MSBA7002 Business Statistics - HW 2

Name: _____ Student ID: _____

9 November 2024

Overview / Instructions

This homework will be *due on 9 November 2024 by 11:55 PM* via Moodle.

You are required to submit 1) an original R Markdown file and 2) a knitted HTML or PDF file. Please provide comments for the R code wherever you see appropriate. Nice formatting of the assignment will have extra points.

In general, be as concise as possible while giving a fully complete answer. All necessary data are available in Moodle.

Remember that the Class Policy strictly applies to homework. You are encouraged to work in groups and discuss with fellow students. However, each student has to know how to answer the questions on her/his own. Note that the final exam is individual based.

Question 0

Review the lectures.

Question 1: Crime Data

We use the crime data at Florida and California to study the prediction of the number of violent crimes (per population). Use the following code to load data.

```
crime <- read.csv("CrimeData_sub.csv", stringsAsFactors = F, na.strings = c("?"))
crime <- na.omit(crime)
```

Our goal is to find the factors/variables which are relate to violent crime. This variable is included in crime as crime$violentcrimes.perpop.

Randomly divide the data into 80% training and 20% testing.

Q1.1

Run the ordinary least square regression with all the variables and with the training data. Get MSE and R^2 for both the training and testing data and see if there is a difference.

Hint: MSE is the average of the RSS (residual sum of squares). Here you evaluate RSS on two types of data: training data and testing data.

Q1.2

Use the training data to fit the Lasso and obtain a reasonable, small model. Re-fit an OLS model with the variables obtained. The final model should only include variables with p-values < 0.05. Note: you may choose to use lambda with the smallest CV error to answer the following questions where applicable.

i. What is the model reported by LASSO? Use 5-fold cross-validation to select the tuning parameter.
ii. What is the model after refitting OLS with the selected variables? What are MSE and R^2 for the training and testing data? Compare them with results in Q1.1.
iii. What is your final model, after excluding high p-value variables? Use the model selection method to obtain this final model. Make it clear what criterion/criteria you have used and justify why they are appropriate.
iv. Try Ridge regression with a 5-fold CV to select the tuning parameter. Compare its training and testing MSE and R^2 with the previous models.

Question 2: Default Data from ISLR

The Default data from ISLR can be loaded by the following R commands:

library(ISLR)

data(Default)

?Default

Q1.1

Fit a logistic regression with the variable ***student*** as the X variable and ***default*** as the response variable. Interpret the coefficients and discuss whether the X variable is significant.

Q1.2

i. Consider all the variables and fit the final logistic regression model in R.

ii. Compare the ROC curves for the model in Q1.1 and Q1.2, along with the corresponding AUC.

iii. Consider a threshold of 0.5 on the probability of default. Calculate the corresponding specificity, sensitivity, false positive rate, and true positive rate for the model in part i.

Question 3: Lost Sales

In many industries throughout the world, suppliers compete for business by submitting quotes for work, services or products. A key criterion used to determine the winning quote is the dollar amount of the quote, but other factors include expected quality, estimated delivery time of the product, or quoted completion time of the work.

The focus of this case is a supplier of equipment to the automotive industry. The products of interest in this case are various precision metal components used in a range of automotive applications, such as braking systems, drive trains, and engines. Some of the products will be used in the manufacture or assembly of new automobiles (i.e. original equipment), while others will be used as replacement parts in automobiles already on the road (i.e. aftermarket).

The supplier wants to increase sales and expand its market position. Many of the quotes provided to prospective customers in the past haven't resulted in orders. Does the data provide any indication of the reasons? Are there certain situations that make it more or less likely that a customer will place an order?

Please fit a model using the available data to explore these questions. Based on the fitted model, please provide your answers to the above questions. Drop insignificant variables for variable selection if you like. It can be based on p.values or AIC.

The data set contains 550 records for quotes provided over a six-month period. The variables in the data set are:

Quote: The quoted price, in dollars, for the order

Time to Delivery: The quoted number of calendar days within which the order is to be delivered

Part Type: OE = original equipment; AM = aftermarket

Status: Whether the quote resulted in a subsequent order within 30 days of receiving the quote: Lost = the order was not placed; Won = the order was placed.

Question 4: Wine Quality

The wine quality data contain information on quality ratings for 6,497 different wines, along with measures of wine properties. The response variable is **Quality**. Please build a model to predict wine quality.

Conduct exploratory data analysis and see if there is redundant or irrelevant information in the data. If so, remove them. Then consider two possible modeling choices:

i. Multinomial logistic regression

ii. Ordinal logistic regression

Write up a summary of the characteristics of **Good** quality wine for each model. Do you roughly get similar conclusions using those two models?