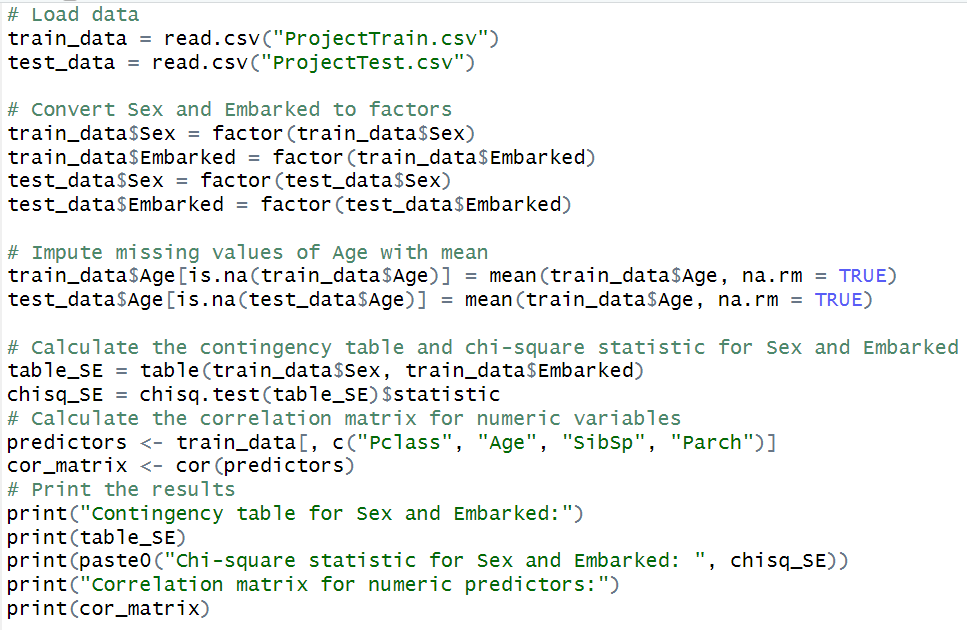
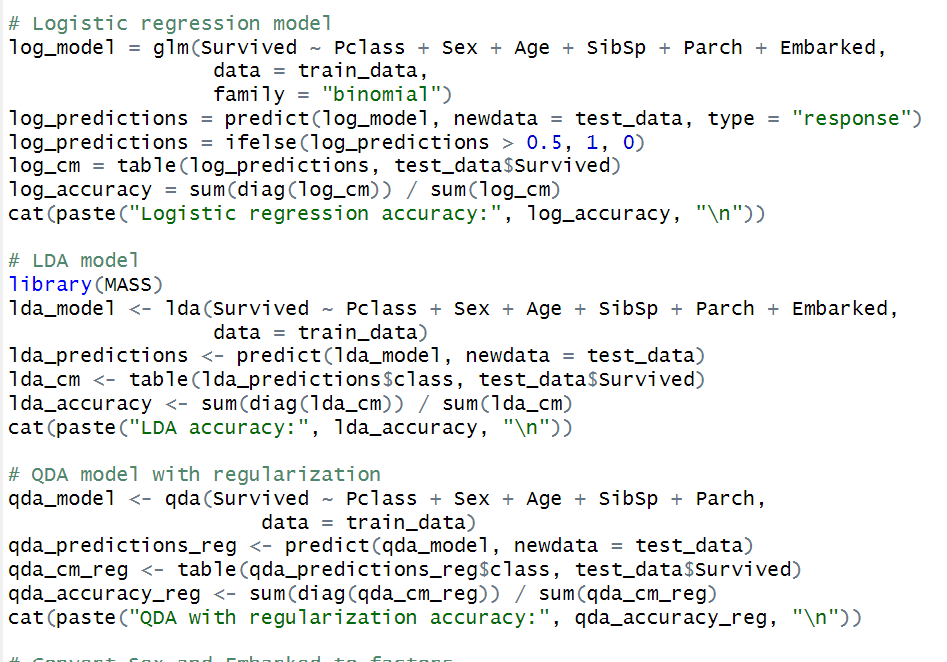
Assignment 6: Titanic Survival Prediction

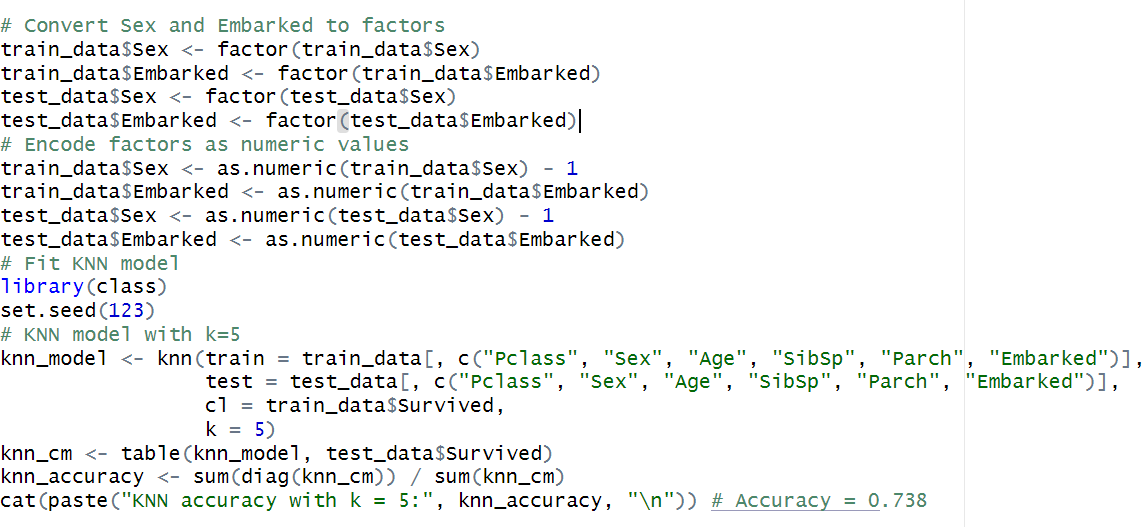
This assignment is for 400 points, four times the normal assignment weight. The goal of the project is to predict what type of persons were more likely to survive? The features available are Name, Age, Gender, Fare Class, etc. Data dictionary is provided in the appendix. Data is partitioned into (1) ProjectTrain.csv, and (2) ProjectTest.csv. Use Train data to develop the model and report performance results on the Test dataset.

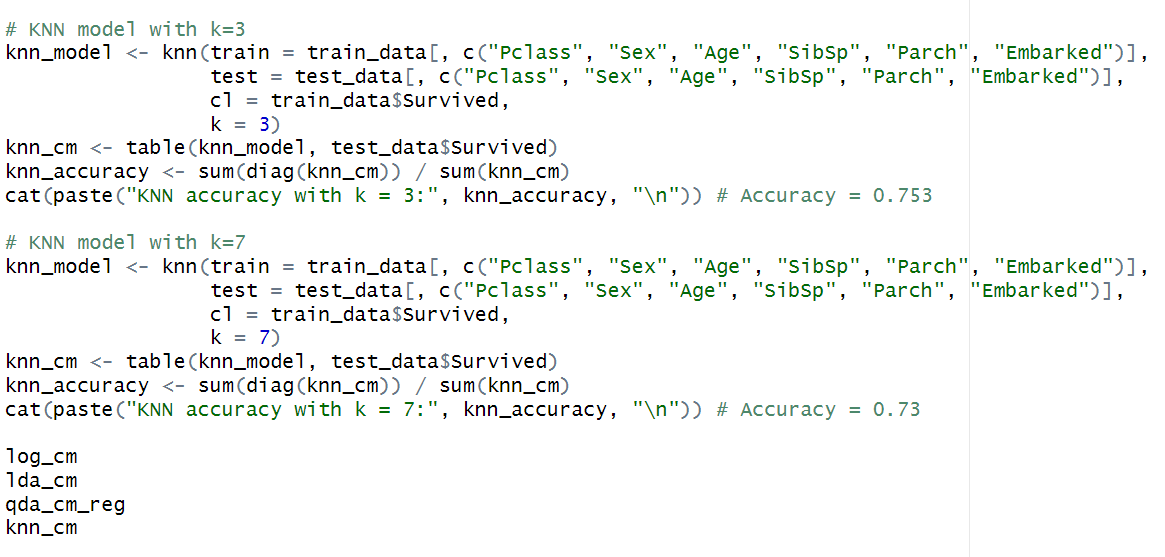
1. Develop Logistic Regression, LDA, QDA and KNN based survival prediction models using Pclass, Sex, Age, SibSp, Parch, and Embarked as predictor variables. Note that some of these variables may need to be defined as categorical (factors in R). Also, Age has lot of missing values. The missing values may need to be imputed (e.g., mean) for using this variable. Try few values of k in KNN to determine suitable value for K. Compare and interpret True Positive (TP) and False Positive (FP) of the different models using test data.

By using following coding, we can fit the model of Logistic Regression, LDA, QDA and KNN:



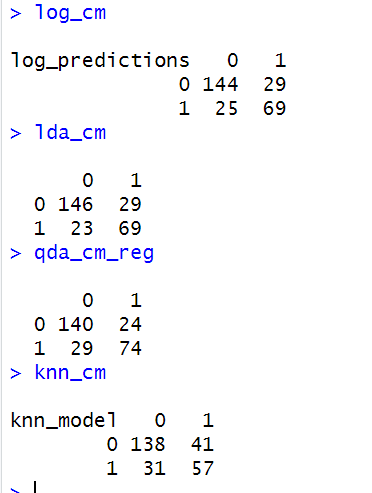






When I was trying to fit the QDA model, it kept give me an error about high correlation about variables. So, I use chi-squared to check the multicollinearity. In this case, the chi-square statistic for Sex and Embarked is 11.547 and the p-value is 0.009, which indicates that there is a statistically significant association between the two variables. I decided to delete the “embarked” variable for QDA fitting model.

By trying different values of K (3,5,7), it seems that when K = 3, the KNN model will have highest accuracy value.

And by printing out their confusion matrix we can see:  


So, for logistic regression: True Positive (TP): 69 False Positive (FP): 29

For LDA: True Positive (TP): 69 False Positive (FP): 29

For QDA: True Positive (TP): 74 False Positive (FP): 24

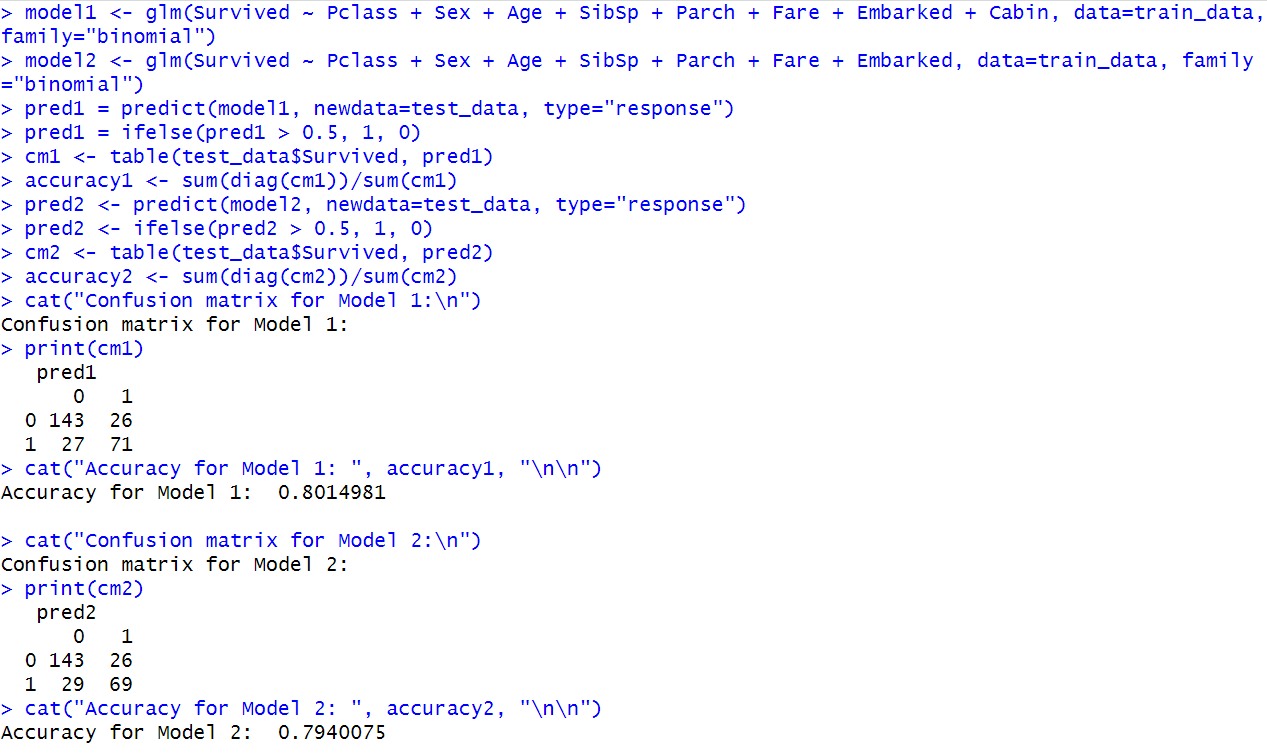
For KNN: True Positive (TP) = 57 False Positive (FP) = 41

Form the above result, we can see that logistic regression, LDA and QDA they both have similar accuracy, but QDA is better when I take off the variable of “Embarked”.

KNN seems works worse for this case, because it has lower true positive and higher false positive value compared with other models.

1. “Cabin” has sparse data content. One approach to handle the missing data is to have a special value “Not Available” for all the missing values. For the Logistic Regression model, evaluate performance improvement with and without including the cabin feature using test data.

By converting missing value to “Not Available” and non-missing value to “Available”, we can get following coding:



Therefore, we can see that there is a little difference of accuracy between adding the indicator “Cabin” or without it.

With the cabin feature:

Precision = 71 / (71 + 26) = 0.731

Recall = 71 / (71 + 27) = 0.724

F1-score = 2 \* (0.731 \* 0.724) / (0.731 + 0.724) = 0.727

Without the cabin feature:

Precision = 69 / (69 + 26) = 0.726

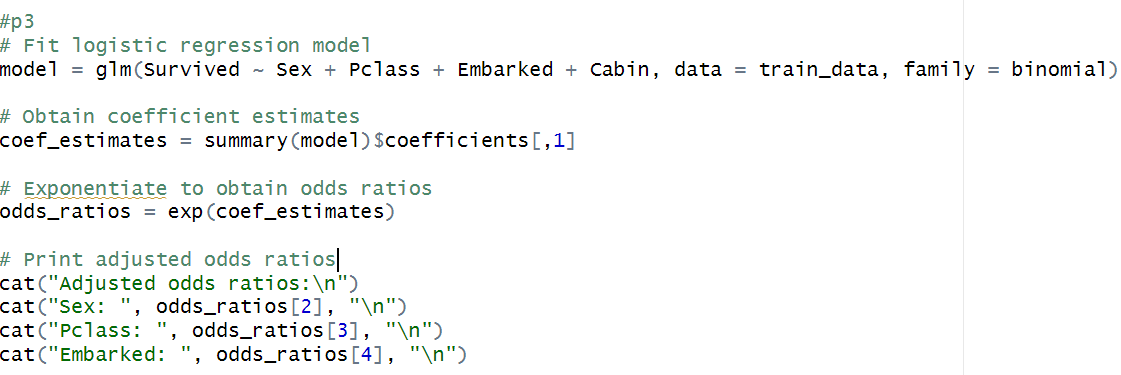
Recall = 69 / (69 + 29) = 0.704

F1-score = 2 \* (0.726 \* 0.704) / (0.726 + 0.704) = 0.715

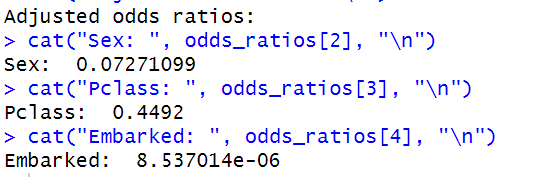
From the above calculations, we can see that Model 1 has a slightly higher overall score than Model 2, so that Model 1 performs slightly better overall on the test data. So, we can consider the cabin as a indicator, but it won’t influence a lot.

1. Like linear regression, Logistic regression (LR) has the advantage of interpretability. Research the concepts of “Unadjusted Odds Ratio” and “Adjusted Odds Ratio”. Determine the adjusted odds ratio for Sex, Pclass, and Embarked using LR. Interpret the results.

By using following coding:



We can get:

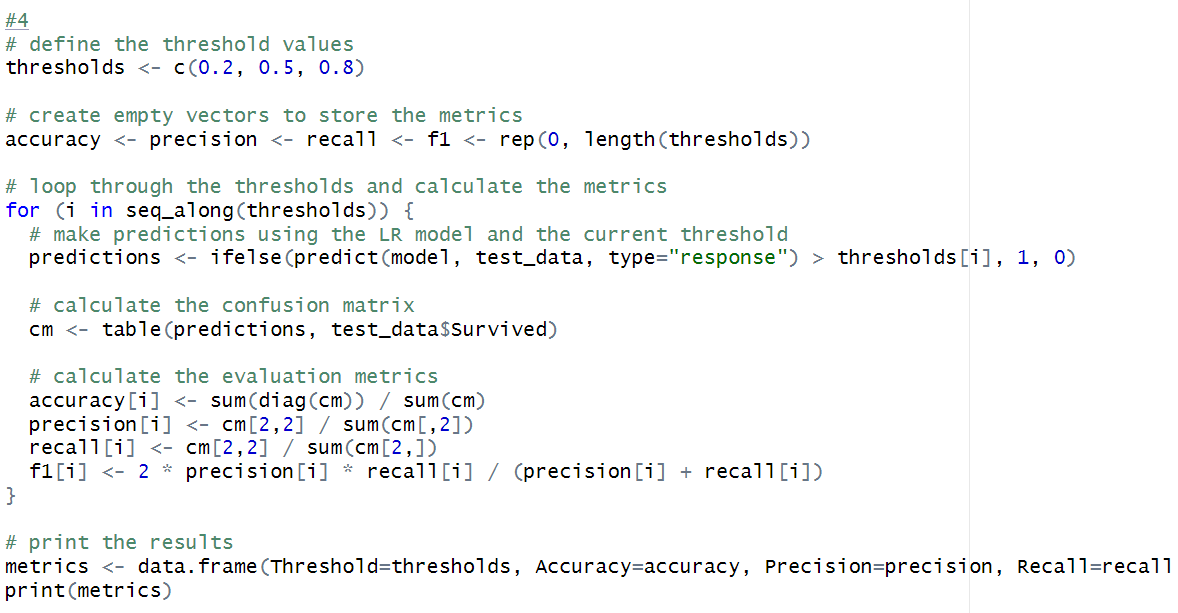


So, for the adjusted odds ratios obtained in this LR model, the odds of survival for females are 0.0727 times the odds of survival for males. This means that females are more likely to survive than males, which is consistent with the common understanding that women were given priority in lifeboats during the Titanic disaster.

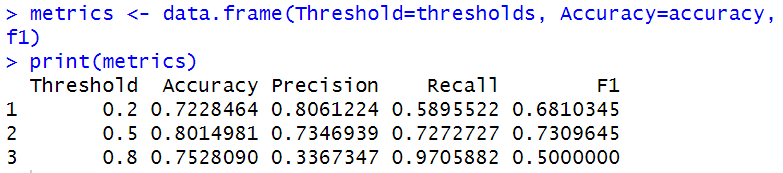
The odds of survival for a passenger in Pclass 2 or Pclass 3 are 0.4492 times the odds of survival for a passenger in Pclass 1. This means that being in a higher passenger class is positively associated with survival.

Finally, the odds of survival for a passenger embarking from ports C or Q are 0.0000085 times the odds of survival for a passenger embarking from port S. This indicates that passengers embarking from ports C or Q are much less likely to survive than those embarking from port S.

1. The default threshold to classify an entity to a class is 0.5. For the LR models, vary the threshold to 0.8, 0.5, and 0.2. Which threshold value do you think is appropriate for survival prediction? Why? Justify your answer with respect to misclassification rate on test data.

By using following coding: 

So, we can get:



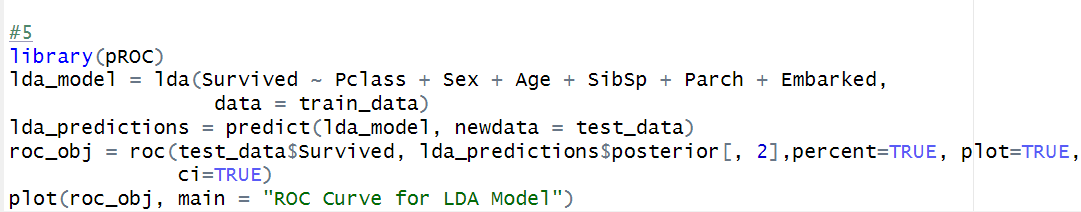
Based on the metrics, it appears that a threshold of 0.5 is the most appropriate for survival prediction. This is because it has the highest accuracy (0.8014981) and F1 score (0.7309645), which are important metrics for binary classification tasks for survival prediction.

Also, the precision and recall values at this threshold are reasonably balanced, that means a good trade-off between correctly predicting positive cases (survivors) and minimizing false positives.

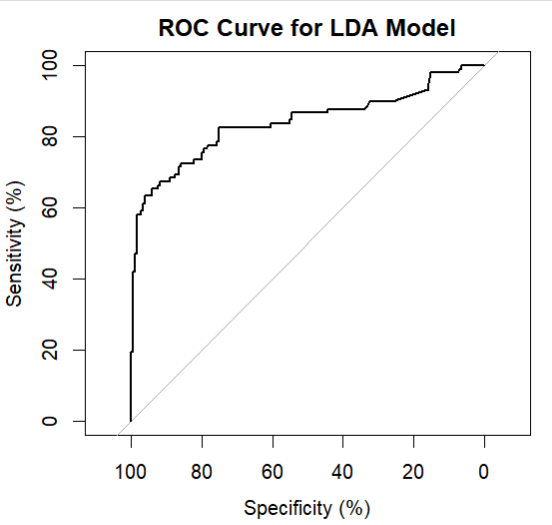
So, in general, based on the misclassification rate on test data and the trade-off between precision and recall, a threshold of 0.5 is the most appropriate for survival prediction.

1. Develop ROC plot for the LDA model.

By using following data:



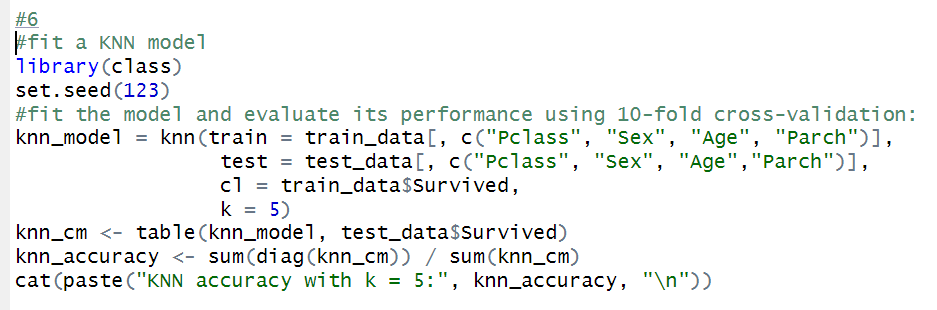
We can get the LDA model:



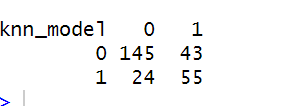
1. What features do you think are important to make the prediction? Why? Evaluate the KNN model performance by including just the important features

I believe that the following features are crucial for prediction: age, sex, passenger class, and the number of parents/children aboard. Passenger class can serve as a proxy for socio-economic status. This is because in the event of a disaster, rescue efforts tend to prioritize young and female individuals. Additionally, if there are wealthy individuals present, they may be able to use their resources to secure rescue. Therefore, I believe that these features are important in predicting survival in such scenarios.

By using following coding to fitting the KNN model:



The confusion matrix is:



So,

Accuracy: (TP+TN)/(TP+TN+FP+FN) = (145+55)/(145+55+43+24) = 0.75

Error rate: (FP+FN)/(TP+TN+FP+FN) = (43+24)/(145+55+43+24) = 0.25

Sensitivity (True Positive Rate): TP/(TP+FN) = 55/(55+24) = 0.696

Specificity (True Negative Rate): TN/(TN+FP) = 145/(145+43) = 0.771

Based on the result, the KNN model has an accuracy of 0.75, an error rate of 0.25, a sensitivity of 0.696, and a specificity of 0.771 when evaluated using the important features of age, sex, passenger class, and number of parents/children aboard.