# 546 Final Project

Alison King

2023-03-09

## Abstract

**Background:** The aim of this analysis was to create a predictive model to classify patients into Alzheimer's Disease and healthy patients by utilizing cerebral cortex measurements as predictors. We had 360 measurements from various areas of the brain for each patient. For this analysis, we wanted to identify those predictors that were most important for predicting the Alzheimer's outcome and use these predictors to fit a predictive logistic regression model.

**Methods:** In order to find those predictors that were most important, feature selection was performed via forward stepwise selection with 10-fold cross-validation. The best model size for each CV method was chosen based on the lowest cross-validation accuracy rate. The two best models were then evaluated on the training set and used to generate predictions for the test set. We then compared the test accuracy of both models and chose the model from the cross-validation method with the best test accuracy to be our final model.

**Findings:** The forward stepwise selection cross-validation method that resulted in the best test accuracy was 10-fold cross-validation. The final model identified via forward stepwise logistic regression with 10-fold CV was that with 23 predictors. The training accuracy of this model was 96.75%, and the test accuracy was 91.75%. The model misclassifications also tended towards false positives over false negatives, so we would expect to not miss many Alzheimer's diagnoses. These results suggest that our model had good predictive abilities for using cerebral cortex measurements to classify patients into Alzheimer's and healthy patients.

## Introduction

### Goal and Data

The goal of this project was to classify adults into Alzheimer's (AD) and healthy patients (C) based on cerebral cortex thickness measurements from 360 brain regions of interest. We had data from 400 patients on which to train our model, and a test set of 400 patients on which to evaluate the accuracy of the model.

### Approach

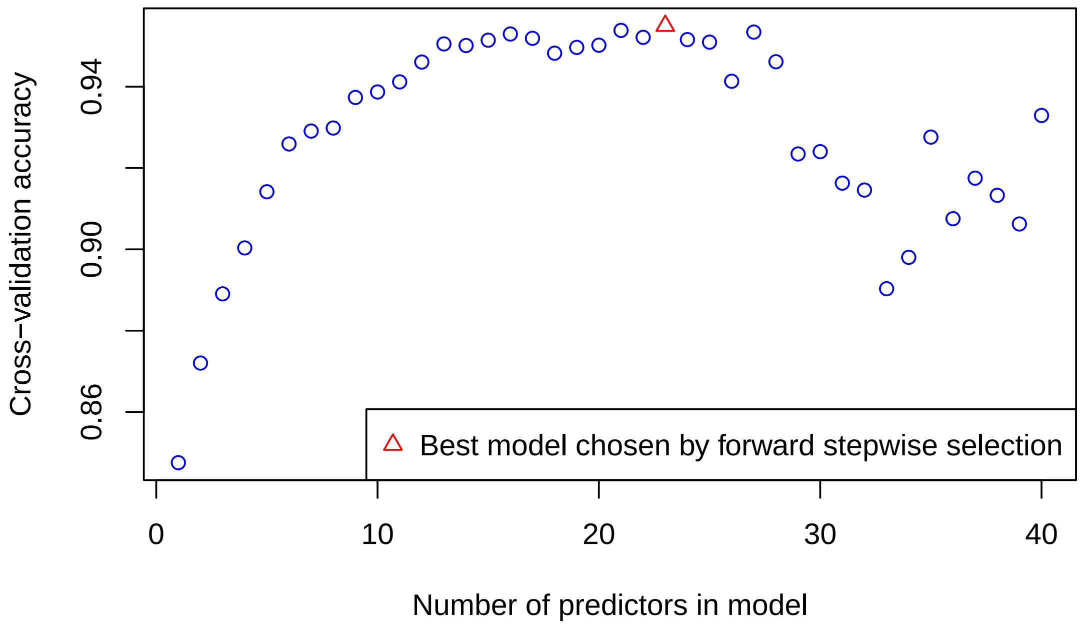
For this project, we implemented forward stepwise selection with logistic regression in order to select the predictors that were most associated with the outcome and train the classification model using these predictors. This method was chosen because logistic regression models produce more easily interpretable results as compared to other machine learning methods. Therefore, we were able to understand how certain features are associated with the disease outcome. The best model from forward stepwise selection was chosen via 10-fold cross validation. Once the best model was chosen and trained with the training data, we generated the 400 predicted classifications of the patients in the blind test set and submitted them for evaluation of test prediction accuracy.

## Analysis

### Feature Selection

First, we used forward stepwise selection to select the most important features from the possible 360 predictors in the training data. In order to do so, we implemented the *regsubsets* function in R on the training data to fit logistic regression models of size 1 to 40 predictors. The max number of variables, 40, was chosen because the cross-validation accuracy peaked before this point and was beginning to worsen with more added predictors. After fitting the 40 potential models, we used 10-fold cross-validation to estimate the cross-validation prediction accuracy of each model on the training set. The CV accuracies for each model are depicted in Figure 1. We then chose the model with the best CV accuracy rate (95.5%), and this was the model with 23 predictors.

**Figure 1: Model size vs. accuracy**



### The Best Model

We re-fit the logistic regression model with 23 predictors selected to be the best model from forward stepwise selection with 10-fold cross-validation on the entire training set. We produced the following model (where $P_X$ is the cerebral cortex thickness measurement for the X region of the brain):

$$\begin{aligned} Pr[AD] = exp(&23.5 - 3.3 * P_{11} - 1.9 * P_{21} + 1.2 * P_{37} + 2.4 * P_{42} + 2.7 * P_{52} \\ &+1.8 * P_{60} - 1.9 * P_{105} - 3.1 * P_{106} - 3.3 * P_{124} - 1.7 * P_{139} - 1.8 * P_{194} \\ &-2.8 * P_{208} - 2.6 * P_{209} + 3.1 * P_{211} + 3.5 * P_{224} - 0.7 * P_{226} + 2.0 * P_{227} \\ &+4.8 * P_{231} - 4.5 * P_{263} + 3.4 * P_{279} - 1.5 * P_{285} - 5.1 * P_{314} + 3.3 * P_{357}) \end{aligned}$$

From this model, if the probability of AD was greater than 0.5, we classified that prediction as “Alzheimer’s”. If the probability of AD was less than 0.5, we classified that prediction as “Control”.

# Results

### Training Accuracy

The prediction accuracy rate of the model on the training set was 96.75%. We can see from the confusion matrix (Table 1) that the model classifies the training data fairly well, with only 3 false negatives and 10 false positives out of 400 patients. In a medical setting, we tend to prefer false positives over false negatives, as we would not want to miss any potentially dangerous diagnoses. Thus, we are satisfied with the classification results.

Table 1: Confusion Matrix for Model on Training Set

| | AD | C |
|---|---:|---:|
| C | 10 | 87 |
| AD | 300 | 3 |

### Test Accuracy

We used the fitted model to predict the classifications of the 400 patients in the test set and submitted them to the blinded predictions evaluation. The prediction accuracy rate on the test data was returned as 91.75%.

# Conclusions

The goal of this project was to create a machine learning model that can classify patients with Alzheimer's Disease versus healthy elderly by analyzing the thickness measurements from their cerebral cortex. Our forward stepwise logistic regression model performed feature selection and 10-fold cross validation to choose the best predictive model with a resulting test accuracy of 91.75% on a blinded test set. This is a good test accuracy, and thus we believe that our model does a good job of classifying patients into Alzheimer's and healthy patients based on their cerebral cortex measurements. We believe that the cerebral cortex measurements in the regions identified by our model are good predictors for Alzheimer's Disease.

## Code Appendix

```r
knitr::opts_chunk$set(echo = FALSE)
load("~/Downloads/ADProj.Rdata")
X_train <- ADProj$X_train
y_train <- ADProj$y_train
X_test <- ADProj$X_test
library(leaps)
library(tidyverse)
library(boot)
# make data frame of training set
train.data <- as.data.frame(cbind(X_train,y_train))
relevel(train.data$Outcome, ref="C")

# forward stepwise with logistic regression
# with leaps::regsubsets
forward <- regsubsets(as.factor(Outcome) ~ .,
                      data=train.data, nvmax = 40, method = "forward")

#get the significant predictor ids
get_model_formula <- function(id, object, outcome){
  # get models data
  models <- summary(object)$which[id,-1]
  # Get model predictors
  predictors <- names(which(models == TRUE))
  predictors <- paste(predictors, collapse = "+")
  # Build model formula
  as.formula(paste0(outcome, "~", predictors))
}

# 10-fold cross-validation to find best model
set.seed(55)
acc_vec2 <- rep(NA,40)
for (i in 1:length(acc_vec2)){
  my_glm_mod = glm(get_model_formula(i,forward,"Outcome"),
                   family="binomial", data = train.data)
  cv.err = cv.glm(train.data, my_glm_mod, K = 10)
  acc_vec2[i]=  1-cv.err$delta[1]
}

maxacc <- max(acc_vec2)
max_ind <- which(acc_vec2==maxacc)

# graph of model size versus CV accuracy
```

```r
mod.size <- seq(1,40,1)
plot(mod.size, acc_vec2, col=ifelse(mod.size==max_ind, 'red', 'blue'),
     pch=ifelse(mod.size==max_ind,2,1),
     xlab="Number of predictors in model",
     ylab="Cross-validation accuracy",
     main="Figure 1: Model size vs. accuracy")
legend("bottomright",
       "Best model chosen by forward stepwise selection",
       col="red", pch=2)
# pick best model with highest prediction accuracy
best.mod <- get_model_formula(max_ind,forward,"Outcome")

# re-fit the logistic regression model on the
# training set with the chosen predictors
glm.mod <- glm(formula=best.mod, data=train.data, family="binomial")

# calculate accuracy rate of the training set
train.prob <- predict(glm.mod, type = "response")
train.label <- as.factor(ifelse(train.prob < .5, "C", "AD"))
train.acc <- mean(train.label == train.data$Outcome)

#confusion matrix
cm.train = table(True = train.data$Outcome, Predicted = train.label)
knitr::kable(cm.train, caption="Confusion Matrix for Model on Training Set")
# predictions using best model (for blind test)
test.prob <- predict(glm.mod, type = "response", newdata=X_test)
test.label <- ifelse(test.prob > .5, "AD", "C")
test.label <- as.data.frame(test.label)

write.table(test.label, file = "march1predictions.txt",
            row.names = FALSE, col.names = FALSE)
```