

**BUDT 758T**

**assignment #2: 100 PTS**

The goal of this homework is to introduce you to classification concepts. You will develop (1) a linear probability (regression) model and (2) a logistic regression model. You will need to create random partitions of a data set, build your model on the training data set, and then compute prediction errors using the test data set. You are required to complete this assignment in R—be sure to include the code you used and output any results you use!

**The Data**

The data in the Airlines data file (“Airlines\_Data.csv” posted on Canvas) contains data from 3,999 airline customers enrolled in East-West Airlines’ customer rewards program. (Note that while East-West Airlines is clearly fictional, this is data from a real airlines reward program; names have been changed to protect the innocent and not-so-innocent alike.) East-West Airlines has two goals with this analysis: (1) identifying if a customer will claim a travel award using their rewards, and (2) identifying factors that lead to customers claiming a travel award. The data contains information about each customer’s history:

|  |  |
| --- | --- |
| **Field Name** | **Description** |
| *ID* | Unique ID |
| *Balance* | Number of miles eligible for award travel (including deductions of any miles for claimed travel awards) |
| *Qual\_miles* | Number of miles counted as qualifying for Topflight status |
| *cc1\_miles* | Number of miles earned with freq. flyer credit card in the past 12 months: |
| *cc2\_miles* | Number of miles earned with Rewards credit card in the past 12 months: |
| *cc3\_miles* | Number of miles earned with Small Business credit card in the past 12 months: |
| *note: miles are binned* | 1 = under 5,000 |
| *for all cc\_miles variables* | 2 = 5,000 - 10,000 |
|  | 3 = 10,001 - 25,000 |
|  | 4 = 25,001 - 50,000 |
|  | 5 = over 50,000 |
| *Bonus\_miles* | Number of miles earned from non-flight bonus transactions in the past 12 months |
| *Bonus\_trans* | Number of non-flight bonus transactions in the past 12 months |
| *Flight\_miles\_12mo* | Number of flight miles in the past 12 months |
| *Flight\_trans\_12* | Number of flight transactions in the past 12 months |
| *Days\_since\_enroll* | Number of days since Enroll\_date |
| *Award* | Dummy variable for travel award claimed (1 = award claimed, 0 = not claimed) |

**Assignment**

Please answer all questions in the dedicated space and upload on Canvas. Please ensure that your numbering of questions matches those below. Include any R code you used to answer each question with your response. You are welcome to include any output you wish to provide either with the appropriate question or at the end of your assignment in an appendix.

Remember: you are allowed to consult with others in the class on this assignment, but all submitted work must be your own (and don’t forget to include the names of anyone you consulted in the last question!).

1. **5 points: Data Preparation**
   1. Read the data set in R.

airlines <- read.csv("Airlines\_Data.csv")

attach(airlines)

* 1. Change the *cc1\_miles, cc2\_miles*, and *cc3\_miles* variables to factor variables using **as.factor()** in R. Use these new variables for the remainder of the assignment. Notes:
     + All cc\_miles variables have the same bins, given under *cc3\_miles.* The differences in the variables are in where the number of miles came from.
     + Make sure you change the variables in the data set rather than creating new variables! This will be important when you partition.

airlines$cc1\_miles=as.factor(airlines$cc1\_miles)

airlines$cc2\_miles=as.factor(airlines$cc2\_miles)

airlines$cc3\_miles=as.factor(airlines$cc3\_miles)

* 1. Set the seed in R to 14632.

set.seed(14632)

* 1. Randomly partition the data set into a training data set and a test data set. Use 65% of the data as training data and hold out the remaining 35% as test data.

num\_obs <- nrow(airlines)

split <- sample(num\_obs, 0.65\*num\_obs)

airlines\_train <- airlines[split,]

airlines\_test <- airlines[-split,]

1. **20 points: Exploratory analysis of the data**
   1. In Assignment 1, we used *Award* to predict *Balance* in our first simple linear regression and found *Award* was a significant positive predictor. Given the definitions of the variables above:
      1. Does this make sense to you? Why or why not?

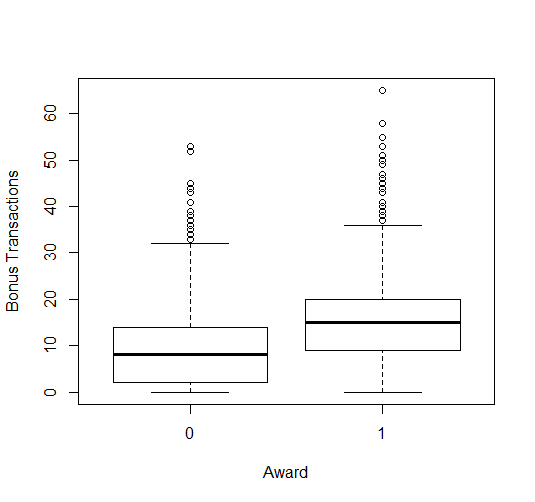
No, because Balance in reality decreases when one claims his Award and this indicates a negative relationship, not a positive one.

* + 1. Would it make more sense to use *Balance* to predict *Award*? Why or why not?

No, it doesn’t make more sense to use Balance to predict Award because the relationship doesn’t change by interchanging the variables.

* 1. Using the training data, construct and report a boxplot of the number of non-flight bonus transactions in the past 12 months (*Bonus\_trans*), broken out by customers who have claimed a travel award vs. customers who have not (*Award*). Does there appear to be a relationship between *Bonus\_trans* and *Award*? How do you know?

boxplot(airlines\_train$Bonus\_trans~airlines\_train$Award, xlab='Award', ylab='Bonus Transactions')

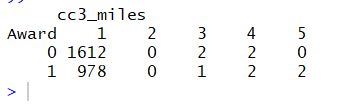


Yes, there appears to be a positive relation i.e for customers who have claimed the Award have higher Non flight bonus transactions in the last 12 months compared to the customers who have not claimed the Award.

We can clearly know this by looking at our boxplot as the min, max and median of bonus\_trans are all higher for Award=1 compared to Award=0.

* 1. Using the training data, construct and report a table for the count of travel award broken up by *cc3\_miles* (i.e. how many customers in each *cc3\_miles* category have claimed a travel award vs. not claimed a travel award).

table(airlines\_train$Award, airlines\_train$cc3\_miles, dnn=c('Award','cc3\_miles'))



* + 1. Without running any models: If you were to create a simple linear or logistic regression to predict *Award* using *cc3\_miles* using the training data, would you expect the *cc3\_miles* variable to be significant? Why or not why?

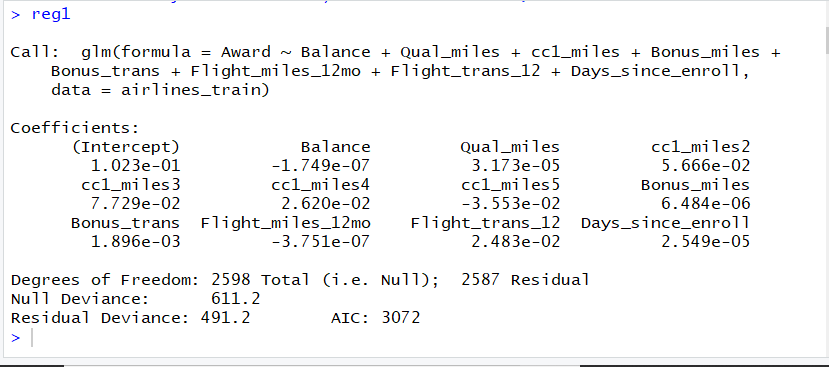
I would expect cc3\_miles variable is not significant as cc3\_miles1 category (the base) has the majority of customers in it and hence the other catergories(2,3,4,5) do not provide any additional information as they do not have enough customers in them.

* + 1. Without running any models: If you ran the regression in part (i) and try to use the model to predict for the testing data, R will give you an error. Based on the training data table you created, why will it give an error? (It may be helpful to create a table for the testing data as well, but it is not necessary.)

The error is given because there are no cc3\_miles2 in the training data but there exists some cc3\_miles2 in the testing data.

1. **40 points: Run a linear regression model to predict *Award* using *Balance, Qual\_miles, cc1\_miles, Bonus\_miles, Bonus\_trans, Flight\_miles\_12mo, Flight\_trans\_12,* and *Days\_since\_enroll*. Use only the training data set for this.**

reg1 = glm(Award~Balance+Qual\_miles+cc1\_miles+Bonus\_miles+Bonus\_trans +Flight\_miles\_12mo+Flight\_trans\_12+Days\_since\_enroll,data=airlines\_train)



1. Compute and report the RMSE using your model for both the training data and the testing data sets. Use predicted values from the regression equation and **do not** do any classification yet.

RMSE\_train <- sqrt(mean((reg1\_pred\_train-airlines\_train$Award)^2))

RMSE\_test <- sqrt(mean((reg1\_pred\_test-airlines\_test$Award)^2))

1. For which data set are these errors smaller? Does this surprise you? Explain.

RMSE of testing data is 0.429032 which is less than RMSE of training data which is 0.434734. This does surprise me because usually RMSE values for most models are less for training data compared to testing data because we used the training data to train our model.

1. What would the RMSE be for this model using the baseline? Based on RMSE, is this model better than using the baseline?

RMSE\_baseline <- sqrt(mean((mean(airlines\_test$Award)-airlines\_test$Award)^2))

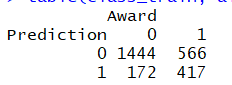
RMSE of baseline model is 0.4787292. Yes, this model is better than using the baseline because the RMSE value is lesser compared to the RMSE of baseline.

1. Now use a cutoff of 0.5 and do the classification. Compute and report the confusion matrix for both training and testing data predictions.

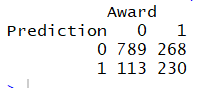
class\_train <- ifelse(reg1\_pred\_train > 0.5,1,0)

class\_test <- ifelse(reg1\_pred\_test > 0.5,1,0)

table(class\_train, airlines\_train$Award, dnn=c('Prediction','Award'))



table(class\_test, airlines\_test$Award, dnn=c('Prediction','Award'))



* + 1. How many false negatives are there in the testing data? In terms of the problem, what is a false negative for this regression?

sum(ifelse((class\_test==0 & airlines\_test$Award==1),1,0))

268 false negatives are there in the testing data.

False negative means predicting 0 where actual label was 1. So, in this regression it means predicting that Award was not claimed when in reality the Award was actually claimed.

* + 1. Does this model seem to perform equally well for both the training data and the testing data? Explain your answer.

This model seems to perform equally well for both the training and testing data because the 4 values in both the above tables look similar accounting for the number of observations in both the datasets.

sum(ifelse(airlines\_train$Award==class\_train,1,0))/nrow(airlines\_train)

sum(ifelse(airlines\_test$Award==class\_test,1,0))/nrow(airlines\_test)

This can also be said seeing at the accuracies as the training accuracy is 71.6% and testing accuracy is 72.8% and both the accuracies are very close.

1. What is the testing accuracy of your model? Calculate the testing accuracy using the most common class baseline. According to accuracy, is this model better than using the baseline?

The testing accuracy for this model is 72.8%

sum(ifelse(airlines\_test$Award==0,1,0))/nrow(airlines\_test)

The testing accuracy using the most common class baseline is 64.4%. Hence, according to accuracy, the model is better than using the baseline because the model provides a greater accuracy.

1. Compute and report the predicted “probability” that the following customer claims a travel award. Does your answer make sense? Why or why not?

Balance = 10000

Qual\_miles = 20000

cc1\_miles = 1

Bonus\_miles = 20000

Bonus\_trans = 50

Flight\_miles\_12mo = 25000

Flight\_trans\_12 = 25

Days\_since\_enroll = 1000

new\_entry <- data.frame(Balance=10000,Qual\_miles=20000, cc1\_miles='1',Bonus\_miles=20000,Bonus\_trans=50, Flight\_miles\_12mo=25000,Flight\_trans\_12=25,Days\_since\_enroll=1000)

predict(reg1,new\_entry)

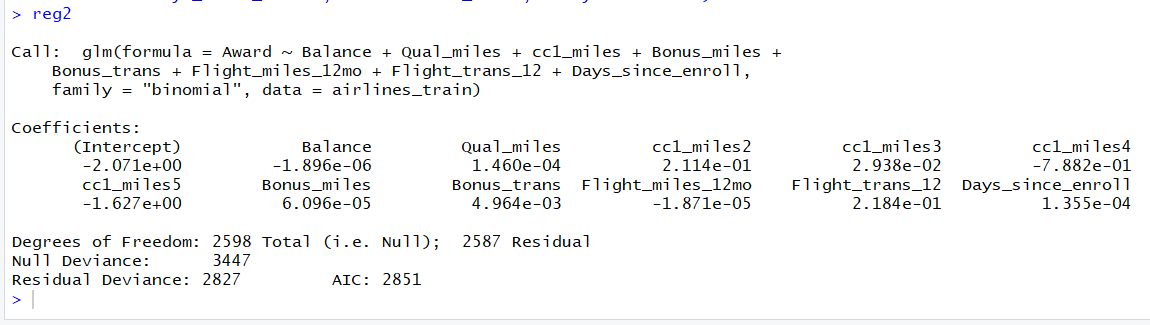
The predicted probablity that the customer claims a travel award is 1.60. This doesn’t make sense because probability can take values from 0 to 1 but not anything beyond 1.

1. **35 points: Run a logistic regression model to predict *Award* using *Balance, Qual\_miles, cc1\_miles, Bonus\_miles, Bonus\_trans, Flight\_miles\_12mo, Flight\_trans\_12,* and *Days\_since\_enroll*. Use only the training data set for this.**

reg2 <- glm(Award~Balance+Qual\_miles+cc1\_miles+Bonus\_miles+Bonus\_trans

+Flight\_miles\_12mo+Flight\_trans\_12+Days\_since\_enroll,data=airlines\_train,

family='binomial')



* 1. What is the coefficient for *Bonus\_trans*? Provide a precise (numerical) interpretation of this coefficient.

The coefficient of Bonus\_trans is 0.004964. For every 1 increase in a non flight bonus transaction, the odds of claiming the Award increases by e^0.004964 times on average and keeping everything else constant.

* 1. What is the coefficient for cc1\_miles4? Provide a precise (numerical) interpretation of this coefficient.

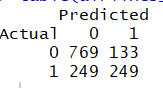
The coefficient of cc1\_miles4 is -0.7882. For every additional mile in the cc1\_miles4 category, the odds of claiming the Award decreases by e^-0.7882 compared a customer in cc1\_miles1 category on average and keeping everything else constant.

* 1. Use a cutoff of 0.5 and do the classification. Compute and report the confusion matrix for your testing data predictions.

reg2\_preds <- predict(reg2, newdata=airlines\_test,type='response')

reg2\_class\_test <- ifelse(reg2\_preds>0.5,1,0)

table(airlines\_test$Award, reg2\_class\_test, dnn=c('Actual','Predicted'))



* 1. Compare this confusion matrix to the one in question 3(c) above. Which model do you think is better to use in practice? Why?

Since it looks like the company is more interested in predicting if someone claimed an Award, the logistic model is better to use as it has got a better True Postive Rate than the linear model.

But if the company is interested in a better True Negative Rate it should go for a linear model as it has got better rate compared to the Logistic model. So, it depends on what we want from our model to say which one is better.

* 1. Compute and report the predicted probability using your logistic model for the same customer from 3(f). Does this make more sense than your answer to 3(f)? Why or why not?

predict(reg2,new\_entry, type='response')

The predicted probability is 0.99. Hence this makes more sense than the answer to 3(f) because the value is in the valid probability range of 0 to 1 and we can assign 1 to the Award based on our cutoff value.