<https://stackoverflow.com/questions/27303051/r-estimating-model-variance>

<https://www.youtube.com/watch?v=c-kqw0Yf6BE>

<https://algorithmia.com/blog/evaluating-machine-learning-models-with-a-confusion-matrix>

<https://stackoverflow.com/questions/24801452/error-in-confusionmatrix-the-data-and-reference-factors-must-have-the-same-numbe>

<https://community.rstudio.com/t/error-data-and-reference-should-be-factors-with-the-same-levels/67965/4>

how to find accuracy of logistic regression in r

<https://rpubs.com/jpmurillo/153750>

<https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/logistic-regression-analysis-r/tutorial/>

<https://stats.stackexchange.com/questions/65244/how-to-determine-the-accuracy-of-logistic-regression-in-r>

<https://www.journaldev.com/46732/confusion-matrix-in-r>

#REGULARIZED REGRESSION: ELASTIC NET

set.seed(123)

en.model <- train(

alc\_consumption ~., data = train.data, method = "glmnet",

trControl = trainControl("cv", number = 10), preProc = c("center", "scale"), tuneLength = 10

)

#Print the values of alpha and lambda that gave best prediction

en.model$bestTune %>% knitr::kable() # 0.6(alpha)| 0.2593237(lambda)| 0.8538419|(accuracy)

#Print all of the options examined. Bc this is a logistic regression we are using the Accuracy. If it was linear regression it would be MSE/RMSE.

en.model$results %>% knitr::kable()

# Model coefficients

coef(en.model$finalModel, en.model$bestTune$lambda)

#Confusion Matrix

confusionMatrix(en.model)

#LOGISTIC REGRESSION

logistic\_control1 <- trainControl(method = "cv", number = 3, savePredictions = T)

set.seed(1000)

logistic <- train(alc\_consumption ~ ., data = alc, method = "glm", family = "binomial", trControl = logistic\_control1)

summary(logistic)

confusionMatrix(logistic)

confusionMatrix(table((logistic$pred)$pred,(logistic$pred)$obs))

In the logistic regression model the average accuracy was 0.8085 and the sensitivity and specificity was 0.8247 and 0.7900, respectively.

In the elastic net model, the average accuracy was 0.8538. The intercept was -0.1364 and the remaining variables went to zero, except for Measure of Impulsivity (impulsiveness\_score = -0.4190) and Measure of Sensation-Seeking Behaviors (sens\_seeking\_score=-0.00096). The best predicting alpha and lambda was 0.6(alpha) and 0.2593(lambda), which resulted in an accuracy of 0.8538.

In the logistic regression model the average accuracy was 0.8015 and the sensitivity and specificity was 0.8193 and 0.7812, respectively.

In the LASSO model, the average accuracy was 0.8538.The intercept was -0.1332 and the remaining variables went to zero, except for Measure of Impulsivity (impulsiveness\_score = -0.2805).The best prediciting alpha and lambda was 1(alpha) and 0.2310(lambda), which resulted in an accuracy of 0.853818.

#Q2

<https://cran.r-project.org/web/packages/predtools/vignettes/calibPlot.html>

train.data$pred <- predict.glm(logistic, type = 'response')

test.data$pred <- predict.glm(logistic, newdata = test.data, type = 'response')

calibration\_plot(data = train.data, obs = logistic, pred = "pred", title = "Calibration plot for development data")

calibration\_plot(data = test.data, obs = "y", pred = "pred", title = "Calibration plot for development data")

train.data$pred <- predict.glm(logistic, type = 'response')

calibration\_plot(data = train.data, obs = "alc\_consumption", pred = "pred", title = "Calibration plot for development data")

test.data$pred <- predict.glm(logistic, newdata = test.data, type = 'response')

calibration\_plot(data = test.data, obs = "y", pred = "pred", title = "Calibration plot for development data")

<http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/#:~:text=Logistic%20regression%20is%20used%20to,multiple%20predictor%20variables%20(x).&text=Logistic%20regression%20belongs%20to%20a,regression))%20to%20other%20situations>.

fitted.results\_model1 <- model1 %

testProbs <- data.frame(obs = test.data$alc\_consumption,

pred.logit = fitted.results\_model1)

calPlotData\_model1<- calibration(obs~as.numeric(pred.logit), data = testProbs)

xyplot(calPlotData\_model1, auto.key = list(columns = 2))

# Q3: Apply your final model in the test set and report your final evaluation metrics

```{r}

# Make predictions in test set

en.pred <- en.model %>% predict(test.data)

en.pred2 = na.omit(en.pred)

confusionMatrix(en.pred2)

#NTS:Caret doesn't work if your binary response is of character/logical so you have to rename the levels as factors.

levels <- levels(en.pred)

levels <- levels[order(levels)]

table(ordered(en.pred,levels), ordered(test.data$alc\_consumption, levels))

# Model prediction performance

confusionMatrix(table(ordered(en.pred,levels), ordered(test.data$alc\_consumption, levels)))

# Q3: Apply your final model in the test set and report your final evaluation metrics

```{r}

#Using the test data to make predictions

en\_test\_data <- train(

alc\_consumption ~., data = test.data, method = "glmnet",

trControl = trainControl("cv", number = 10), preProc = c("center", "scale"), tuneLength = 10

)

#Print the values of alpha and lambda that gave best prediction

en\_test\_data$bestTune %>% knitr::kable() # 0.6(alpha)| 0.257764(lambda)| 0.8495538|(accuracy)

#Print all of the options examined. Bc this is a logistic regression we are using the Accuracy. If it was linear regression it would be MSE/RMSE.

en\_test\_data$results %>% knitr::kable()

# Model coefficients

coef(en\_test\_data$finalModel, en\_test\_data$bestTune$lambda)

#Confusion Matrix

confusionMatrix(en\_test\_data)

# Make predictions in test set

en\_pred <- en\_test\_data %>% predict(test.data)

confusionMatrix(en\_pred)

Q5:

an online survey related to drug and alcohol use and personality traits. Individuals answered standardized questions which were used to calculate continuous scores on personality traits. Individuals were also asked about consumption of alcohol and multiple drugs.

Below is a list of the 7 features and outcome variable within the dataset. Note the dataset also contains an ID variable. In general, the higher value of the score, the greater the personality trait observed within the individual based on the questionnaire.

1. alc\_consumption: CurrentUse, NotCurrentUse
2. neurotocism\_score: Measure of Neuroticism
3. extroversion\_score: Measure of Extroversion
4. openness\_score: Measure of Openness to Experiences
5. agreeableness\_score: Measure of Agreeableness
6. conscientiousness\_score: Measure of Conscientiousness
7. impulsiveness\_score: Measure of Impulsivity
8. sens\_seeking\_score: Measure of Sensation-Seeking Behaviors.

What research questions could this analysis either a) directly address or b) indirectly help to address by providing information that could be used in subsequent analyses? Limit this response to no more than 1 paragraph. Be sure to use complete sentences.

Directly address: Does an individual’s measure of impulsiveness affect their current use of alcohol?

Indirectly: Does an individual’s sensation-seeking behaviors related to the rate of drunk driving incidents amongst young adults (18-25 years old)?

Are sensation-seeking behaviors amongst NYC young adults (18-25 years old) related to the rate of drunk driving deaths in NYC?

This research can be used for a plethora of research questions. This research can be used to directly address: does an individual’s measure of impulsiveness affect their current use of alcohol? This research can be used to subsequently address