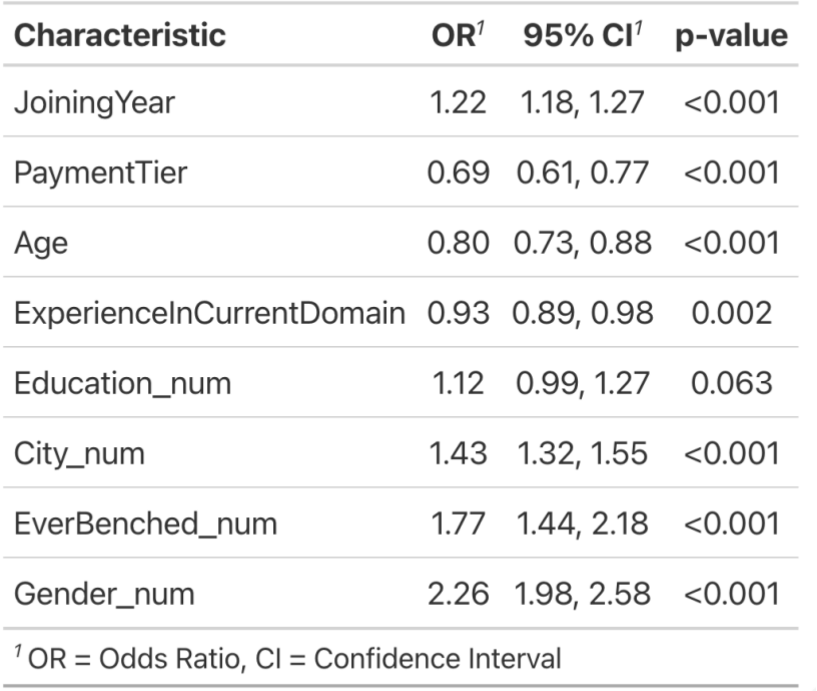
IS 507 Assignment 8
Due by Sunday, December 5, 2021 11:59PM

Problem 1: The Excel spreadsheet Employee.csv contains one sheet named Employee, which is data attempting to explain employee retention. These are data from a sample of 4,654 employees and consists of 9 variables for each employee. These are:
1) Education Level (Categorical)
2) Joining Year (Year Employee joined the company)
3) City (Office Location)
4) Payment Tier (1-Highest Level; 2-Mid Level; 3-Lowest Level)
5) Age (Current Age of Employee)
6) Gender (Gender of Employee-Male/Female)
7) Ever Benched (Left off of a project for 1 or more months)
8) ExperienceInCurrentDomain (Level of Experience)
9) LeaveOrNot (1-Leaves Company in 2 years; 0-Does not Leave Company in 2 years)-Dep. Var.

Note: To use the categorical variables you will first need to convert them into numeric data. There are several different ways of completing this task, make sure you know your reference category.

Develop a **Logistic Regression model** to classify the LeaveOrNot event from the other variables.

a) Create a logistic regression model and explain the significant odds ratios in terms of LeaveOrNot.



Characteristic	OR[1]	95% CI[1]	p-value
JoiningYear	1.22	1.18, 1.27	<0.001
PaymentTier	0.69	0.61, 0.77	<0.001
Age	0.80	0.73, 0.88	<0.001
ExperienceInCurrentDomain	0.93	0.89, 0.98	0.002
Education_num	1.12	0.99, 1.27	0.063
City_num	1.43	1.32, 1.55	<0.001
EverBenched_num	1.77	1.44, 2.18	<0.001
Gender_num	2.26	1.98, 2.58	<0.001

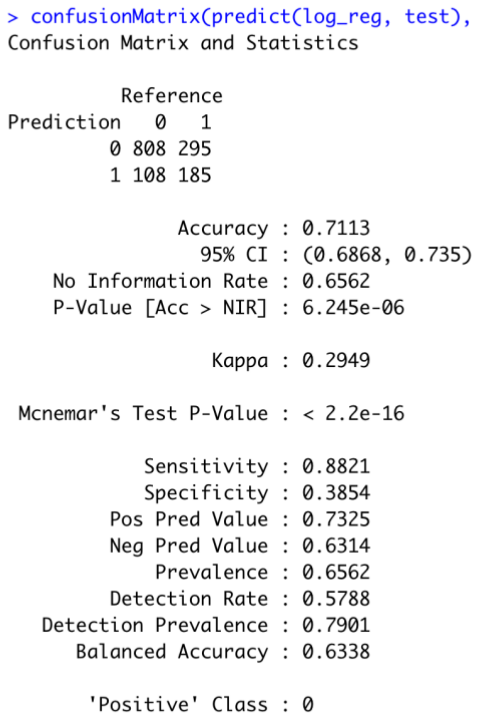
[1] OR = Odds Ratio, CI = Confidence Interval

An odds ratio greater than 1, shows an increased likelihood of an event occurring, an odds ratio less than 1 shows a decreased likelihood of the event occurring. The higher the score the higher the impact on LeaveOrNot.

For a variable to be significant it must have a p value less than 0.005 and cannot have a confidence interval in the range of 1, therefore education_num is not significant.

Gender_num is the most significant factor and has the highest impact on LeaveOrNot. Followed by EverBenched_num, City_num, and Joining_year. On the other hand, as values for ExperienceInCurrentDomain, Age and PaymentTier increase, the likehood of leaving within 2 years decreases.

b) Create a confusion matrix and explain how well the model is classifying the Leaves Company in 2 years events.

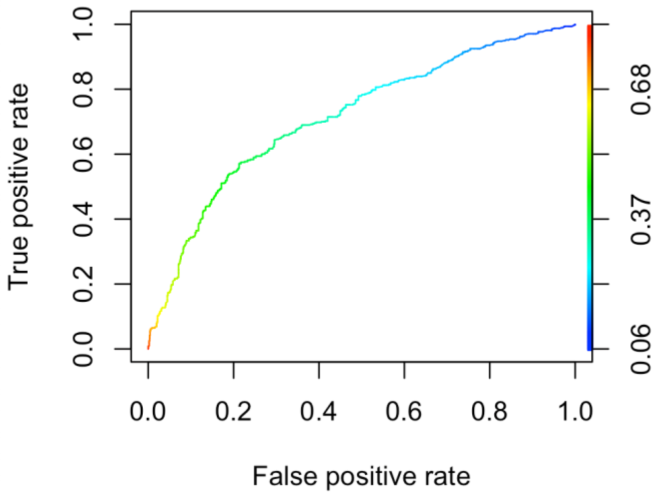


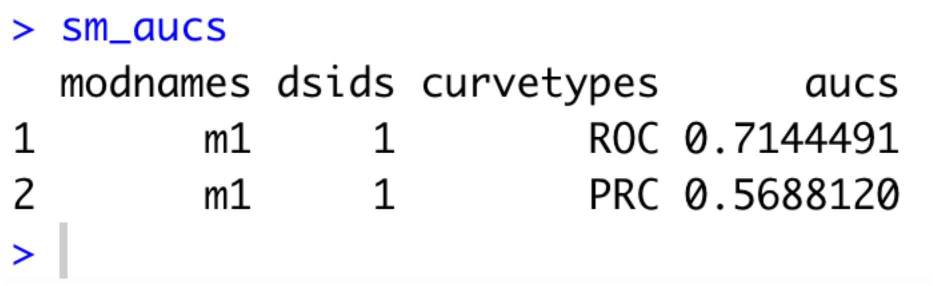
```
> confusionMatrix(predict(log_reg, test), as.factor(test$LeaveOrNot))
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 808 295
         1 108 185
                                          
               Accuracy : 0.7113          
                 95% CI : (0.6868, 0.735) 
    No Information Rate : 0.6562          
    P-Value [Acc > NIR] : 6.245e-06       
                                          
                  Kappa : 0.2949          
                                          
 Mcnemar's Test P-Value : < 2.2e-16       
                                          
            Sensitivity : 0.8821          
            Specificity : 0.3854          
         Pos Pred Value : 0.7325          
         Neg Pred Value : 0.6314          
             Prevalence : 0.6562          
         Detection Rate : 0.5788          
   Detection Prevalence : 0.7901          
      Balanced Accuracy : 0.6338          
                                          
       'Positive' Class : 0               
```

The matrix has an accuracy of ~71% and a sensitivity of 88%. Balanced accuracy of 63% is low which might mean that a higher than expected number of employees may leave within 2 years.

c) Create an ROC curve and calculate the c-statistic (auc). What does this mean about the model?





```
> sm_aucs
  modnames dsids curvetypes      aucs
1       m1     1        ROC 0.7144491
2       m1     1        PRC 0.5688120
> 
```

The c-statistic/AUC has a value of 0.71. This indicates that the model can classify true positives and true negatives better than false positives and false negatives.

d) What are the differences between the information in part a and part b?

Part A shows the odds ratio and the confidence interval of factors that impact the likelihood of an employee leaving. Part A focuses more on the feature selection and dimensionality reduction. Part B focuses on the model's performance using the confusion matrix which enables the user to view the accuracy sensitivity specificity of the model. Part B shows us the number of true positives and true negatives as well as type 1 and type 2 errors.

e) How does this model differ from the linear discriminate analysis you ran in Assignment 7?

The LDA model had an accuracy of ~71% and an accuracy of around 61%. The logistic regression modelacheived a higher sensitivity of 88% which enables the logistic regression model to have fewer type 1 and type 2 errors.