Project 2 Emma Taylor

11571537

1. Introduction

The central topic of this project is logistic regression and its application to classification questions. We are tasked with using the SPAM dataset to build a LASSO penalized logistic regression model that can determine whether or not an email should be classified as spam based on the presence of spam key words and characters and the length of strings of capitalized characters.

A LASSO penalized logistic regression model is preferred here for both the LASSO method and the logistic regression elements of the model. First, a logistic regression model is necessary here because we want to predict a qualitative response. In most cases, linear regression can only be used to predict quantitative responses. Logistic regression uses the logistic function to determine probabilities that lead to qualitative classification. In the case of this project with a binary choice for “spam” or “not spam”, this is also possible with linear regression, but logistic regression has the added benefit of forcing every probability to be between 0 and 1, the range which is physically possible for probabilities.

The LASSO penalty is useful in this model because we are dealing with many variables that each have very small coefficients and are not all significant in a standard logistic regression model, meaning it could improve the accuracy of our model’s predictions by removing unnecessary or even harmful variables. As discussed in other projects and homeworks, the LASSO method offers reduced variance for a slightly higher bias that is often a beneficial trade-off.

Penalized logistic regression is a useful skill in modern data analysis, which statisticians Shofiyah and Sofro demonstrate in their conference presentation entitled *Split and Conquer Method in Penalized Logistic Regression with Lasso (Application on Credit Scoring Data).* Like the textbook’s demonstrations with credit, the authors showed how a special technique within penalized logistic regression can produce very accurate classification predictions with a smaller computational footprint than other Big Data techniques.

1. Methods
2. Tasks

The first task of this project was to remove the missing values from the data, which was accomplished by the “na.omit” command that was introduced in the Chapter 6 Lab on Best Subset Selection. Next, we needed to decide if we wanted to include all three capitalization string length variables, “crl.tot”, “crl.ave”, and “crl.long”. The possible problem with including all three of these variables would be multicollinearity, meaning the model would be overly optimistic due to the fact that those three variables are likely to be highly dependent on one another. In order to see if collinearity would be a problem, I created a correlation matrix and found that each capitalization string length variable was at least 40% correlated to the other capitalization string length variables, which was a much higher value than any other pairwise correlation. Because of those high values, I decided to omit “crl.tot” and “crl.long”, preferring to just use “crl.ave”.

With the data processing task completed, I moved on to creating the LASSO penalized logistic regression model. I began by creating random samples for a training set and a test set, of sizes 1500 observations and 300 observations, respectively. Then, in order to begin creating a LASSO penalized logistic regression model, I reformatted my training and testing sets using the ‘model.matrix’ command and built a model that incorporated every feature/variable except “testid”, “crl.tot”, and “crl.long”. Then, I used the “cv.glmnet” cross-validation command with the logistic regression options to find a best lambda for my LASSO penalized model. Next, I applied the best lambda to my model and looked at the remaining coefficients and their values. After that, I finally applied my test data to my model and obtained the predicted probabilities from the model. Then, in accordance with the procedure in the assignment description, I assigned all predictions with value >0.5 as spam, leaving the rest to be classified as non-spam. I then produced a 2 by 2 confusion table showing the accurate predictions and the mistakes made by the model and used the “high-dimensional inference” package to check the significance of the model.

b. Code

SPAM <- read.csv("C:/Users/emmat/Downloads/SPAM.csv", header = TRUE, na.strings=c(""))

SPAM=na.omit(SPAM)

attach(SPAM)

options(max.print = 20000)

cor(SPAM)

set.seed(123)

train=sample(SPAM$spam,1500,replace=FALSE)

train.RData=SPAM[train,]

set.seed(123)

test=sample(SPAM$spam,300,replace=FALSE)

test.RData=SPAM[test,]

library(glmnet)

x=model.matrix(spam~.-testid -crl.tot -crl.long,data=train.RData)

testx=model.matrix(spam~.-testid -crl.tot -crl.long,data=test.RData)

y=train.RData$spam

testy=test.RData$spam

set.seed(123)

cvfit = cv.glmnet(x,y,family = "binomial", type.measure = "class")

plot(cvfit)

cvfit$lambda.min

coefEst = coef(cvfit, s = "lambda.min")

coefEst = as.vector(coefEst)

coefEst[coefEst!=0]

ProbPrd = predict(cvfit, newx = testx, s = "lambda.min", type = "response")

as.vector(ProbPrd)

PredClass = as.numeric(ProbPrd >0.5)

table(PredClass,testy)

ClassPrd = predict(cvfit, newx = testx, s = "lambda.min", type = "class")

table(as.vector(ClassPrd),testy)

library(hdi)

outRes=lasso.proj(x, y, family = "binomial", standardize = TRUE, multiplecorr.method = "none")

outRes$pval

1. Results

I find that the LASSO penalized logistic regression model is able to predict if emails are spam or not with minimal type I and type II error. The table below is the confusion table produced by my code that demonstrates the accuracy of the model:

Text

Description automatically generated

1. Discussion

I think that further improvement of this model could be possible with more advanced and updated data collection. This data set is from 1999, and spam has changed in its contents as well as its sophistication since then. In order to build a very accurate spam filter, one would need to collect modern data, which in my opinion should also include information about sender address characteristics. Furthermore, as the spam emails that are sent to WSU addresses show, some types of spam emails do follow a very specific formula that is not accounted for in the features of this model. It would be extremely beneficial if WSU could build a spam filter for emails that target WSU students that often contain phishing scams!

1. References

F Shofiyah and A Sofro. *Split and Conquer Method in Penalized Logistic Regression with Lasso (Application on Credit Scoring Data).* IOP Conf. Series: Journal of Physics: Conf. Series 1108 (2018) 012107

Chapters 5 and 6 of *An Introduction to Statistical Learning: with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani