'N'),
    diabetes.mellitus.type.2 = replace_na(diabetes.mellitus.type.2, 'N'),
    myocardial.infarction = replace_na(myocardial.infarction, 'N'),
    alzheimer.s.disease = replace_na(alzheimer.s.disease, 'N'),
    hyperlipidemia = replace_na(hyperlipidemia, 'N'),
    atrial.fibrillation = replace_na(atrial.fibrillation, 'N'),
    chronic.heart.disease = replace_na(chronic.heart.disease, 'N'),
    chronic.kidney.disease = replace_na(chronic.kidney.disease, 'N'),
    carotid.stenosis = replace_na(carotid.stenosis, 'N'),
    Coronary.artery.disease = replace_na(Coronary.artery.disease, 'N'),
    Heart.failure = replace_na(Heart.failure, 'N'),
    Peripheral.vascular.disease = replace_na(Peripheral.vascular.disease, 'N'),
    Dysphagia_outcome = replace_na(Dysphagia_outcome, 'N'),
    ispregnancyDoc = replace_na(ispregnancyDoc, 'N'),
    ispregnancyICD = replace_na(ispregnancyICD, 'N'),
    isTransferEvent = replace_na(isTransferEvent, 'N'))
```

Examining Missingness

My next step wil be to remove columns that present 80% or more missingness, as they will likely not contribute to our analyses, and any attempts to impute values for these columns may generate unreliable data (we may have to consider the same for columns that present 50% or more missing values)

```
miss_percent <- colSums(is.na(stroke) / 29662 * 100)

miss_percent_80 <- as.data.frame(miss_percent)

miss_percent_80 <- miss_percent_80 %>% filter(miss_percent > 79)

print(miss_percent_80)
```

```
##                        miss_percent
## alcohol_use_frequency     93.40908
## evt                       94.56207
## evt_status                97.52545
## tici_score                98.49976
```

Exploratory Data Analysis

```
stroke <- stroke %>%
  select(-alcohol_use_frequency, -evt, -evt_status, -tici_score)
```

Splitting the Data into Training/Test Sets

Before modifying the dataset any further, I will split the data into train/test partitions for the purposes of model validation (I will use a standard 80/20 split. To start, however, I want to see how evenly the binary outcomes of our target variable occur in our dataset (I'm assuming there will be an uneven split that is more biased toward negative outcomes)

```
table(stroke$TARGET)
```

```
##
##     0     1
## 25162  4500
```

As expected, there is a bias toward negative outcomes, presenting the issue of imbalance in our data. As a result, we may need to perform an oversampling or undersampling procedure to account for this, or otherwise balance observations while constructing our models.

```
set.seed(12345)

train_test <- createDataPartition(stroke$TARGET, p = 0.8, list = F)

stroke_train <- stroke[train_test, ]
stroke_test <- stroke[-train_test, ]
```

Exploratory Data Analysis

Frequencies

```
stroke_train_cat <- select_if(stroke_train, is.character)

stroke_freq <- dfSummary(stroke_train_cat, stats = 'freq')

view(stroke_freq)
```

Descriptive Statistics

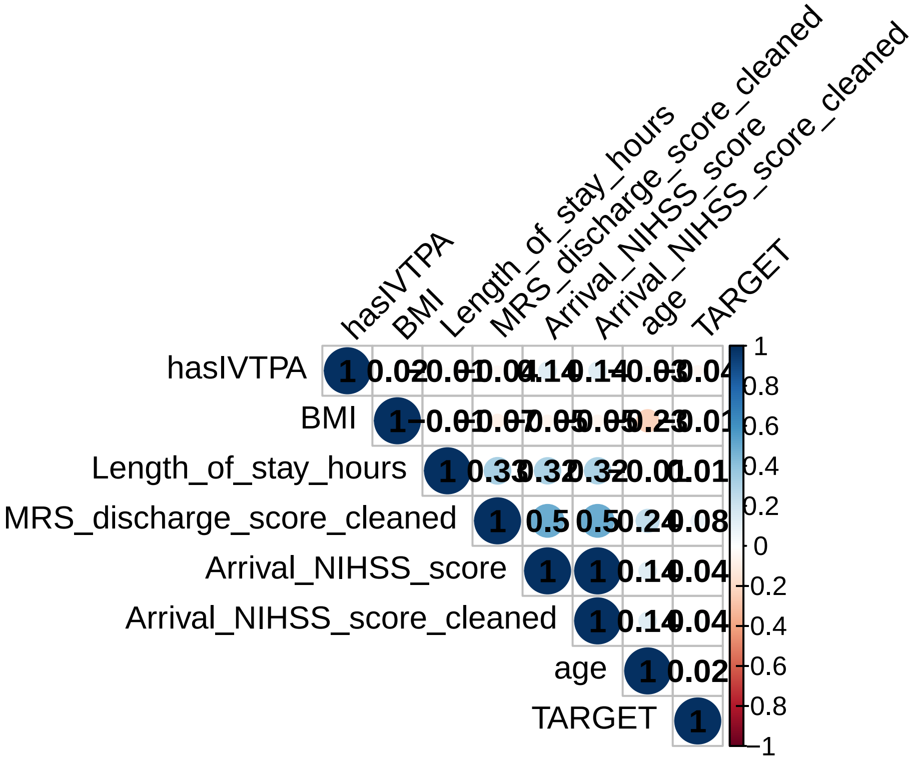
```
stroke_train_quant <- select_if(stroke_train, is.numeric)

stroke_train_quant <- stroke_train_quant %>% select(-IsIschaemicStrokeEvent) #Removing because the only

stroke_sum <- dfSummary(stroke_train_quant, stats = c("mean", "sd", "med", "IQR", "min", "max", "valid"

view(stroke_sum)
```

Correlation Matrix

```
cor_matrix = cor(stroke_train_quant, use = "complete.obs")

print(cor_matrix)
```

```
##                                      age Length_of_stay_hours
## age                          1.000000000         -0.006671007
## Length_of_stay_hours        -0.006671007          1.000000000
## MRS_discharge_score_cleaned  0.242601132          0.326620579
## Arrival_NIHSS_score          0.136791598          0.318459998
## Arrival_NIHSS_score_cleaned  0.136811194          0.318461121
## hasIVTPA                    -0.028511163         -0.013943908
## BMI                         -0.225076828         -0.007849217
## TARGET                       0.020567145          0.011010547
##                             MRS_discharge_score_cleaned Arrival_NIHSS_score
## age                                          0.24260113          0.13679160
## Length_of_stay_hours                         0.32662058          0.31846000
## MRS_discharge_score_cleaned                  1.00000000          0.49871769
## Arrival_NIHSS_score                          0.49871769          1.00000000
## Arrival_NIHSS_score_cleaned                  0.49872728          0.99999923
## hasIVTPA                                    -0.03933799          0.13865775
## BMI                                         -0.07427320         -0.05462951
## TARGET                                       0.08236076          0.04261962
##                             Arrival_NIHSS_score_cleaned    hasIVTPA
## age                                          0.13681119 -0.02851116
## Length_of_stay_hours                            0.31846112 -0.01394391
## MRS_discharge_score_cleaned                     0.49872728 -0.03933799
## Arrival_NIHSS_score                             0.99999923  0.13865775
## Arrival_NIHSS_score_cleaned                     1.00000000  0.13867002
## hasIVTPA                                        0.13867002  1.00000000
## BMI                                            -0.05462692  0.01900534
## TARGET                                          0.04262645 -0.04337908
##                                      BMI      TARGET
## age                         -0.225076828  0.02056714
## Length_of_stay_hours        -0.007849217  0.01101055
## MRS_discharge_score_cleaned -0.074273198  0.08236076
## Arrival_NIHSS_score         -0.054629515  0.04261962
## Arrival_NIHSS_score_cleaned -0.054626924  0.04262645
## hasIVTPA                     0.019005339 -0.04337908
## BMI                          1.000000000 -0.01062292
## TARGET                      -0.010622923  1.00000000
```

```
corrplot(cor_matrix, method = "circle", type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black")
```



As we can see, most of the correlations present are rather weak. The exception would be the correlations among Arrival_NHISS_score and the cleaned version (I will drop the original), and some moderate correlations among MRS_discharge_score_cleaned and the NHISS scores. Aside from that, there seems to be no real concern regarding multicollinearity among these variables.

Interestingly, there seem to be no high correlations among the predictors and target variables, suggesting that our features may be weak predictors, by themselves, of recurrent strokes, which is rather interesting, since these factors should intuitively be related to the presence of recurrent strokes.

```
stroke_train <- stroke_train %>% select(-Arrival_NIHSS_score)
stroke_test <- stroke_test %>% select(-Arrival_NIHSS_score)
```

Data Handling and Cleaning

Missing data

```
missing_percentage <- stroke_train %>%
  summarise_all(~ mean(is.na(.)) * 100)

print(missing_percentage)
```

```
##      race ethnicity     gender vital_status age age_group Length_of_stay_hours
## 1 4.71555  7.770754 0.03792668            0   0         0           0.00842815
##   visit_type IsIschaemicStrokeEvent arrival_mode arrival_from
## 1          0                      0     12.27981      2.93721
##   discharge_disposition visit_data_dispo MRS_discharge_score_cleaned
## 1            0.08849558        0.2697008                    14.26464
##   Arrival_NIHSS_score_cleaned hasIVTPA      BMI Tobacco_current_use_indicator
## 1                    25.04846        0 18.68521                      33.45976
##   Tobacco_prior_use_indicator CovidAtVisitFlag FamilyHistoryStrokeFlag
## 1                    33.45976                0                       0
##   prior.COVID.19 hypertension diabetes.mellitus diabetes.mellitus.type.2
## 1              0            0                 0                        0
##   myocardial.infarction alzheimer.s.disease hyperlipidemia atrial.fibrillation
## 1                     0                   0              0                   0
##   chronic.heart.disease chronic.kidney.disease carotid.stenosis
## 1                     0                      0                0
##   Coronary.artery.disease Heart.failure Peripheral.vascular.disease
## 1                       0             0                           0
##   Dysphagia_outcome InsuranceCategory ispregnancyDoc ispregnancyICD
## 1                 0        0.05899705              0              0
##   isTransferEvent HasAntiDepressionMedWithin1yr discharge_disposition_regex
## 1               0                             0                    12.07754
##   TARGET
## 1      0
```

The following variables have missing data. I broke them up based on the type and provided the percent of missing data to inform the best method to impute the missing data.

Continuous Variables: Length_of_stay_hours (<1%), MRS_discharge_score_cleaned (14%), Arrival_NIHSS_score_cleaned (25%), BMI(18.7%). These variables are not normally distributed so I will use median imputation

Categorical or Ordinal Variables: Race, Ethnicity, Gender, Arrival_mode, Arrival_from, Discharge_disposition, Visit_data_dispo, Tobacco_current_use_indicator, Tobacco_prior_use_indicator, InsuranceCategory, Discharge_disposition_regex. To preserve the nature of these variables, I will use mode imputation as it replaces missing values with the most frequent category.

```
mode_impute <- function(x) {
  mode_val <- names(sort(table(x), decreasing = TRUE))[1]
  x[is.na(x)] <- mode_val
  return(x)
}
```

```r
columns_to_impute <- c("race", "ethnicity", "gender", "arrival_mode", "arrival_from", "discharge_dispos

stroke_train <- stroke_train %>%
  mutate_at(.vars = columns_to_impute, .funs = mode_impute)

stroke_train$MRS_discharge_score_cleaned <- ifelse(
    is.na(stroke_train$MRS_discharge_score_cleaned),
    median(stroke_test$MRS_discharge_score_cleaned, na.rm = TRUE),
    stroke_train$MRS_discharge_score_cleaned
)

stroke_train$Arrival_NIHSS_score_cleaned <- ifelse(
    is.na(stroke_train$Arrival_NIHSS_score_cleaned),
    median(stroke_test$Arrival_NIHSS_score_cleaned, na.rm = TRUE),
    stroke_train$Arrival_NIHSS_score_cleaned
)

stroke_train$Length_of_stay_hours <- ifelse(
    is.na(stroke_train$Length_of_stay_hours),
    median(stroke_test$Length_of_stay_hours, na.rm = TRUE),
    stroke_train$Length_of_stay_hours
)

stroke_train$BMI <- ifelse(
    is.na(stroke_train$BMI),
    median(stroke_test$BMI, na.rm = TRUE),
    stroke_train$BMI
)
```

Now there is no missing data in stroke_train dataset

```r
missing_data_report = stroke_train %>%
  summarise_all(~sum(is.na(.)))

print(missing_data_report)
```

```
##   race ethnicity gender vital_status age age_group Length_of_stay_hours
## 1    0         0      0            0   0         0                    0
##   visit_type IsIschaemicStrokeEvent arrival_mode arrival_from
## 1          0                      0            0            0
##   discharge_disposition visit_data_dispo MRS_discharge_score_cleaned
## 1                     0                0                           0
##   Arrival_NIHSS_score_cleaned hasIVTPA BMI Tobacco_current_use_indicator
## 1                           0        0   0                             0
##   Tobacco_prior_use_indicator CovidAtVisitFlag FamilyHistoryStrokeFlag
## 1                           0                0                       0
##   prior.COVID.19 hypertension diabetes.mellitus diabetes.mellitus.type.2
## 1              0            0                 0                        0
##   myocardial.infarction alzheimer.s.disease hyperlipidemia atrial.fibrillation
## 1                     0                   0              0                   0
##   chronic.heart.disease chronic.kidney.disease carotid.stenosis
## 1                     0                      0                0
##   Coronary.artery.disease Heart.failure Peripheral.vascular.disease
## 1                       0             0                           0
##   Dysphagia_outcome InsuranceCategory ispregnancyDoc ispregnancyICD
## 1                 0                 0              0              0
##   isTransferEvent HasAntiDepressionMedWithin1yr discharge_disposition_regex
## 1               0                             0                           0
##   TARGET
## 1      0
```

```r
#imputing testing dataset
mode_impute <- function(x) {
  mode_val <- names(sort(table(x), decreasing = TRUE))[1]
  x[is.na(x)] <- mode_val
  return(x)
}
columns_to_impute <- c("race", "ethnicity", "gender", "arrival_mode", "arrival_from", "discharge_dispos

stroke_test <- stroke_test %>%
  mutate_at(.vars = columns_to_impute, .funs = mode_impute)

stroke_test$MRS_discharge_score_cleaned <- ifelse(is.na(stroke_test$MRS_discharge_score_cleaned), median
stroke_test$Arrival_NIHSS_score_cleaned <- ifelse(is.na(stroke_test$Arrival_NIHSS_score_cleaned), median
stroke_test$Length_of_stay_hours <- ifelse(is.na(stroke_test$Length_of_stay_hours), median(stroke_test$
stroke_test$BMI <- ifelse(is.na(stroke_test$BMI), median(stroke_test$BMI, na.rm = TRUE), stroke_test$BMI
```

```r
missing_data_test = stroke_test%>%
  summarise_all(~sum(is.na(.)))

print(missing_data_test)
```

```
##   race ethnicity gender vital_status age age_group Length_of_stay_hours
## 1    0         0      0            0   0         0                    0
##   visit_type IsIschaemicStrokeEvent arrival_mode arrival_from
## 1          0                      0            0            0
##   discharge_disposition visit_data_dispo MRS_discharge_score_cleaned
## 1                     0                0                           0
##   Arrival_NIHSS_score_cleaned hasIVTPA BMI Tobacco_current_use_indicator
## 1                           0        0   0                             0
##   Tobacco_prior_use_indicator CovidAtVisitFlag FamilyHistoryStrokeFlag
## 1                           0                0                       0
##   prior.COVID.19 hypertension diabetes.mellitus diabetes.mellitus.type.2
## 1              0            0                 0                        0
##   myocardial.infarction alzheimer.s.disease hyperlipidemia atrial.fibrillation
## 1                     0                   0              0                   0
##   chronic.heart.disease chronic.kidney.disease carotid.stenosis
## 1                     0                      0                0
##   Coronary.artery.disease Heart.failure Peripheral.vascular.disease
## 1                       0             0                           0
##   Dysphagia_outcome InsuranceCategory ispregnancyDoc ispregnancyICD
## 1                 0                 0              0              0
##   isTransferEvent HasAntiDepressionMedWithin1yr discharge_disposition_regex
## 1               0                             0                           0
##   TARGET
## 1      0
```

###Dummy Coding Categorical Variables

Creating dummy coding for categorical variables, in both training and testing datasets, results in a format that helps prepare data for further analysis. The '-1' part of the code was done to avoid multicollinearity issues.

```
# Function to create dummy variables with consistent naming
create_dummies <- function(data, variable_name) {
  dummies <- model.matrix(~ get(variable_name) - 1, data=data)
  colnames(dummies) <- paste("dummy", variable_name, gsub("(Intercept)|get\\(variable_name\\)", "", col
  return(dummies)
}

# List of categorical variables
variables_list <- c("race", "ethnicity", "gender", "vital_status", "age_group", "visit_type",
                    "Tobacco_current_use_indicator", "Tobacco_prior_use_indicator",
                    "FamilyHistoryStrokeFlag", "hypertension", "diabetes.mellitus",
                    "diabetes.mellitus.type.2", "myocardial.infarction", "hyperlipidemia",
                    "atrial.fibrillation", "chronic.heart.disease", "chronic.kidney.disease",
                    "Coronary.artery.disease", "Heart.failure", "Dysphagia_outcome",
                    "isTransferEvent")

# Apply the function to both datasets using a loop to create dummy variables
for (var in variables_list) {
  stroke_train[paste("dummy", var, sep="_")] <- create_dummies(stroke_train, var)
  stroke_test[paste("dummy", var, sep="_")] <- create_dummies(stroke_test, var)
}
```

Transformation

Log transformation on the variable BMI should prove to be helpful since the range of 2 to 259 is unrealistic in real world metrics (on both the higher and smaller end). The same transformation on Length_of_stay_hours would also be useful as there likely should not be negative hours nor 9,666 hours (max value) which estimates to over a year.

```
stroke_train[] <- lapply(stroke_train, function(x) {
    if(is.factor(x)) factor(x) else x
})



stroke_train$log_BMI <- log(stroke_train$BMI + 1)
stroke_train$log_Length_of_stay_hours <- log(stroke_train$Length_of_stay_hours + 1)
```

```
## Warning in log(stroke_train$Length_of_stay_hours + 1): NaNs produced
```

```
stroke_test$log_BMI <- log(stroke_test$BMI + 1)
stroke_test$log_Length_of_stay_hours <- log(stroke_test$Length_of_stay_hours + 1)

print(sum(is.na(stroke_train$log_BMI)))
```

```
## [1] 0
```

```
print(sum(is.na(stroke_train$log_Length_of_stay_hours)))
```

```
## [1] 1
```

```
print(sum(is.na(stroke_test$log_BMI)))
```

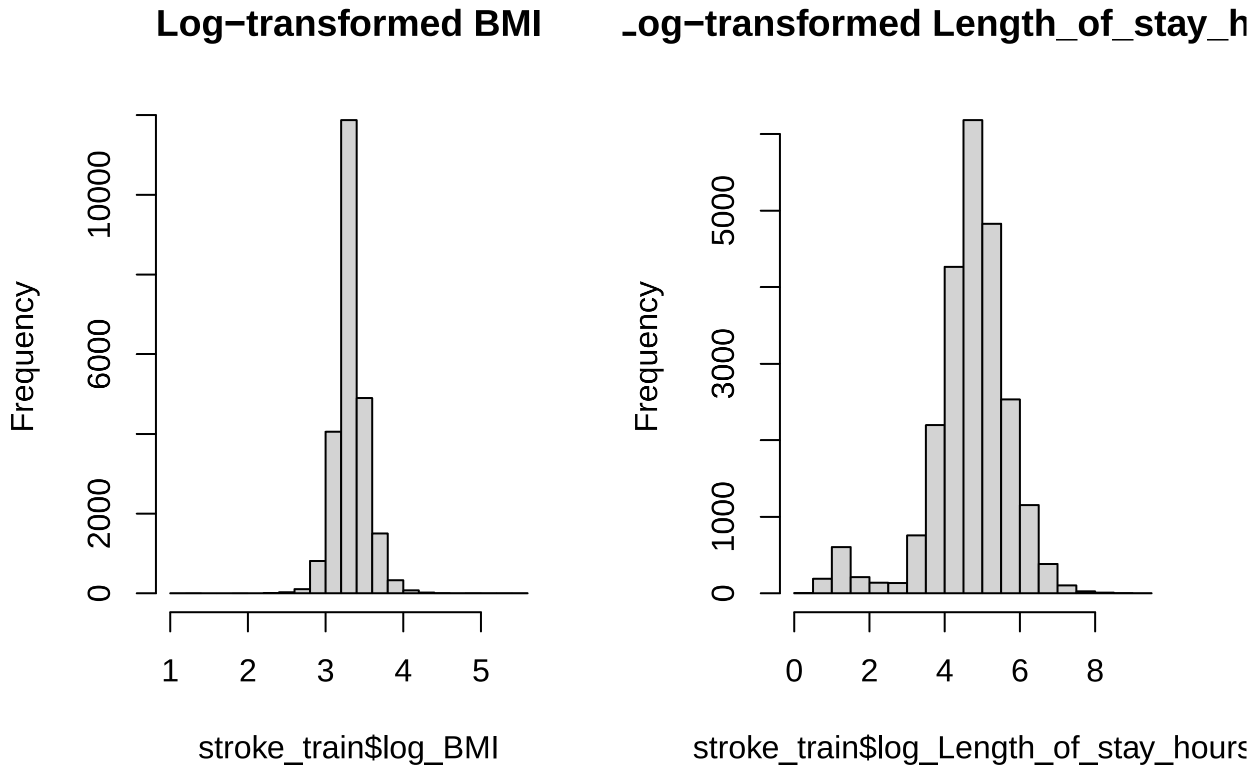
```
## [1] 0
```

```
print(sum(is.na(stroke_test$log_Length_of_stay_hours)))
```

```
## [1] 0
```

```
train_stats <- dfSummary(stroke_train, stats = c("mean", "sd", "med", "IQR", "min", "max", "valid", "n.r

view(train_stats)
```

```
# Check the histograms of the log-transformed variables
par(mfrow=c(1,2))
hist(stroke_train$log_BMI, main = "Log-transformed BMI")
hist(stroke_train$log_Length_of_stay_hours, main = "Log-transformed Length_of_stay_hours")
```



Building the Model

```
stroke_train = stroke_train %>%
  select(where(~!is.character(.)))

stroke_test = stroke_test %>%
  select(where(~!is.character(.)))
```

Filtering out all the categorical variables

```
model <- glm(TARGET ~ .,
             data = stroke_train,
             family = binomial)

summary(model)
```

Logistic Regression:

```
##
## Call:
## glm(formula = TARGET ~ ., family = binomial, data = stroke_train)
##
## Coefficients: (22 not defined because of singularities)
##                                                                               Estimate
## (Intercept)                                                                  4.4821495
## age                                                                         -0.0144018
## Length_of_stay_hours                                                        -0.0003143
## IsIschaemicStrokeEvent                                                              NA
## MRS_discharge_score_cleaned                                                  0.1050894
## Arrival_NIHSS_score_cleaned                                                  0.0112489
## hasIVTPA                                                                    -0.3499272
## BMI                                                                          0.0105910
## dummy_racedummy_race_American Indian or Alaska Native                       -0.3474007
## dummy_racedummy_race_Asian                                                  -0.2273886
## dummy_racedummy_race_Black or African American                               0.2359376
## dummy_racedummy_race_More Than One Race                                     -0.1151115
## dummy_racedummy_race_Native Hawaiian or Other Pacific Islander              -0.3275489
## dummy_racedummy_race_White                                                          NA
## dummy_ethnicitydummy_ethnicity_Hispanic or Latino                            0.2064781
## dummy_ethnicitydummy_ethnicity_Not Hispanic or Latino                               NA
## dummy_genderdummy_gender_Female                                              0.0929291
## dummy_genderdummy_gender_Male                                                       NA
## dummy_vital_statusdummy_vital_status_Alive                                  -0.1010330
## dummy_vital_statusdummy_vital_status_Dead                                           NA
## dummy_age_groupdummy_age_group_<50                                           0.0107823
## dummy_age_groupdummy_age_group_>=80                                          0.0119313
## dummy_age_groupdummy_age_group_50-79                                                NA
## dummy_visit_typedummy_visit_type_Emergency                                  -0.4717882
## dummy_visit_typedummy_visit_type_Inpatient                                          NA
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_No   0.3983462
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_Yes         NA
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_No      -0.0641784
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_Yes             NA
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_N                 0.5097498
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_Y                        NA
## dummy_hypertensiondummy_hypertension_N                                      -0.9819220
## dummy_hypertensiondummy_hypertension_Y                                              NA
## dummy_diabetes.mellitusdummy_diabetes.mellitus_N                             0.3699231
## dummy_diabetes.mellitusdummy_diabetes.mellitus_Y                                    NA
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_N              -0.6016662
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_Y                      NA
## dummy_myocardial.infarctiondummy_myocardial.infarction_N                    -0.4022660
## dummy_myocardial.infarctiondummy_myocardial.infarction_Y                            NA
## dummy_hyperlipidemiadummy_hyperlipidemia_N                                  -1.0997547
## dummy_hyperlipidemiadummy_hyperlipidemia_Y                                          NA
## dummy_atrial.fibrillationdummy_atrial.fibrillation_N                        -0.2125255
## dummy_atrial.fibrillationdummy_atrial.fibrillation_Y                               NA
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_N                   -0.5262765
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_Y                           NA
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_N                 -0.5033149
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_Y                         NA
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_N               -0.4363701
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_Y                       NA
## dummy_Heart.failuredummy_Heart.failure_N                                    0.0757810
## dummy_Heart.failuredummy_Heart.failure_Y                                           NA
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_N                           -0.3298076
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_Y                                   NA
## dummy_isTransferEventdummy_isTransferEvent_N                                0.2652456
## dummy_isTransferEventdummy_isTransferEvent_Y                                       NA
## log_BMI                                                                    -0.9666890
## log_Length_of_stay_hours                                                   -0.1478540
##                                                                            Std. Error
## (Intercept)                                                                 0.7291355
## age                                                                         0.0029604
## Length_of_stay_hours                                                        0.0001723
## IsIschaemicStrokeEvent                                                             NA
## MRS_discharge_score_cleaned                                                 0.0141039
## Arrival_NIHSS_score_cleaned                                                 0.0032510
## hasIVTPA                                                                    0.0700594
## BMI                                                                         0.0068155
## dummy_racedummy_race_American Indian or Alaska Native                       0.2623641
## dummy_racedummy_race_Asian                                                  0.0775305
## dummy_racedummy_race_Black or African American                              0.0493272
## dummy_racedummy_race_More Than One Race                                     0.0729466
## dummy_racedummy_race_Native Hawaiian or Other Pacific Islander              0.6238266
## dummy_racedummy_race_White                                                         NA
## dummy_ethnicitydummy_ethnicity_Hispanic or Latino                           0.0808706
## dummy_ethnicitydummy_ethnicity_Not Hispanic or Latino                              NA
## dummy_genderdummy_gender_Female                                             0.0413190
## dummy_genderdummy_gender_Male                                                      NA
## dummy_vital_statusdummy_vital_status_Alive                                  0.0516526
## dummy_vital_statusdummy_vital_status_Dead                                          NA
## dummy_age_groupdummy_age_group_<50                                          0.1129058
## dummy_age_groupdummy_age_group_>=80                                         0.0699231
## dummy_age_groupdummy_age_group_50-79                                               NA
## dummy_visit_typedummy_visit_type_Emergency                                  0.1760342
## dummy_visit_typedummy_visit_type_Inpatient                                         NA
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_No   0.1054387
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_Yes         NA
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_No       0.0633770
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_Yes             NA
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_N                0.1150478
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_Y                       NA
## dummy_hypertensiondummy_hypertension_N                                      0.0630915
## dummy_hypertensiondummy_hypertension_Y                                             NA
## dummy_diabetes.mellitusdummy_diabetes.mellitus_N                            0.3438371
## dummy_diabetes.mellitusdummy_diabetes.mellitus_Y                                   NA
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_N              0.3439960
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_Y                     NA
## dummy_myocardial.infarctiondummy_myocardial.infarction_N                    0.0737301
## dummy_myocardial.infarctiondummy_myocardial.infarction_Y                           NA
## dummy_hyperlipidemiadummy_hyperlipidemia_N                                  0.0436181
## dummy_hyperlipidemiadummy_hyperlipidemia_Y                                         NA
## dummy_atrial.fibrillationdummy_atrial.fibrillation_N                        0.0501355
## dummy_atrial.fibrillationdummy_atrial.fibrillation_Y                               NA
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_N                    0.0688670
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_Y                           NA
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_N                  0.0538533
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_Y                         NA
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_N                0.0500726
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_Y                       NA
## dummy_Heart.failuredummy_Heart.failure_N                                    0.0666364
## dummy_Heart.failuredummy_Heart.failure_Y                                           NA
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_N                            0.0521607
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_Y                                   NA
## dummy_isTransferEventdummy_isTransferEvent_N                                0.0756136
## dummy_isTransferEventdummy_isTransferEvent_Y                                       NA
## log_BMI                                                                     0.2401457
## log_Length_of_stay_hours                                                    0.0409921
##                                                                             z value
## (Intercept)                                                                   6.147
## age                                                                          -4.865
## Length_of_stay_hours                                                         -1.824
## IsIschaemicStrokeEvent                                                           NA
## MRS_discharge_score_cleaned                                                   7.451
## Arrival_NIHSS_score_cleaned                                                   3.460
## hasIVTPA                                                                     -4.995
## BMI                                                                           1.554
## dummy_racedummy_race_American Indian or Alaska Native                        -1.324
## dummy_racedummy_race_Asian                                                   -2.933
## dummy_racedummy_race_Black or African American                                4.783
## dummy_racedummy_race_More Than One Race                                      -1.578
## dummy_racedummy_race_Native Hawaiian or Other Pacific Islander               -0.525
## dummy_racedummy_race_White                                                       NA
## dummy_ethnicitydummy_ethnicity_Hispanic or Latino                             2.553
## dummy_ethnicitydummy_ethnicity_Not Hispanic or Latino                            NA
## dummy_genderdummy_gender_Female                                               2.249
## dummy_genderdummy_gender_Male                                                    NA
## dummy_vital_statusdummy_vital_status_Alive                                   -1.956
## dummy_vital_statusdummy_vital_status_Dead                                        NA
## dummy_age_groupdummy_age_group_<50                                            0.095
## dummy_age_groupdummy_age_group_>=80                                           0.171
## dummy_age_groupdummy_age_group_50-79                                             NA
## dummy_visit_typedummy_visit_type_Emergency                                   -2.680
## dummy_visit_typedummy_visit_type_Inpatient                                       NA
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_No    3.778
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_Yes      NA
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_No       -1.013
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_Yes          NA
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_N                  4.431
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_Y                     NA
## dummy_hypertensiondummy_hypertension_N                                      -15.563
## dummy_hypertensiondummy_hypertension_Y                                           NA
## dummy_diabetes.mellitusdummy_diabetes.mellitus_N                              1.076
## dummy_diabetes.mellitusdummy_diabetes.mellitus_Y                                NA
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_N              -1.749
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_Y                  NA
## dummy_myocardial.infarctiondummy_myocardial.infarction_N                    -5.456
## dummy_myocardial.infarctiondummy_myocardial.infarction_Y                        NA
## dummy_hyperlipidemiadummy_hyperlipidemia_N                                 -25.213
## dummy_hyperlipidemiadummy_hyperlipidemia_Y                                      NA
## dummy_atrial.fibrillationdummy_atrial.fibrillation_N                        -4.239
## dummy_atrial.fibrillationdummy_atrial.fibrillation_Y                            NA
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_N                    -7.642
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_Y                        NA
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_N                  -9.346
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_Y                      NA
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_N                -8.715
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_Y                    NA
## dummy_Heart.failuredummy_Heart.failure_N                                     1.137
## dummy_Heart.failuredummy_Heart.failure_Y                                        NA
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_N                            -6.323
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_Y                                NA
## dummy_isTransferEventdummy_isTransferEvent_N                                 3.508
## dummy_isTransferEventdummy_isTransferEvent_Y                                    NA
## log_BMI                                                                     -4.025
## log_Length_of_stay_hours                                                    -3.607
##                                                                            Pr(>|z|)
## (Intercept)                                                                7.89e-10
## age                                                                        1.15e-06
## Length_of_stay_hours                                                       0.068083
## IsIschaemicStrokeEvent                                                           NA
## MRS_discharge_score_cleaned                                                9.26e-14
## Arrival_NIHSS_score_cleaned                                                0.000540
## hasIVTPA                                                                   5.89e-07
## BMI                                                                        0.120196
## dummy_racedummy_race_American Indian or Alaska Native                      0.185464
## dummy_racedummy_race_Asian                                                 0.003358
## dummy_racedummy_race_Black or African American                             1.73e-06
## dummy_racedummy_race_More Than One Race                                    0.114560
## dummy_racedummy_race_Native Hawaiian or Other Pacific Islander             0.599539
## dummy_racedummy_race_White                                                       NA
## dummy_ethnicitydummy_ethnicity_Hispanic or Latino                          0.010674
## dummy_ethnicitydummy_ethnicity_Not Hispanic or Latino                            NA
## dummy_genderdummy_gender_Female                                            0.024508
## dummy_genderdummy_gender_Male                                                    NA
## dummy_vital_statusdummy_vital_status_Alive                                 0.050464
## dummy_vital_statusdummy_vital_status_Dead                                        NA
## dummy_age_groupdummy_age_group_<50                                         0.923919
## dummy_age_groupdummy_age_group_>=80                                        0.864511
## dummy_age_groupdummy_age_group_50-79                                             NA
## dummy_visit_typedummy_visit_type_Emergency                                 0.007360
## dummy_visit_typedummy_visit_type_Inpatient                                       NA
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_No  0.000158
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_Yes       NA
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_No      0.311230
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_Yes           NA
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_N               9.39e-06
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_Y                     NA
## dummy_hypertensiondummy_hypertension_N                                      < 2e-16
## dummy_hypertensiondummy_hypertension_Y                                           NA
## dummy_diabetes.mellitusdummy_diabetes.mellitus_N                           0.281987
## dummy_diabetes.mellitusdummy_diabetes.mellitus_Y                                 NA
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_N             0.080282
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_Y                   NA
## dummy_myocardial.infarctiondummy_myocardial.infarction_N                   4.87e-08
## dummy_myocardial.infarctiondummy_myocardial.infarction_Y                         NA
## dummy_hyperlipidemiadummy_hyperlipidemia_N                                  < 2e-16
## dummy_hyperlipidemiadummy_hyperlipidemia_Y                                       NA
## dummy_atrial.fibrillationdummy_atrial.fibrillation_N                       2.24e-05
## dummy_atrial.fibrillationdummy_atrial.fibrillation_Y                             NA
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_N                   2.14e-14
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_Y                         NA
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_N                  < 2e-16
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_Y                       NA
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_N                < 2e-16
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_Y                     NA
## dummy_Heart.failuredummy_Heart.failure_N                                   0.255442
## dummy_Heart.failuredummy_Heart.failure_Y                                         NA
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_N                           2.57e-10
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_Y                                 NA
## dummy_isTransferEventdummy_isTransferEvent_N                               0.000452
## dummy_isTransferEventdummy_isTransferEvent_Y                                     NA
## log_BMI                                                                    5.69e-05
## log_Length_of_stay_hours                                                   0.000310
##
## (Intercept)                                                                ***
## age                                                                        ***
## Length_of_stay_hours                                                       .
## IsIschaemicStrokeEvent
## MRS_discharge_score_cleaned                                                ***
## Arrival_NIHSS_score_cleaned                                                ***
## hasIVTPA                                                                   ***
## BMI
## dummy_racedummy_race_American Indian or Alaska Native
## dummy_racedummy_race_Asian                                                 **
## dummy_racedummy_race_Black or African American                             ***
## dummy_racedummy_race_More Than One Race
## dummy_racedummy_race_Native Hawaiian or Other Pacific Islander
## dummy_racedummy_race_White
## dummy_ethnicitydummy_ethnicity_Hispanic or Latino                          *
## dummy_ethnicitydummy_ethnicity_Not Hispanic or Latino
## dummy_genderdummy_gender_Female                                            *
## dummy_genderdummy_gender_Male
## dummy_vital_statusdummy_vital_status_Alive                                 .
## dummy_vital_statusdummy_vital_status_Dead
## dummy_age_groupdummy_age_group_<50
## dummy_age_groupdummy_age_group_>=80
## dummy_age_groupdummy_age_group_50-79
## dummy_visit_typedummy_visit_type_Emergency                                 **
## dummy_visit_typedummy_visit_type_Inpatient
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_No  ***
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_Yes
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_No
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_Yes
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_N               ***
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_Y
## dummy_hypertensiondummy_hypertension_N                                     ***
## dummy_hypertensiondummy_hypertension_Y
## dummy_diabetes.mellitusdummy_diabetes.mellitus_N
## dummy_diabetes.mellitusdummy_diabetes.mellitus_Y
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_N             .
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_Y
## dummy_myocardial.infarctiondummy_myocardial.infarction_N                   ***
## dummy_myocardial.infarctiondummy_myocardial.infarction_Y
## dummy_hyperlipidemiadummy_hyperlipidemia_N                                 ***
## dummy_hyperlipidemiadummy_hyperlipidemia_Y
## dummy_atrial.fibrillationdummy_atrial.fibrillation_N                       ***
## dummy_atrial.fibrillationdummy_atrial.fibrillation_Y
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_N                   ***
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_Y
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_N                 ***
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_Y
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_N               ***
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_Y
## dummy_Heart.failuredummy_Heart.failure_N
## dummy_Heart.failuredummy_Heart.failure_Y
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_N                           ***
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_Y
## dummy_isTransferEventdummy_isTransferEvent_N                               ***
## dummy_isTransferEventdummy_isTransferEvent_Y
## log_BMI                                                                    ***
## log_Length_of_stay_hours                                                   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 20077  on 23728  degrees of freedom
## Residual deviance: 16983  on 23694  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 17053
##
## Number of Fisher Scoring iterations: 5
```

The summary yields the result "Coefficients: (28 not defined because of singularities)." In an effort to counter this negative result, I will perform both Ridge and Lasso Regressions.

```
# Extract predictor variables and the target variable from the stroke_train data
# Impute the missing value using the median of the column
#stroke_train$log_Length_of_stay_hours[is.na(stroke_train$log_Length_of_stay_hours)] <- median(stroke_t

stroke_train[] <- lapply(stroke_train, function(x) {
    if(is.factor(x)) factor(x) else x
})
```

```r
# Impute the missing value using the median of the column
stroke_train$log_Length_of_stay_hours[is.na(stroke_train$log_Length_of_stay_hours)] <- median(stroke_tra

X <- model.matrix(~ . - 1 - TARGET, data = stroke_train)  # Remove intercept term
Y <- stroke_train$TARGET

# Fit the Lasso logistic regression model
lasso_model <- glmnet(X, Y, family = "binomial", alpha = 1)  # Set alpha = 1 for Lasso penalty

# Get the lambda values from the Lasso model
lambda_values <- lasso_model$lambda

# Perform cross-validation
cv_model <- cv.glmnet(X, Y, alpha = 1, lambda = lambda_values, nfolds = 10)

# Select optimal lambda
optimal_lambda <- cv_model$lambda.min
print(paste("Optimal Lambda:", optimal_lambda))
```

Lasso:

```
## [1] "Optimal Lambda: 0.00029892056083166"
```

```r
# Refit final model
lasso_model <- glmnet(X, Y, family = "binomial", alpha = 1, lambda = optimal_lambda)

print(lasso_model)
```

```
##
## Call:  glmnet(x = X, y = Y, family = "binomial", alpha = 1, lambda = optimal_lambda)
##
##   Df %Dev    Lambda
## 1 48 15.4 0.0002989
```

Lasso Regression is helpful here as the variable selection is almost "automatic" and given that the dataset has a lot of predictors, this is extremely valuable. Lasso regression also helps with possible over-fitting as well.

```r

# Extract predictor variables and the target variable from the stroke_train data
X <- model.matrix(~ . - 1 - TARGET, data = stroke_train)  # Remove intercept term
Y <- stroke_train$TARGET

# Perform cross-validation to select optimal lambda value
cv_model <- cv.glmnet(X, Y, family = "binomial", alpha = 0, type.measure = "deviance")

# Select optimal lambda
optimal_lambda <- cv_model$lambda.min
print(paste("Optimal Lambda:", optimal_lambda))
```

Ridge Regression:

```
## [1] "Optimal Lambda: 0.00956326727080695"
```

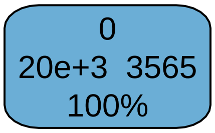
```r
# Fit the ridge logistic regression model with the selected lambda value
ridge_model <- glmnet(X, Y, family = "binomial", alpha = 0, lambda = optimal_lambda)

```

```
# Summary of the ridge logistic regression model
print(ridge_model)
```

```
##
## Call:  glmnet(x = X, y = Y, family = "binomial", alpha = 0, lambda = optimal_lambda)
##
##   Df  %Dev   Lambda
## 1 55 15.36 0.009563
```

Ridge regression is also a great method for "removing" the affect caused by irrelevant predictors in the dataset and therefore in the model. Though, the difference with Ridge regression is that Ridge regression does not "remove" the more irrelevant predictors, which could help with multicollinearity in a more graceful manner.

```

tree_model <- rpart(TARGET ~ ., data = stroke_train, method = "class")

rpart.plot(tree_model, type = 4, extra = 101)
```



Decision Trees

Given that the goal is predict a binary "TARGET" variable with many different variables, using a decision tree may be an advantageous method given that decision trees automatically select the most relevant variables.

Model Validation and Selection

I will use mean standard error to conduct model validation.

```
# Logistic Regression
predicted_vals_lr <- predict(model, newdata = stroke_train, type = "response")
observed_vals_lr <- stroke_train$TARGET
```

```r
res_lr <- observed_vals_lr - predicted_vals_lr
mse_lr <- mean(res_lr^2)

# Lasso Regression
predicted_vals_lasso <- predict(lasso_model, newx = X, s = optimal_lambda, type = "response")
observed_vals_lasso <- Y
res_lasso <- observed_vals_lasso - predicted_vals_lasso
mse_lasso <- mean(res_lasso^2)

# Ridge Regression
predicted_vals_ridge <- predict(ridge_model, newx = X, s = optimal_lambda, type = "response")
observed_vals_ridge <- Y
res_ridge <- observed_vals_ridge - predicted_vals_ridge
mse_ridge <- mean(res_ridge^2)

# Decision Tree Model
tree_model <- rpart(TARGET ~ ., data = stroke_train, method = "class")
predicted_vals_tree <- predict(tree_model, newdata = stroke_train, type = "prob")[,2]  # assuming TARGE
observed_vals_tree <- stroke_train$TARGET
res_tree <- observed_vals_tree - predicted_vals_tree
mse_tree <- mean(res_tree^2)

# Print MSE of each model
cat("Logistic Regression MSE:", mse_lr, "\n")
```

```
## Logistic Regression MSE: 0.1094173
```

```r
cat("Lasso Regression MSE:", mse_lasso, "\n")
```

```
## Lasso Regression MSE: 0.1094412
```

```r
cat("Ridge Regression MSE:", mse_ridge, "\n")
```

```
## Ridge Regression MSE: 0.1095066
```

```r
cat("Decision Tree MSE:", mse_tree, "\n")
```

```
## Decision Tree MSE: 0.1276622
```

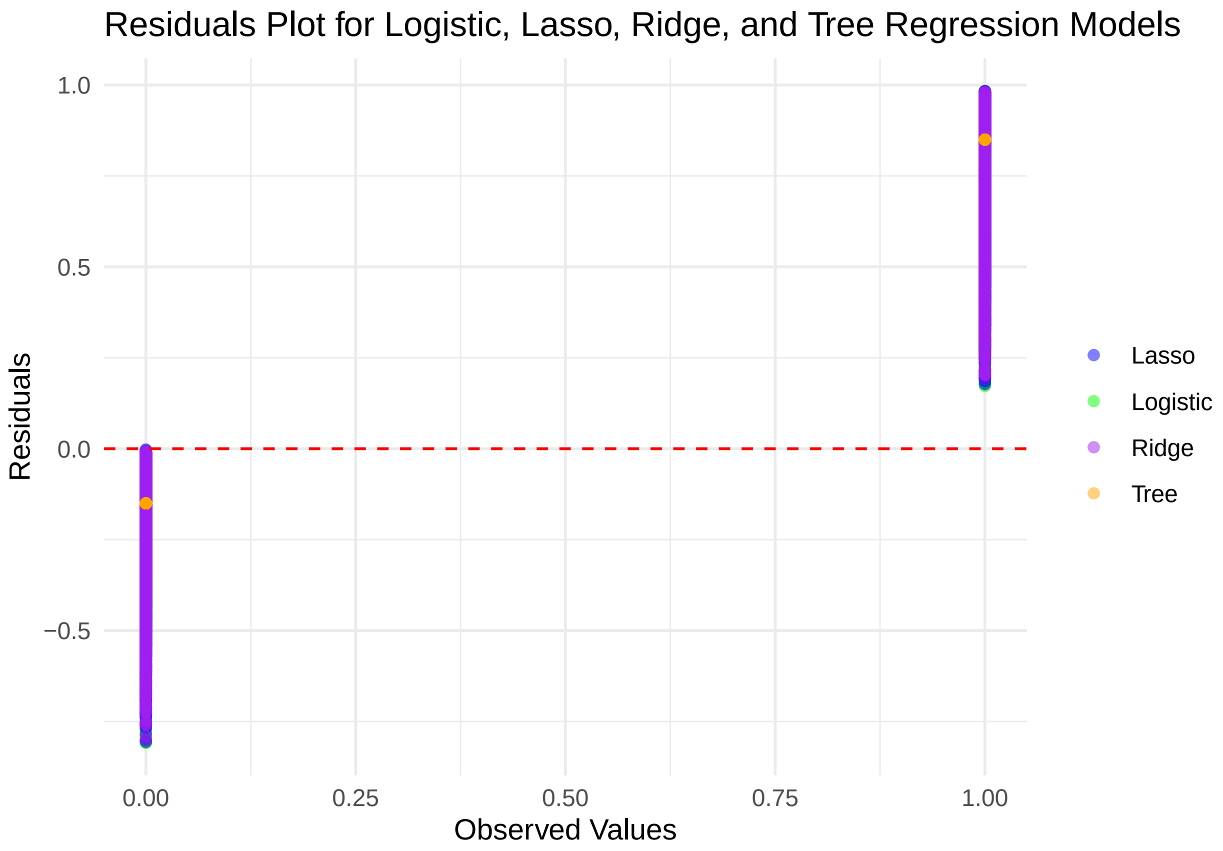
The mean squared error (MSE) results from the different models in our analysis present a close comparison, particularly among the logistic, Lasso, and Ridge regression models, with the Decision Tree model performing slightly worse. The Logistic Regression model achieved the lowest MSE at 0.1094177, indicating it was the most accurate in predicting the target variable among the four models. This suggests that the logistic model, despite its simplicity relative to the regularized models, managed to fit the data slightly better without overfitting, as the regularization in Lasso and Ridge did not significantly enhance the model accuracy in this case.

The Lasso Regression is almost identical to that of the Logistic Regression, at 0.1094384, showing that the penalty applied to reduce the coefficients of less important predictors did not substantially improve prediction accuracy. Similarly, the Ridge Regression, which also applies a penalty but does not reduce coefficients to zero, showed a marginally higher MSE of 0.1095066. This implies that the penalty in Ridge, which aims to handle multicollinearity and reduce model complexity, was also not significantly beneficial in this context.

The Decision Tree model had the highest MSE at 0.1276622, suggesting it was less effective at predicting the target compared to the regression-based models. This could be due to the model overfitting the training data or not capturing the linear relationships as effectively as the regression models. Decision trees are typically more sensitive to the specific structure of the training data and can lead to higher variance if not properly tuned or if the data does not support the tree's split criteria well.

Now, I will plot the residuals to visualize:

```
residuals_data <- data.frame(
  Observed = c(observed_vals_lr, observed_vals_lasso, observed_vals_ridge, observed_vals_tree),
  Residuals = c(res_lr, res_lasso, res_ridge, res_tree),
  Model = factor(c(rep("Logistic", length(res_lr)),
                   rep("Lasso", length(res_lasso)),
                   rep("Ridge", length(res_ridge)),
                   rep("Tree", length(res_tree)))
                 )
)

ggplot(residuals_data, aes(x = Observed, y = Residuals, color = Model)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals Plot for Logistic, Lasso, Ridge, and Tree Regression Models",
       x = "Observed Values",
       y = "Residuals") +
  theme_minimal() +
  scale_color_manual(values = c("blue", "green", "purple", "orange")) +
  theme(legend.title = element_blank())
```



Moving forward, lets use the Lgistic model

Now lets make predictions using the test dataset.

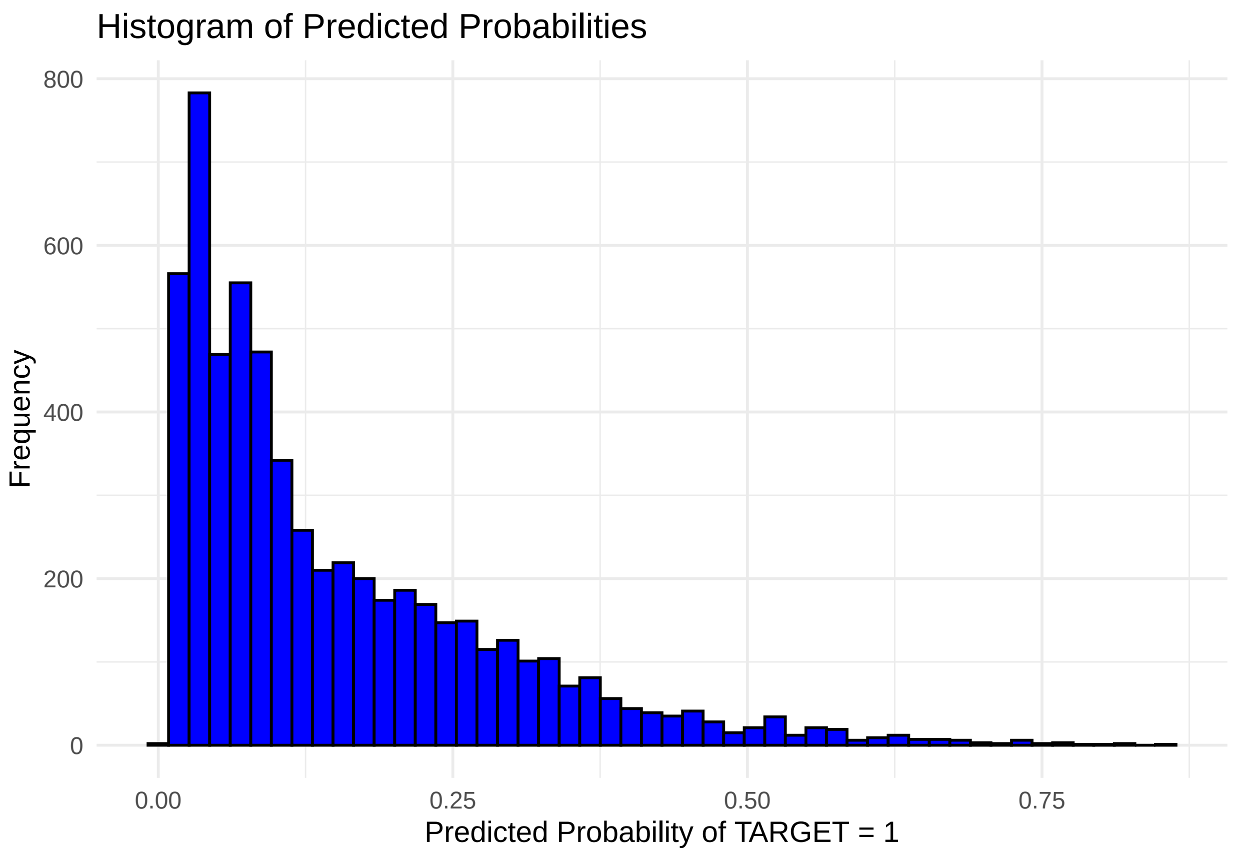
```r
names(stroke_test) <- gsub("test", "train", names(stroke_test))
print(colnames(stroke_test))
```

```
##  [1] "age"                                 "Length_of_stay_hours"
##  [3] "IsIschaemicStrokeEvent"              "MRS_discharge_score_cleaned"
##  [5] "Arrival_NIHSS_score_cleaned"         "hasIVTPA"
##  [7] "BMI"                                 "TARGET"
##  [9] "dummy_race"                          "dummy_ethnicity"
## [11] "dummy_gender"                        "dummy_vital_status"
## [13] "dummy_age_group"                     "dummy_visit_type"
## [15] "dummy_Tobacco_current_use_indicator" "dummy_Tobacco_prior_use_indicator"
## [17] "dummy_FamilyHistoryStrokeFlag"       "dummy_hypertension"
## [19] "dummy_diabetes.mellitus"             "dummy_diabetes.mellitus.type.2"
## [21] "dummy_myocardial.infarction"         "dummy_hyperlipidemia"
## [23] "dummy_atrial.fibrillation"           "dummy_chronic.heart.disease"
## [25] "dummy_chronic.kidney.disease"        "dummy_Coronary.artery.disease"
## [27] "dummy_Heart.failure"                 "dummy_Dysphagia_outcome"
## [29] "dummy_isTransferEvent"               "log_BMI"
## [31] "log_Length_of_stay_hours"
```

```r
# Create model matrix for the test dataset
X_test <- model.matrix(~ . - 1, data = stroke_test)

# Recheck column names to ensure they match those used in the training model
print(colnames(X_test))
```

```
##  [1] "age"
##  [2] "Length_of_stay_hours"
##  [3] "IsIschaemicStrokeEvent"
##  [4] "MRS_discharge_score_cleaned"
##  [5] "Arrival_NIHSS_score_cleaned"
##  [6] "hasIVTPA"
##  [7] "BMI"
##  [8] "TARGET"
##  [9] "dummy_racedummy_race_American Indian or Alaska Native"
## [10] "dummy_racedummy_race_Asian"
## [11] "dummy_racedummy_race_Black or African American"
## [12] "dummy_racedummy_race_More Than One Race"
## [13] "dummy_racedummy_race_Native Hawaiian or Other Pacific Islander"
## [14] "dummy_racedummy_race_White"
## [15] "dummy_ethnicitydummy_ethnicity_Hispanic or Latino"
## [16] "dummy_ethnicitydummy_ethnicity_Not Hispanic or Latino"
## [17] "dummy_genderdummy_gender_Female"
## [18] "dummy_genderdummy_gender_Male"
## [19] "dummy_vital_statusdummy_vital_status_Alive"
## [20] "dummy_vital_statusdummy_vital_status_Dead"
## [21] "dummy_age_groupdummy_age_group_<50"
## [22] "dummy_age_groupdummy_age_group_>=80"
## [23] "dummy_age_groupdummy_age_group_50-79"
## [24] "dummy_visit_typedummy_visit_type_Emergency"
## [25] "dummy_visit_typedummy_visit_type_Inpatient"
## [26] "dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_No"
## [27] "dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_Yes"
## [28] "dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_No"
## [29] "dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_Yes"
## [30] "dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_N"
## [31] "dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_Y"
## [32] "dummy_hypertensiondummy_hypertension_N"
## [33] "dummy_hypertensiondummy_hypertension_Y"
## [34] "dummy_diabetes.mellitusdummy_diabetes.mellitus_N"
## [35] "dummy_diabetes.mellitusdummy_diabetes.mellitus_Y"
## [36] "dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_N"
## [37] "dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_Y"
## [38] "dummy_myocardial.infarctiondummy_myocardial.infarction_N"
## [39] "dummy_myocardial.infarctiondummy_myocardial.infarction_Y"
## [40] "dummy_hyperlipidemiadummy_hyperlipidemia_N"
## [41] "dummy_hyperlipidemiadummy_hyperlipidemia_Y"
## [42] "dummy_atrial.fibrillationdummy_atrial.fibrillation_N"
## [43] "dummy_atrial.fibrillationdummy_atrial.fibrillation_Y"
## [44] "dummy_chronic.heart.diseasedummy_chronic.heart.disease_N"
## [45] "dummy_chronic.heart.diseasedummy_chronic.heart.disease_Y"
## [46] "dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_N"
## [47] "dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_Y"
## [48] "dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_N"
## [49] "dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_Y"
## [50] "dummy_Heart.failuredummy_Heart.failure_N"
## [51] "dummy_Heart.failuredummy_Heart.failure_Y"
## [52] "dummy_Dysphagia_outcomedummy_Dysphagia_outcome_N"
## [53] "dummy_Dysphagia_outcomedummy_Dysphagia_outcome_Y"
## [54] "dummy_isTransferEventdummy_isTransferEvent_N"
## [55] "dummy_isTransferEventdummy_isTransferEvent_Y"
## [56] "log_BMI"
## [57] "log_Length_of_stay_hours"
```

```r
# Predict using the logistic regression model
predicted_probs <- predict(model, newdata = stroke_test, type = "response")

# Create a data frame for visualization
predictions <- data.frame(Probability = predicted_probs)


ggplot(predictions, aes(x = Probability)) +
  geom_histogram(bins = 50, fill = "blue", color = "black") +
  ggtitle("Histogram of Predicted Probabilities") +
  xlab("Predicted Probability of TARGET = 1") +
  ylab("Frequency") +
  theme_minimal()
```



Model Interpretation

We therefore select the logistic regression model as the model to interpret for our project.

```
coefficients <- coef(model)

print(summary(model))
```

```
##
## Call:
## glm(formula = TARGET ~ ., family = binomial, data = stroke_train)
##
## Coefficients: (22 not defined because of singularities)
##                                                                 Estimate
## (Intercept)                                                    4.4821495
## age                                                           -0.0144018
## Length_of_stay_hours                                          -0.0003143
## IsIschaemicStrokeEvent                                                NA
## MRS_discharge_score_cleaned                                    0.1050894
## Arrival_NIHSS_score_cleaned                                    0.0112489
## hasIVTPA                                                      -0.3499272
## BMI                                                            0.0105910
## dummy_racedummy_race_American Indian or Alaska Native         -0.3474007
## dummy_racedummy_race_Asian                                    -0.2273886
## dummy_racedummy_race_Black or African American                 0.2359376
## dummy_racedummy_race_More Than One Race                       -0.1151115
## dummy_racedummy_race_Native Hawaiian or Other Pacific Islander -0.3275489
## dummy_racedummy_race_White                                                            NA
## dummy_ethnicitydummy_ethnicity_Hispanic or Latino                             0.2064781
## dummy_ethnicitydummy_ethnicity_Not Hispanic or Latino                                NA
## dummy_genderdummy_gender_Female                                               0.0929291
## dummy_genderdummy_gender_Male                                                        NA
## dummy_vital_statusdummy_vital_status_Alive                                   -0.1010330
## dummy_vital_statusdummy_vital_status_Dead                                            NA
## dummy_age_groupdummy_age_group_<50                                            0.0107823
## dummy_age_groupdummy_age_group_>=80                                           0.0119313
## dummy_age_groupdummy_age_group_50-79                                                 NA
## dummy_visit_typedummy_visit_type_Emergency                                   -0.4717882
## dummy_visit_typedummy_visit_type_Inpatient                                           NA
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_No   0.3983462
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_Yes         NA
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_No      -0.0641784
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_Yes             NA
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_N                 0.5097498
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_Y                        NA
## dummy_hypertensiondummy_hypertension_N                                       -0.9819220
## dummy_hypertensiondummy_hypertension_Y                                               NA
## dummy_diabetes.mellitusdummy_diabetes.mellitus_N                              0.3699231
## dummy_diabetes.mellitusdummy_diabetes.mellitus_Y                                     NA
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_N               -0.6016662
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_Y                       NA
## dummy_myocardial.infarctiondummy_myocardial.infarction_N                     -0.4022660
## dummy_myocardial.infarctiondummy_myocardial.infarction_Y                             NA
## dummy_hyperlipidemiadummy_hyperlipidemia_N                                   -1.0997547
## dummy_hyperlipidemiadummy_hyperlipidemia_Y                                           NA
## dummy_atrial.fibrillationdummy_atrial.fibrillation_N                         -0.2125255
## dummy_atrial.fibrillationdummy_atrial.fibrillation_Y                                 NA
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_N                     -0.5262765
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_Y                             NA
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_N                   -0.5033149
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_Y                           NA
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_N                 -0.4363701
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_Y                         NA
## dummy_Heart.failuredummy_Heart.failure_N                                      0.0757810
## dummy_Heart.failuredummy_Heart.failure_Y                                             NA
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_N                             -0.3298076
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_Y                                     NA
## dummy_isTransferEventdummy_isTransferEvent_N                                  0.2652456
## dummy_isTransferEventdummy_isTransferEvent_Y                                         NA
## log_BMI                                                                      -0.9666890
## log_Length_of_stay_hours                                                     -0.1478540
##                                                                              Std. Error
## (Intercept)                                                                   0.7291355
## age                                                                           0.0029604
## Length_of_stay_hours                                                          0.0001723
## IsIschaemicStrokeEvent                                                               NA
## MRS_discharge_score_cleaned                                                   0.0141039
## Arrival_NIHSS_score_cleaned                                                   0.0032510
## hasIVTPA                                                                      0.0700594
## BMI                                                                           0.0068155
## dummy_racedummy_race_American Indian or Alaska Native                         0.2623641
## dummy_racedummy_race_Asian                                                  0.0775305
## dummy_racedummy_race_Black or African American                              0.0493272
## dummy_racedummy_race_More Than One Race                                     0.0729466
## dummy_racedummy_race_Native Hawaiian or Other Pacific Islander              0.6238266
## dummy_racedummy_race_White                                                         NA
## dummy_ethnicitydummy_ethnicity_Hispanic or Latino                           0.0808706
## dummy_ethnicitydummy_ethnicity_Not Hispanic or Latino                              NA
## dummy_genderdummy_gender_Female                                             0.0413190
## dummy_genderdummy_gender_Male                                                      NA
## dummy_vital_statusdummy_vital_status_Alive                                  0.0516526
## dummy_vital_statusdummy_vital_status_Dead                                          NA
## dummy_age_groupdummy_age_group_<50                                          0.1129058
## dummy_age_groupdummy_age_group_>=80                                         0.0699231
## dummy_age_groupdummy_age_group_50-79                                               NA
## dummy_visit_typedummy_visit_type_Emergency                                  0.1760342
## dummy_visit_typedummy_visit_type_Inpatient                                         NA
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_No   0.1054387
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_Yes         NA
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_No      0.0633770
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_Yes             NA
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_N              0.1150478
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_Y                     NA
## dummy_hypertensiondummy_hypertension_N                                      0.0630915
## dummy_hypertensiondummy_hypertension_Y                                             NA
## dummy_diabetes.mellitusdummy_diabetes.mellitus_N                            0.3438371
## dummy_diabetes.mellitusdummy_diabetes.mellitus_Y                                   NA
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_N              0.3439960
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_Y                     NA
## dummy_myocardial.infarctiondummy_myocardial.infarction_N                    0.0737301
## dummy_myocardial.infarctiondummy_myocardial.infarction_Y                           NA
## dummy_hyperlipidemiadummy_hyperlipidemia_N                                  0.0436181
## dummy_hyperlipidemiadummy_hyperlipidemia_Y                                         NA
## dummy_atrial.fibrillationdummy_atrial.fibrillation_N                        0.0501355
## dummy_atrial.fibrillationdummy_atrial.fibrillation_Y                               NA
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_N                    0.0688670
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_Y                           NA
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_N                  0.0538533
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_Y                         NA
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_N                0.0500726
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_Y                       NA
## dummy_Heart.failuredummy_Heart.failure_N                                    0.0666364
## dummy_Heart.failuredummy_Heart.failure_Y                                           NA
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_N                            0.0521607
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_Y                                   NA
## dummy_isTransferEventdummy_isTransferEvent_N                                0.0756136
## dummy_isTransferEventdummy_isTransferEvent_Y                                       NA
## log_BMI                                                                     0.2401457
## log_Length_of_stay_hours                                                    0.0409921
##                                                                             z value
## (Intercept)                                                                   6.147
## age                                                                          -4.865
## Length_of_stay_hours                                                         -1.824
## IsIschaemicStrokeEvent                                                           NA
## MRS_discharge_score_cleaned                                                   7.451
## Arrival_NIHSS_score_cleaned                                                  3.460
## hasIVTPA                                                                    -4.995
## BMI                                                                          1.554
## dummy_racedummy_race_American Indian or Alaska Native                       -1.324
## dummy_racedummy_race_Asian                                                  -2.933
## dummy_racedummy_race_Black or African American                               4.783
## dummy_racedummy_race_More Than One Race                                     -1.578
## dummy_racedummy_race_Native Hawaiian or Other Pacific Islander              -0.525
## dummy_racedummy_race_White                                                      NA
## dummy_ethnicitydummy_ethnicity_Hispanic or Latino                            2.553
## dummy_ethnicitydummy_ethnicity_Not Hispanic or Latino                           NA
## dummy_genderdummy_gender_Female                                              2.249
## dummy_genderdummy_gender_Male                                                   NA
## dummy_vital_statusdummy_vital_status_Alive                                  -1.956
## dummy_vital_statusdummy_vital_status_Dead                                       NA
## dummy_age_groupdummy_age_group_<50                                           0.095
## dummy_age_groupdummy_age_group_>=80                                          0.171
## dummy_age_groupdummy_age_group_50-79                                            NA
## dummy_visit_typedummy_visit_type_Emergency                                  -2.680
## dummy_visit_typedummy_visit_type_Inpatient                                      NA
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_No    3.778
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_Yes      NA
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_No       -1.013
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_Yes          NA
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_N                 4.431
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_Y                    NA
## dummy_hypertensiondummy_hypertension_N                                     -15.563
## dummy_hypertensiondummy_hypertension_Y                                          NA
## dummy_diabetes.mellitusdummy_diabetes.mellitus_N                             1.076
## dummy_diabetes.mellitusdummy_diabetes.mellitus_Y                                NA
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_N              -1.749
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_Y                  NA
## dummy_myocardial.infarctiondummy_myocardial.infarction_N                    -5.456
## dummy_myocardial.infarctiondummy_myocardial.infarction_Y                        NA
## dummy_hyperlipidemiadummy_hyperlipidemia_N                                 -25.213
## dummy_hyperlipidemiadummy_hyperlipidemia_Y                                      NA
## dummy_atrial.fibrillationdummy_atrial.fibrillation_N                        -4.239
## dummy_atrial.fibrillationdummy_atrial.fibrillation_Y                            NA
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_N                    -7.642
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_Y                        NA
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_N                  -9.346
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_Y                      NA
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_N                -8.715
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_Y                    NA
## dummy_Heart.failuredummy_Heart.failure_N                                     1.137
## dummy_Heart.failuredummy_Heart.failure_Y                                        NA
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_N                            -6.323
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_Y                                NA
## dummy_isTransferEventdummy_isTransferEvent_N                                 3.508
## dummy_isTransferEventdummy_isTransferEvent_Y                                    NA
## log_BMI                                                                     -4.025
## log_Length_of_stay_hours                                                    -3.607
##                                                                            Pr(>|z|)
## (Intercept)                                                                7.89e-10
## age                                                                        1.15e-06
## Length_of_stay_hours                                                       0.068083
## IsIschaemicStrokeEvent                                                           NA
## MRS_discharge_score_cleaned                                                9.26e-14
## Arrival_NIHSS_score_cleaned                                                0.000540
## hasIVTPA                                                                   5.89e-07
## BMI                                                                        0.120196
## dummy_racedummy_race_American Indian or Alaska Native                      0.185464
## dummy_racedummy_race_Asian                                                 0.003358
## dummy_racedummy_race_Black or African American                             1.73e-06
## dummy_racedummy_race_More Than One Race                                    0.114560
## dummy_racedummy_race_Native Hawaiian or Other Pacific Islander             0.599539
## dummy_racedummy_race_White                                                       NA
## dummy_ethnicitydummy_ethnicity_Hispanic or Latino                          0.010674
## dummy_ethnicitydummy_ethnicity_Not Hispanic or Latino                            NA
## dummy_genderdummy_gender_Female                                            0.024508
## dummy_genderdummy_gender_Male                                                    NA
## dummy_vital_statusdummy_vital_status_Alive                                 0.050464
## dummy_vital_statusdummy_vital_status_Dead                                        NA
## dummy_age_groupdummy_age_group_<50                                         0.923919
## dummy_age_groupdummy_age_group_>=80                                        0.864511
## dummy_age_groupdummy_age_group_50-79                                             NA
## dummy_visit_typedummy_visit_type_Emergency                                 0.007360
## dummy_visit_typedummy_visit_type_Inpatient                                       NA
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_No  0.000158
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_Yes       NA
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_No      0.311230
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_Yes           NA
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_N               9.39e-06
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_Y                     NA
## dummy_hypertensiondummy_hypertension_N                                      < 2e-16
## dummy_hypertensiondummy_hypertension_Y                                           NA
## dummy_diabetes.mellitusdummy_diabetes.mellitus_N                           0.281987
## dummy_diabetes.mellitusdummy_diabetes.mellitus_Y                                 NA
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_N             0.080282
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_Y                   NA
## dummy_myocardial.infarctiondummy_myocardial.infarction_N                   4.87e-08
## dummy_myocardial.infarctiondummy_myocardial.infarction_Y                         NA
## dummy_hyperlipidemiadummy_hyperlipidemia_N                                  < 2e-16
## dummy_hyperlipidemiadummy_hyperlipidemia_Y                                       NA
## dummy_atrial.fibrillationdummy_atrial.fibrillation_N                       2.24e-05
## dummy_atrial.fibrillationdummy_atrial.fibrillation_Y                             NA
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_N                   2.14e-14
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_Y                         NA
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_N                  < 2e-16
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_Y                       NA
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_N                < 2e-16
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_Y                     NA
## dummy_Heart.failuredummy_Heart.failure_N                                   0.255442
## dummy_Heart.failuredummy_Heart.failure_Y                                         NA
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_N                           2.57e-10
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_Y                                 NA
## dummy_isTransferEventdummy_isTransferEvent_N                               0.000452
## dummy_isTransferEventdummy_isTransferEvent_Y                                     NA
## log_BMI                                                                    5.69e-05
## log_Length_of_stay_hours                                                   0.000310
##                                                                                   
## (Intercept)                                                                ***
## age                                                                        ***
## Length_of_stay_hours                                                       .  
## IsIschaemicStrokeEvent                                                        
## MRS_discharge_score_cleaned                                                ***
## Arrival_NIHSS_score_cleaned                                                ***
## hasIVTPA                                                                   ***
## BMI                                                                           
## dummy_racedummy_race_American Indian or Alaska Native                         
## dummy_racedummy_race_Asian                                                 ** 
## dummy_racedummy_race_Black or African American                             ***
## dummy_racedummy_race_More Than One Race                                       
## dummy_racedummy_race_Native Hawaiian or Other Pacific Islander                
## dummy_racedummy_race_White                                                    
## dummy_ethnicitydummy_ethnicity_Hispanic or Latino                          *  
## dummy_ethnicitydummy_ethnicity_Not Hispanic or Latino                         
## dummy_genderdummy_gender_Female                                            *  
## dummy_genderdummy_gender_Male                                                 
## dummy_vital_statusdummy_vital_status_Alive                                 .  
## dummy_vital_statusdummy_vital_status_Dead                                     
## dummy_age_groupdummy_age_group_<50                                            
## dummy_age_groupdummy_age_group_>=80                                           
## dummy_age_groupdummy_age_group_50-79                                          
## dummy_visit_typedummy_visit_type_Emergency                                 ** 
## dummy_visit_typedummy_visit_type_Inpatient                                    
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_No  ***
## dummy_Tobacco_current_use_indicatordummy_Tobacco_current_use_indicator_Yes    
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_No         
## dummy_Tobacco_prior_use_indicatordummy_Tobacco_prior_use_indicator_Yes        
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_N               ***
## dummy_FamilyHistoryStrokeFlagdummy_FamilyHistoryStrokeFlag_Y                  
## dummy_hypertensiondummy_hypertension_N                                     ***
## dummy_hypertensiondummy_hypertension_Y                                        
## dummy_diabetes.mellitusdummy_diabetes.mellitus_N                              
## dummy_diabetes.mellitusdummy_diabetes.mellitus_Y                              
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_N             .  
## dummy_diabetes.mellitus.type.2dummy_diabetes.mellitus.type.2_Y                
## dummy_myocardial.infarctiondummy_myocardial.infarction_N                   ***
## dummy_myocardial.infarctiondummy_myocardial.infarction_Y                      
## dummy_hyperlipidemiadummy_hyperlipidemia_N                                 ***
## dummy_hyperlipidemiadummy_hyperlipidemia_Y                                    
## dummy_atrial.fibrillationdummy_atrial.fibrillation_N                       ***
## dummy_atrial.fibrillationdummy_atrial.fibrillation_Y                          
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_N                   ***
## dummy_chronic.heart.diseasedummy_chronic.heart.disease_Y                      
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_N                 ***
## dummy_chronic.kidney.diseasedummy_chronic.kidney.disease_Y                    
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_N               ***
## dummy_Coronary.artery.diseasedummy_Coronary.artery.disease_Y                  
## dummy_Heart.failuredummy_Heart.failure_N                                      
## dummy_Heart.failuredummy_Heart.failure_Y                                      
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_N                            ***
## dummy_Dysphagia_outcomedummy_Dysphagia_outcome_Y
## dummy_isTransferEventdummy_isTransferEvent_N                                ***
## dummy_isTransferEventdummy_isTransferEvent_Y
## log_BMI                                                                     ***
## log_Length_of_stay_hours                                                    ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 20077  on 23728  degrees of freedom
## Residual deviance: 16983  on 23694  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 17053
##
## Number of Fisher Scoring iterations: 5
```

Our logistic regression model reveals a mixture of significant and non-significant predictors. Notably, the model adjusts for various clinical and demographic factors.

Significant variables like age, hasIVTPA, and MRS discharge score cleaned have direct implications on our model's ability to predict the TARGET.

Age and hasIVTPA (intravenous thrombolysis treatment) are inversely related to the likelihood of the TARGET, indicating that younger ages and those not receiving IVTPA might have different outcomes relative to the base case. On the other hand, an increase in MRS discharge score, which measures the degree of disability or dependence in daily activities, positively influences the TARGET, suggesting higher scores (more disability) are associated with the outcome.

Interestingly, several dummy variables representing racial and ethnic categories were significant but showed no consistent trend in influence across the groups, highlighting the complexity of how these socio-demographic factors interact with medical outcomes.

Notable among the findings is the variable 'dummy__hypertension__N', which significantly predicts the TARGET when hypertension is absent, reflecting a strong protective effect against the condition modeled. Conversely, variables like dummy__hyperlipidemia__N and other chronic conditions displayed strong positive associations with the TARGET, suggesting that these conditions might increase the likelihood of the outcome.

The coefficients for log-transformed BMI and log-transformed Length of stay hours were also significant, suggesting that as these values increase, they have a discernible impact on the likelihood of the TARGET, though the relationship with BMI was negative, indicating a complex interaction potentially mediated by other factors.

The presence of 'NA' across numerous coefficients indicates issues of multicollinearity or perfect separation, where some predictors perfectly predict the outcome, thus are not included in the final model due to redundancy or statistical indefiniteness.

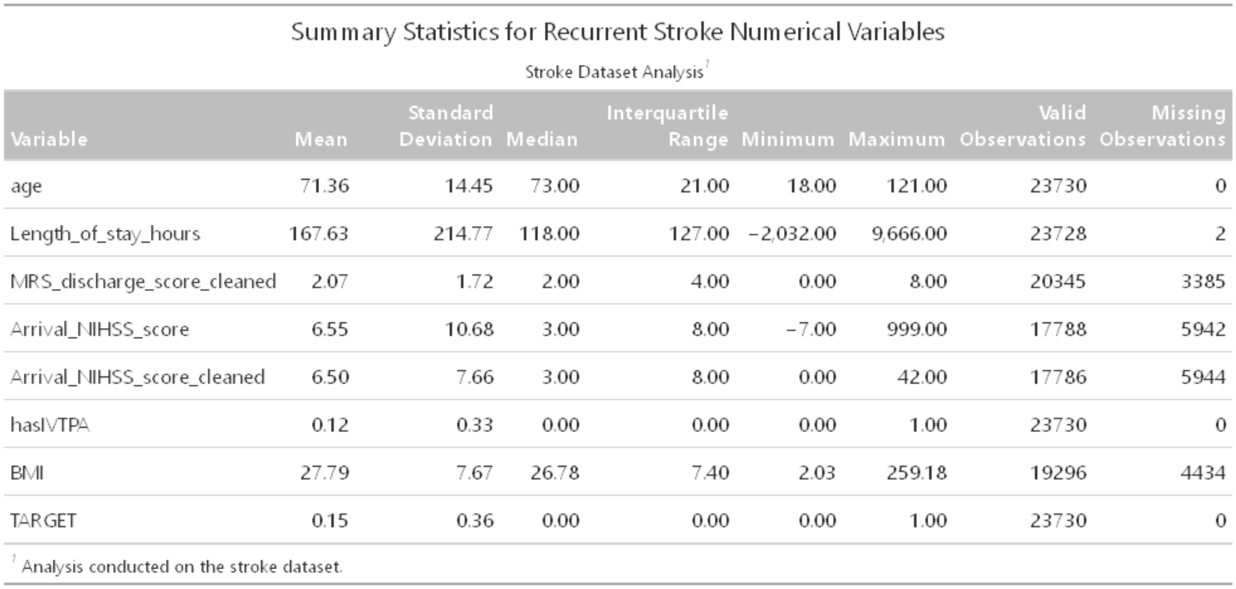
Figures and Tables

Table 1

```r
summary_stats <- stroke_train_quant %>%
  summarise(across(
    everything(),
    list(
      Mean = ~mean(., na.rm = TRUE),
      SD = ~sd(., na.rm = TRUE),
      Median = ~median(., na.rm = TRUE),
      IQR = ~IQR(., na.rm = TRUE),
      Min = ~min(., na.rm = TRUE),
      Max = ~max(., na.rm = TRUE),
      N_Valid = ~sum(!is.na(.)),
      N_Missing = ~sum(is.na(.))
    ),
    .names = "{.col}_{.fn}"  # Constructs names based on variable and function
  ))

# Transform to long format to manage separate statistic columns per variable
summary_stats_long <- pivot_longer(summary_stats, cols = everything(), names_to = "Measure", values_to

# Use regex to properly separate the variable names and statistic types
summary_stats_long <- summary_stats_long %>%
  mutate(
    Variable = sub("_(Mean|SD|Median|IQR|Min|Max|N_Valid|N_Missing)$", "", Measure),
    Statistic = sub(".*_", "", Measure)
  )

# Pivot to wide format for easier readability in gt
summary_stats_wide <- pivot_wider(summary_stats_long, names_from = "Statistic", values_from = "Value",

gt_table <- gt(summary_stats_wide) %>%
  cols_label(
    Variable = "Variable",
    Mean = "Mean",
    SD = "Standard Deviation",
    Median = "Median",
    IQR = "Interquartile Range",
    Min = "Minimum",
    Max = "Maximum",
    Valid = "Valid Observations",
    Missing = "Missing Observations"
  ) %>%
  tab_header(
    title = "Summary Statistics for Recurrent Stroke Numerical Variables",
    subtitle = "Stroke Dataset Analysis"
  ) %>%
  fmt_number(
    columns = vars(Mean, SD, Median, IQR, Min, Max),
    decimals = 2
  ) %>%
  tab_style(
    style = cell_fill(color = "gray"),
    locations = cells_column_labels()
  ) %>%
  tab_style(
    style = cell_text(color = "white", weight = "bold"),
    locations = cells_column_labels()
  ) %>%
  tab_footnote(
    footnote = "Analysis conducted on the stroke dataset.",
    locations = cells_title(groups = "subtitle")
  )
```

```
## Warning: Since gt v0.3.0, `columns = vars(...)` has been deprecated.
## * Please use `columns = c(...)` instead.
## Since gt v0.3.0, `columns = vars(...)` has been deprecated.
## * Please use `columns = c(...)` instead.
```

```
gtsave(gt_table, "stroke_summary_stats.html")

#webshot::install_phantomjs()

webshot("stroke_summary_stats.html",
                 "stroke_summary_stats.png",
                 delay = 2)
```



Summary Statistics for Recurrent Stroke Numerical Variables

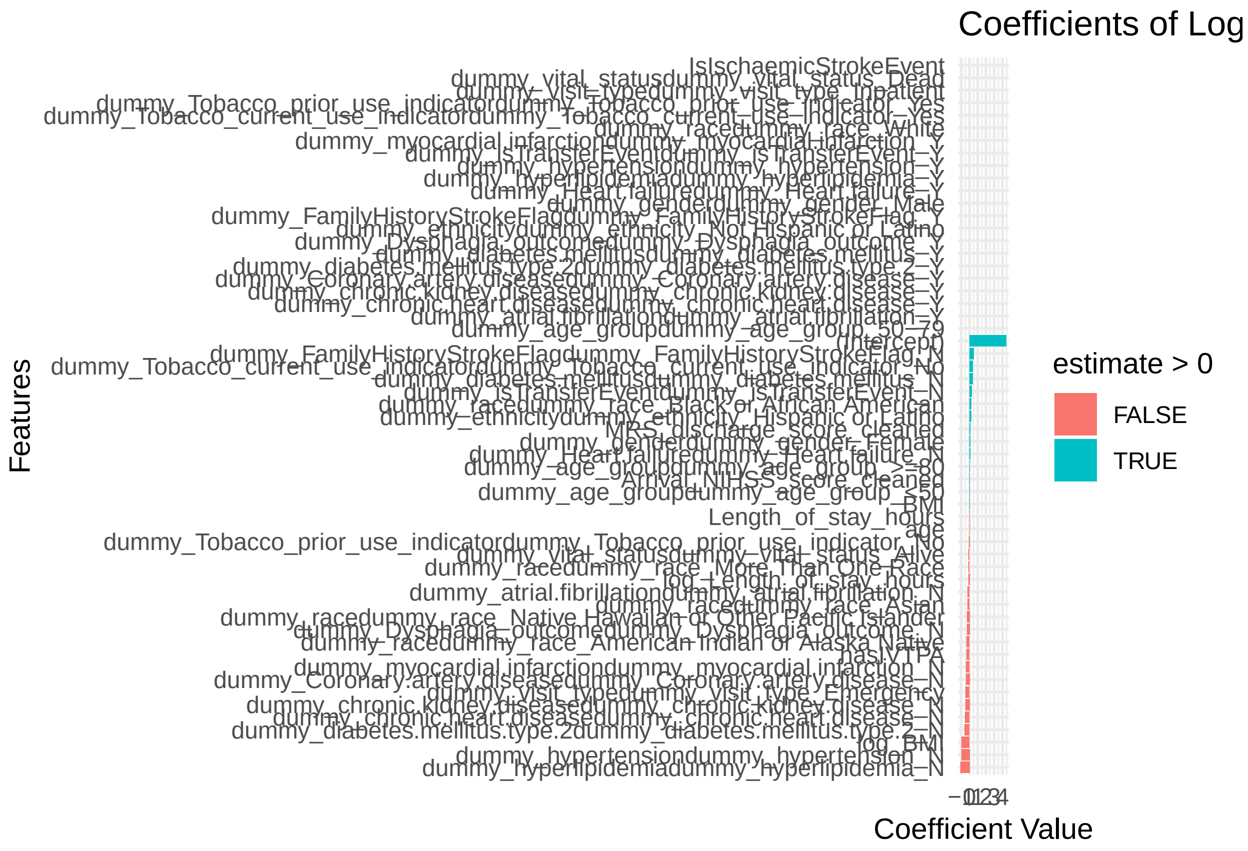
Stroke Dataset Analysis[1]

Variable	Mean	Standard Deviation	Median	Interquartile Range	Minimum	Maximum	Valid Observations	Missing Observations
age	71.36	14.45	73.00	21.00	18.00	121.00	23730	0
Length_of_stay_hours	167.63	214.77	118.00	127.00	−2,032.00	9,666.00	23728	2
MRS_discharge_score_cleaned	2.07	1.72	2.00	4.00	0.00	8.00	20345	3385
Arrival_NIHSS_score	6.55	10.68	3.00	8.00	−7.00	999.00	17788	5942
Arrival_NIHSS_score_cleaned	6.50	7.66	3.00	8.00	0.00	42.00	17786	5944
hasIVTPA	0.12	0.33	0.00	0.00	0.00	1.00	23730	0
BMI	27.79	7.67	26.78	7.40	2.03	259.18	19296	4434
TARGET	0.15	0.36	0.00	0.00	0.00	1.00	23730	0

[1] Analysis conducted on the stroke dataset.

```
coefs <- broom::tidy(model)
ggplot(coefs, aes(x = reorder(term, estimate), y = estimate, fill = estimate > 0)) +
  geom_col() +
  coord_flip() +
  labs(title = "Coefficients of Logistic Regression Model",
       x = "Features",
       y = "Coefficient Value") +
  theme_minimal()
```

```
## Warning: Removed 22 rows containing missing values (`position_stack()`).
```



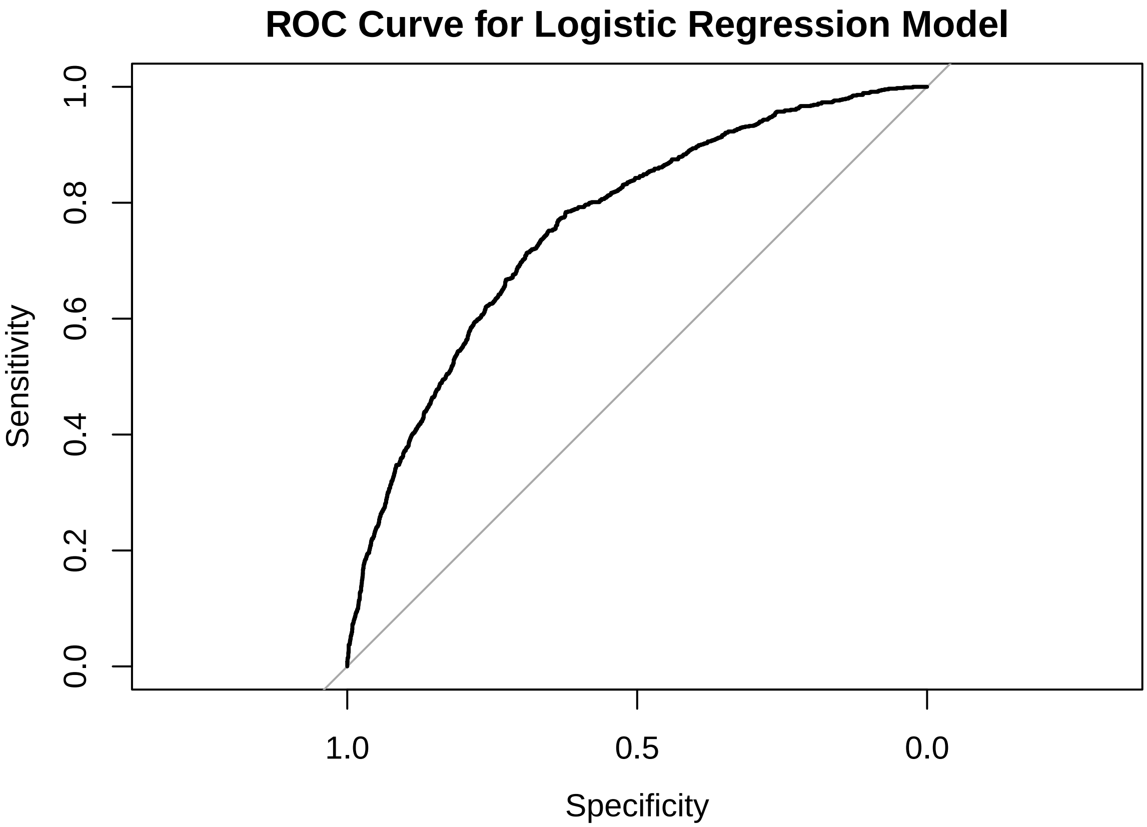
```r
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.3.2

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

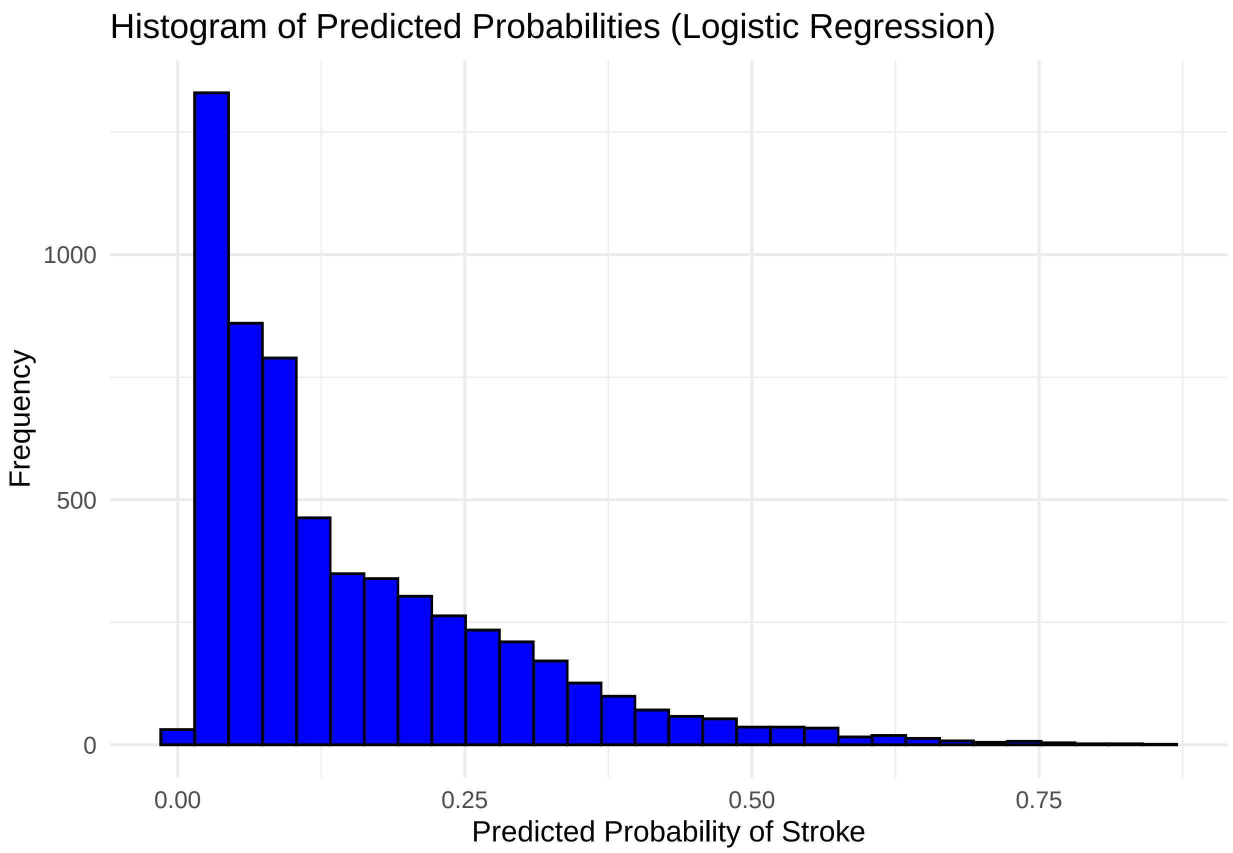
```r
predicted_probs <- predict(model, newdata = stroke_test, type = "response")
roc_curve <- roc(stroke_test$TARGET, predicted_probs)
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```

```
plot(roc_curve, main = "ROC Curve for Logistic Regression Model")
```



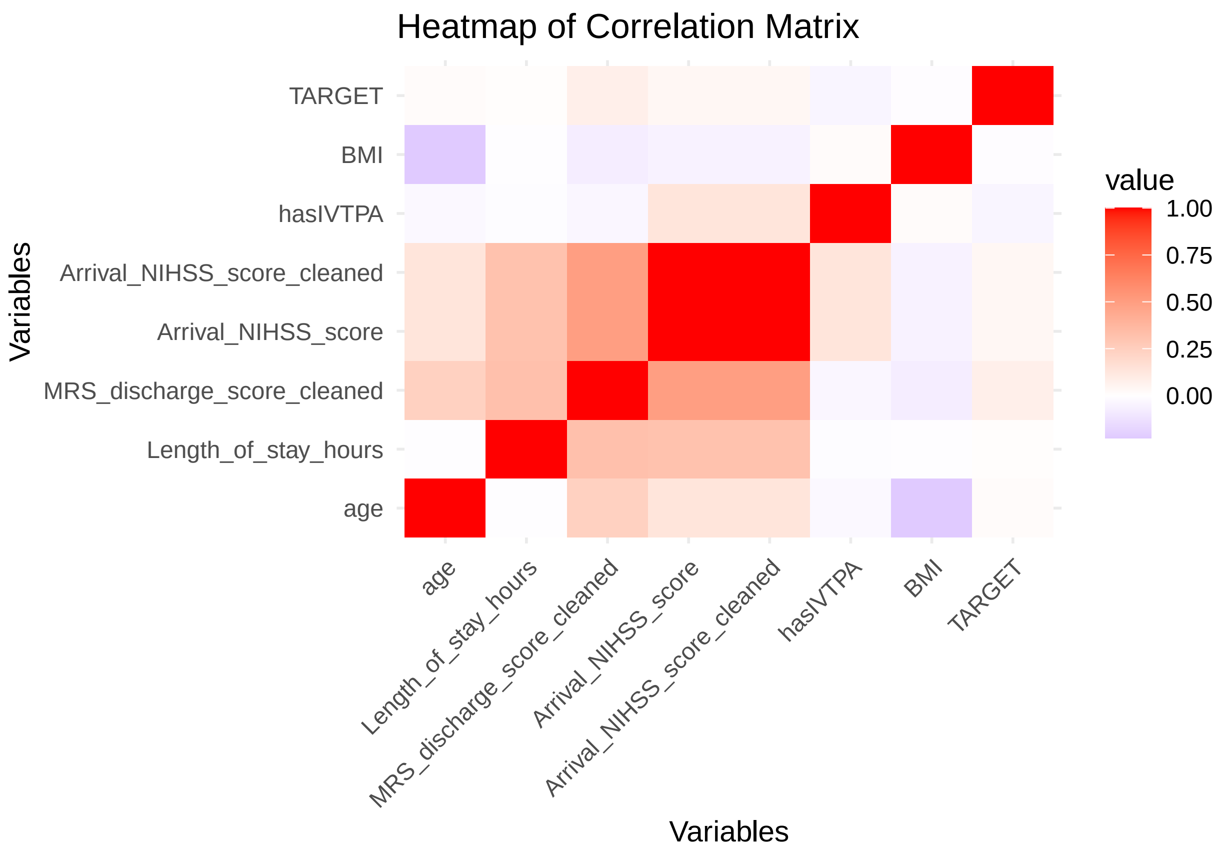
```
predicted_probs <- predict(model, newdata = stroke_test, type = "response")
ggplot(data.frame(Probability = predicted_probs), aes(x = Probability)) +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  ggtitle("Histogram of Predicted Probabilities (Logistic Regression)") +
  xlab("Predicted Probability of Stroke") +
  ylab("Frequency") +
  theme_minimal()
```



```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
cor_matrix <- cor(stroke_train_quant, use = "complete.obs")


melted_cor_matrix <- melt(cor_matrix)


ggplot(melted_cor_matrix, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Heatmap of Correlation Matrix", x = "Variables", y = "Variables")
```



```
mse_data <- data.frame(
  Model = c("Logistic Regression", "Lasso Regression", "Ridge Regression", "Decision Tree"),
  MSE = c(mse_lr, mse_lasso, mse_ridge, mse_tree)
)

mse_table <- gt(mse_data) %>%
  tab_header(
    title = "Model Comparison",
    subtitle = "Mean Squared Errors of Predictive Models"
  ) %>%
  cols_label(
    Model = "Model",
    MSE = "Mean Squared Error"
  ) %>%
  fmt_number(
    columns = vars(MSE),
    decimals = 4
  ) %>%
  tab_style(
    style = list(
      cell_fill(color = "gray"),
      cell_text(color = "white", weight = "bold")
    ),
    locations = cells_column_labels(columns = TRUE)
  )
```

```
## Warning: Since gt v0.3.0, `columns = vars(...)` has been deprecated.
## * Please use `columns = c(...)` instead.
## Since gt v0.3.0, `columns = vars(...)` has been deprecated.
## * Please use `columns = c(...)` instead.
```

```
## Warning: Since gt v0.3.0, `columns = TRUE` has been deprecated.
## * Please use `columns = everything()` instead.
```

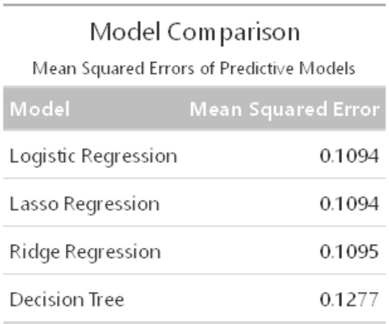
```
# Print or save the table
print(mse_table)
```

```
## <div id="iquzrlmiyx" style="padding-left:0px;padding-right:0px;padding-top:10px;padding-bottom:10px;
##   <style>#iquzrlmiyx table {
##   font-family: system-ui, 'Segoe UI', Roboto, Helvetica, Arial, sans-serif, 'Apple Color Emoji', 'Se
##   -webkit-font-smoothing: antialiased;
##   -moz-osx-font-smoothing: grayscale;
## }
##
## #iquzrlmiyx thead, #iquzrlmiyx tbody, #iquzrlmiyx tfoot, #iquzrlmiyx tr, #iquzrlmiyx td, #iquzrlmiyx
##   border-style: none;
## }
##
## #iquzrlmiyx p {
##   margin: 0;
##   padding: 0;
## }
##
## #iquzrlmiyx .gt_table {
##   display: table;
##   border-collapse: collapse;
##   line-height: normal;
##   margin-left: auto;
##   margin-right: auto;
##   color: #333333;
##   font-size: 16px;
##   font-weight: normal;
##   font-style: normal;
##   background-color: #FFFFFF;
##   width: auto;
##   border-top-style: solid;
##   border-top-width: 2px;
##   border-top-color: #A8A8A8;
##   border-right-style: none;
##   border-right-width: 2px;
##   border-right-color: #D3D3D3;
##   border-bottom-style: solid;
##   border-bottom-width: 2px;
##   border-bottom-color: #A8A8A8;
##   border-left-style: none;
##   border-left-width: 2px;
##   border-left-color: #D3D3D3;
## }
##
## #iquzrlmiyx .gt_caption {
##   padding-top: 4px;
##   padding-bottom: 4px;
## }
##
## #iquzrlmiyx .gt_title {
##   color: #333333;
##   font-size: 125%;
##   font-weight: initial;
##   padding-top: 4px;
##   padding-bottom: 4px;
##   padding-left: 5px;
##   padding-right: 5px;
##   border-bottom-color: #FFFFFF;
##   border-bottom-width: 0;
## }
##
## #iquzrlmiyx .gt_subtitle {
##   color: #333333;
##   font-size: 85%;
##   font-weight: initial;
##   padding-top: 3px;
##   padding-bottom: 5px;
##   padding-left: 5px;
##   padding-right: 5px;
##   border-top-color: #FFFFFF;
##   border-top-width: 0;
## }
##
## #iquzrlmiyx .gt_heading {
##   background-color: #FFFFFF;
##   text-align: center;
##   border-bottom-color: #FFFFFF;
##   border-left-style: none;
##   border-left-width: 1px;
##   border-left-color: #D3D3D3;
##   border-right-style: none;
##   border-right-width: 1px;
##   border-right-color: #D3D3D3;
## }
##
## #iquzrlmiyx .gt_bottom_border {
##   border-bottom-style: solid;
##   border-bottom-width: 2px;
##   border-bottom-color: #D3D3D3;
## }
##
## #iquzrlmiyx .gt_col_headings {
##   border-top-style: solid;
##   border-top-width: 2px;
##   border-top-color: #D3D3D3;
##   border-bottom-style: solid;
##   border-bottom-width: 2px;
##   border-bottom-color: #D3D3D3;
##   border-left-style: none;
##   border-left-width: 1px;
##   border-left-color: #D3D3D3;
##   border-right-style: none;
##   border-right-width: 1px;
##   border-right-color: #D3D3D3;
## }
##
## #iquzrlmiyx .gt_col_heading {
##   color: #333333;
##   background-color: #FFFFFF;
##   font-size: 100%;
##   font-weight: normal;
##   text-transform: inherit;
##   border-left-style: none;
##   border-left-width: 1px;
##   border-left-color: #D3D3D3;
##   border-right-style: none;
##   border-right-width: 1px;
##   border-right-color: #D3D3D3;
##   vertical-align: bottom;
##   padding-top: 5px;
##   padding-bottom: 6px;
##   padding-left: 5px;
##   padding-right: 5px;
##   overflow-x: hidden;
## }
##
## #iquzrlmiyx .gt_column_spanner_outer {
##   color: #333333;
##   background-color: #FFFFFF;
##   font-size: 100%;
##   font-weight: normal;
##   text-transform: inherit;
##   padding-top: 0;
##   padding-bottom: 0;
##   padding-left: 4px;
##   padding-right: 4px;
## }
##
## #iquzrlmiyx .gt_column_spanner_outer:first-child {
##   padding-left: 0;
## }
##
## #iquzrlmiyx .gt_column_spanner_outer:last-child {
##   padding-right: 0;
## }
##
## #iquzrlmiyx .gt_column_spanner {
##   border-bottom-style: solid;
##   border-bottom-width: 2px;
##   border-bottom-color: #D3D3D3;
##   vertical-align: bottom;
##   padding-top: 5px;
##   padding-bottom: 5px;
##   overflow-x: hidden;
##   display: inline-block;
##   width: 100%;
## }
##
## #iquzrlmiyx .gt_spanner_row {
##   border-bottom-style: hidden;
## }
##
## #iquzrlmiyx .gt_group_heading {
##   padding-top: 8px;
##   padding-bottom: 8px;
##   padding-left: 5px;
##   padding-right: 5px;
##   color: #333333;
##   background-color: #FFFFFF;
##   font-size: 100%;
##   font-weight: initial;
##   text-transform: inherit;
##   border-top-style: solid;
##   border-top-width: 2px;
##   border-top-color: #D3D3D3;
##   border-bottom-style: solid;
##   border-bottom-width: 2px;
##   border-bottom-color: #D3D3D3;
##   border-left-style: none;
##   border-left-width: 1px;
##   border-left-color: #D3D3D3;
##   border-right-style: none;
##   border-right-width: 1px;
##   border-right-color: #D3D3D3;
##   vertical-align: middle;
##   text-align: left;
## }
##
## #iquzrlmiyx .gt_empty_group_heading {
##   padding: 0.5px;
##   color: #333333;
##   background-color: #FFFFFF;
##   font-size: 100%;
##   font-weight: initial;
##   border-top-style: solid;
##   border-top-width: 2px;
##   border-top-color: #D3D3D3;
##   border-bottom-style: solid;
##   border-bottom-width: 2px;
##   border-bottom-color: #D3D3D3;
##   vertical-align: middle;
## }
##
## #iquzrlmiyx .gt_from_md > :first-child {
##   margin-top: 0;
## }
##
## #iquzrlmiyx .gt_from_md > :last-child {
```

```
##   margin-bottom: 0;
## }
##
## #iquzrlmiyx .gt_row {
##   padding-top: 8px;
##   padding-bottom: 8px;
##   padding-left: 5px;
##   padding-right: 5px;
##   margin: 10px;
##   border-top-style: solid;
##   border-top-width: 1px;
##   border-top-color: #D3D3D3;
##   border-left-style: none;
##   border-left-width: 1px;
##   border-left-color: #D3D3D3;
##   border-right-style: none;
##   border-right-width: 1px;
##   border-right-color: #D3D3D3;
##   vertical-align: middle;
##   overflow-x: hidden;
## }
##
## #iquzrlmiyx .gt_stub {
##   color: #333333;
##   background-color: #FFFFFF;
##   font-size: 100%;
##   font-weight: initial;
##   text-transform: inherit;
##   border-right-style: solid;
##   border-right-width: 2px;
##   border-right-color: #D3D3D3;
##   padding-left: 5px;
##   padding-right: 5px;
## }
##
## #iquzrlmiyx .gt_stub_row_group {
##   color: #333333;
##   background-color: #FFFFFF;
##   font-size: 100%;
##   font-weight: initial;
##   text-transform: inherit;
##   border-right-style: solid;
##   border-right-width: 2px;
##   border-right-color: #D3D3D3;
##   padding-left: 5px;
##   padding-right: 5px;
##   vertical-align: top;
## }
##
## #iquzrlmiyx .gt_row_group_first td {
##   border-top-width: 2px;
## }
##
## #iquzrlmiyx .gt_row_group_first th {
```

```
##   border-top-width: 2px;
## }
##
## #iquzrlmiyx .gt_summary_row {
##   color: #333333;
##   background-color: #FFFFFF;
##   text-transform: inherit;
##   padding-top: 8px;
##   padding-bottom: 8px;
##   padding-left: 5px;
##   padding-right: 5px;
## }
##
## #iquzrlmiyx .gt_first_summary_row {
##   border-top-style: solid;
##   border-top-color: #D3D3D3;
## }
##
## #iquzrlmiyx .gt_first_summary_row.thick {
##   border-top-width: 2px;
## }
##
## #iquzrlmiyx .gt_last_summary_row {
##   padding-top: 8px;
##   padding-bottom: 8px;
##   padding-left: 5px;
##   padding-right: 5px;
##   border-bottom-style: solid;
##   border-bottom-width: 2px;
##   border-bottom-color: #D3D3D3;
## }
##
## #iquzrlmiyx .gt_grand_summary_row {
##   color: #333333;
##   background-color: #FFFFFF;
##   text-transform: inherit;
##   padding-top: 8px;
##   padding-bottom: 8px;
##   padding-left: 5px;
##   padding-right: 5px;
## }
##
## #iquzrlmiyx .gt_first_grand_summary_row {
##   padding-top: 8px;
##   padding-bottom: 8px;
##   padding-left: 5px;
##   padding-right: 5px;
##   border-top-style: double;
##   border-top-width: 6px;
##   border-top-color: #D3D3D3;
## }
##
## #iquzrlmiyx .gt_last_grand_summary_row_top {
##   padding-top: 8px;
##   padding-bottom: 8px;
##   padding-left: 5px;
##   padding-right: 5px;
##   border-bottom-style: double;
##   border-bottom-width: 6px;
##   border-bottom-color: #D3D3D3;
## }
##
## #iquzrlmiyx .gt_striped {
##   background-color: rgba(128, 128, 128, 0.05);
## }
##
## #iquzrlmiyx .gt_table_body {
##   border-top-style: solid;
##   border-top-width: 2px;
##   border-top-color: #D3D3D3;
##   border-bottom-style: solid;
##   border-bottom-width: 2px;
##   border-bottom-color: #D3D3D3;
## }
##
## #iquzrlmiyx .gt_footnotes {
##   color: #333333;
##   background-color: #FFFFFF;
##   border-bottom-style: none;
##   border-bottom-width: 2px;
##   border-bottom-color: #D3D3D3;
##   border-left-style: none;
##   border-left-width: 2px;
##   border-left-color: #D3D3D3;
##   border-right-style: none;
##   border-right-width: 2px;
##   border-right-color: #D3D3D3;
## }
##
## #iquzrlmiyx .gt_footnote {
##   margin: 0px;
##   font-size: 90%;
##   padding-top: 4px;
##   padding-bottom: 4px;
##   padding-left: 5px;
##   padding-right: 5px;
## }
##
## #iquzrlmiyx .gt_sourcenotes {
##   color: #333333;
##   background-color: #FFFFFF;
##   border-bottom-style: none;
##   border-bottom-width: 2px;
##   border-bottom-color: #D3D3D3;
##   border-left-style: none;
##   border-left-width: 2px;
##   border-left-color: #D3D3D3;
##   border-right-style: none;
##   border-right-width: 2px;
##   border-right-color: #D3D3D3;
## }
##
## #iquzrlmiyx .gt_sourcenote {
##   font-size: 90%;
##   padding-top: 4px;
##   padding-bottom: 4px;
##   padding-left: 5px;
##   padding-right: 5px;
## }
##
## #iquzrlmiyx .gt_left {
##   text-align: left;
## }
##
## #iquzrlmiyx .gt_center {
##   text-align: center;
## }
##
## #iquzrlmiyx .gt_right {
##   text-align: right;
##   font-variant-numeric: tabular-nums;
## }
##
## #iquzrlmiyx .gt_font_normal {
##   font-weight: normal;
## }
##
## #iquzrlmiyx .gt_font_bold {
##   font-weight: bold;
## }
##
## #iquzrlmiyx .gt_font_italic {
##   font-style: italic;
## }
##
## #iquzrlmiyx .gt_super {
##   font-size: 65%;
## }
##
## #iquzrlmiyx .gt_footnote_marks {
##   font-size: 75%;
##   vertical-align: 0.4em;
##   position: initial;
## }
##
## #iquzrlmiyx .gt_asterisk {
##   font-size: 100%;
##   vertical-align: 0;
## }
##
## #iquzrlmiyx .gt_indent_1 {
##   text-indent: 5px;
## }
##
## #iquzrlmiyx .gt_indent_2 {
##   text-indent: 10px;
## }
##
## #iquzrlmiyx .gt_indent_3 {
##   text-indent: 15px;
## }
##
## #iquzrlmiyx .gt_indent_4 {
##   text-indent: 20px;
## }
##
## #iquzrlmiyx .gt_indent_5 {
##   text-indent: 25px;
## }
## </style>
##   <table class="gt_table" data-quarto-disable-processing="false" data-quarto-bootstrap="false">
##   <thead>
##     <tr class="gt_heading">
##       <td colspan="2" class="gt_heading gt_title gt_font_normal" style>Model Comparison</td>
##     </tr>
##     <tr class="gt_heading">
##       <td colspan="2" class="gt_heading gt_subtitle gt_font_normal gt_bottom_border" style>Mean Squa
##     </tr>
##     <tr class="gt_col_headings">
##       <th class="gt_col_heading gt_columns_bottom_border gt_left" rowspan="1" colspan="1" style="bac
##       <th class="gt_col_heading gt_columns_bottom_border gt_right" rowspan="1" colspan="1" style="ba
##     </tr>
##   </thead>
##   <tbody class="gt_table_body">
##     <tr><td headers="Model" class="gt_row gt_left">Logistic Regression</td>
## <td headers="MSE" class="gt_row gt_right">0.1094</td></tr>
##     <tr><td headers="Model" class="gt_row gt_left">Lasso Regression</td>
## <td headers="MSE" class="gt_row gt_right">0.1094</td></tr>
##     <tr><td headers="Model" class="gt_row gt_left">Ridge Regression</td>
## <td headers="MSE" class="gt_row gt_right">0.1095</td></tr>
##     <tr><td headers="Model" class="gt_row gt_left">Decision Tree</td>
## <td headers="MSE" class="gt_row gt_right">0.1277</td></tr>
##   </tbody>
##
##
## </table>
## </div>
```

```r
# Optional: Save the table as HTML or PNG
gtsave(mse_table, "model_mse_comparison.html")
webshot("model_mse_comparison.html", "model_mse_comparison.png", delay = 2)
```



Model Comparison

Mean Squared Errors of Predictive Models

Model	Mean Squared Error
Logistic Regression	0.1094
Lasso Regression	0.1094
Ridge Regression	0.1095
Decision Tree	0.1277