ACTSC 632 – Project for Module 4 – due on July 10

This project will consider the German Credit data set available in the `CASdatasets` package in which it is named `credit`. As a team of data analysts for a credit card company, you have been asked to compare different classification methods and provide a recommendation to your manager for which one is the best to identify the "good" (i.e., creditworthy, no non-payments) vs the "bad" (not credit-worthy, having existing non-payments) credit files.

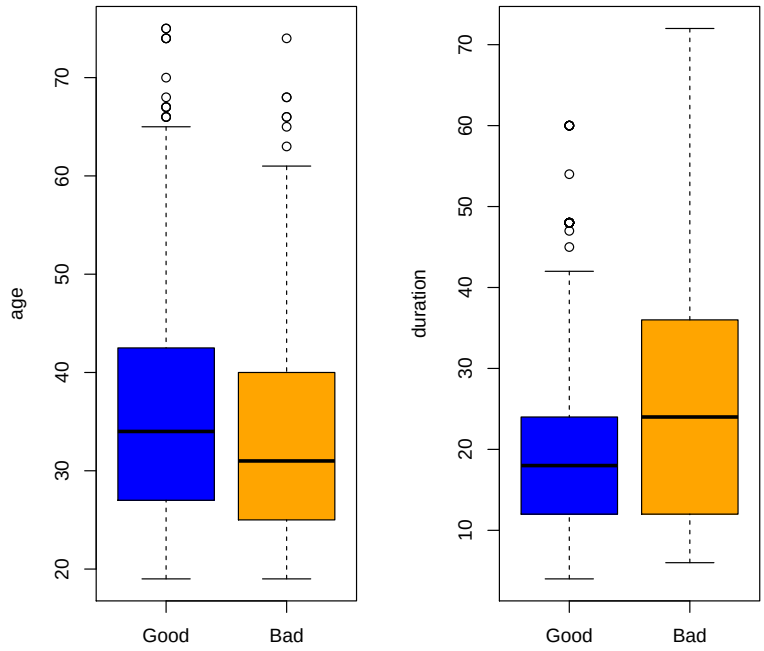
This project is to be complete in teams. Information about teams' membership will be emailed on LEARN.

1. Determine the number of predictors in this data set, and whether they are quantitative of qualitative. How many observations are there? How many observations are classified as a "bad credit"? "good credit"?

   Solution: There are 20 predictors, duration, credit_amount, age are quantitative, and the other are qualitative. Depending on the encoding used for the data set (which differs from source to source) the predictors installment_rate, residence_since, existing_credits and num_dependents may be encoded as integers and thus be a priori be considered quantitative. But when looking at the meaning of the different values these predictors can take, it seems best to treat them as categorical. It won't matter in our analysis though, as these 4 predictors do not end up being considered in our models. and the others are categorical. There are 300 observations out of 1000 classified as bad credit.

2. To explore the data further, produce the following plots: (i) for each of the predictors age and duration, make a box plot showing the distribution of the observations, separately for the "good" and "bad" observations.; (ii) for the pairs duration & savings and duration & credit history, plot the observations (using duration on the x axis) and use different symbols for the "good" and "bad" observations. Comments on the plots you obtained.

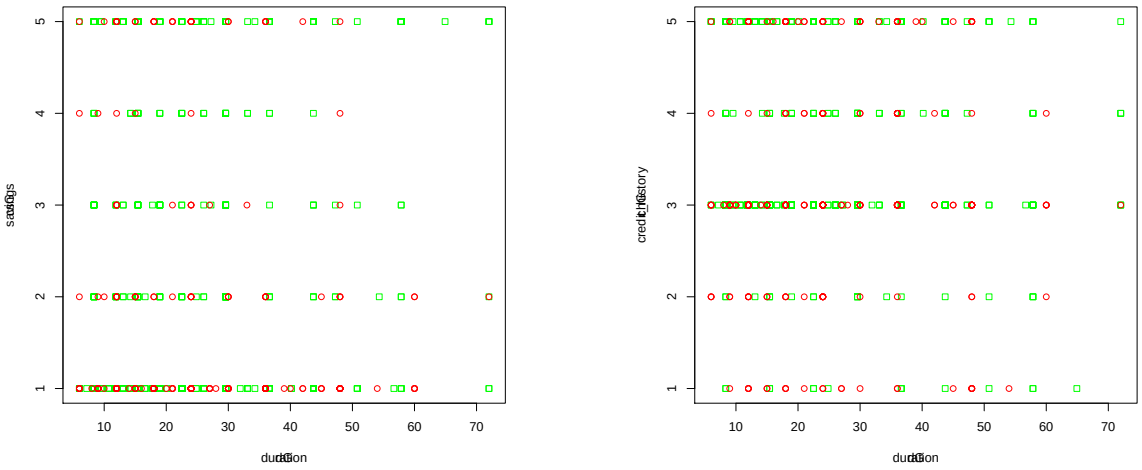
   Solution: For (i) we get



Clearly the bad credit observations seems to have a longer duration, and the distribution of the duration has a larger variance; they also seem to be slightly younger compared to the good credit.

For (ii) we get

Figure 1: Bad credits shown in red, good shown in green



It is harder to see a clear trend in these two graphs, in that the good and bad credits are not easily classified according to duration for a given level of credit history or savings. However

3. Initially, you should work with the following predictors: age, duration, purpose, credit_history, and savings. For each of (i) logistic regression (ii) linear discriminant analysis (iii) quadratic discriminant analysis, do the following:

   (a) Determine the confusion matrix, the overall error rate, Type-I error, Type-II error.
   Solution: We get

			overall	type 1	type 2
logistic	0	1			
0	646	192			
1	54	108	0.246	0.07714	0.64
lda	0	1			
0	644	191			
1	56	109	0.247	0.08	0.6367
qda	0	1			
0	552	133			
1	148	167	0.281	0.2114	0.4433

   (b) Plot the ROC curve for the three methods considered and determine the AUC (area under the curve).
   Solution: the 3 ROC curves have respective AUC of 0.7609, 0.7613, 0.7461 and are given in Figure 3.

4. The manager who asked you to recommend a classifier for this problem says you'd be allowed to use up to two more predictors to build your classifier. Choose one or two additional predictors and determine how their inclusion affects the performance of the three classifiers (by repeating steps (a) and (b) in the previous question). Explain your choice of predictors.

   Solution: We choose to add checking status and other parties after running a forward selection.

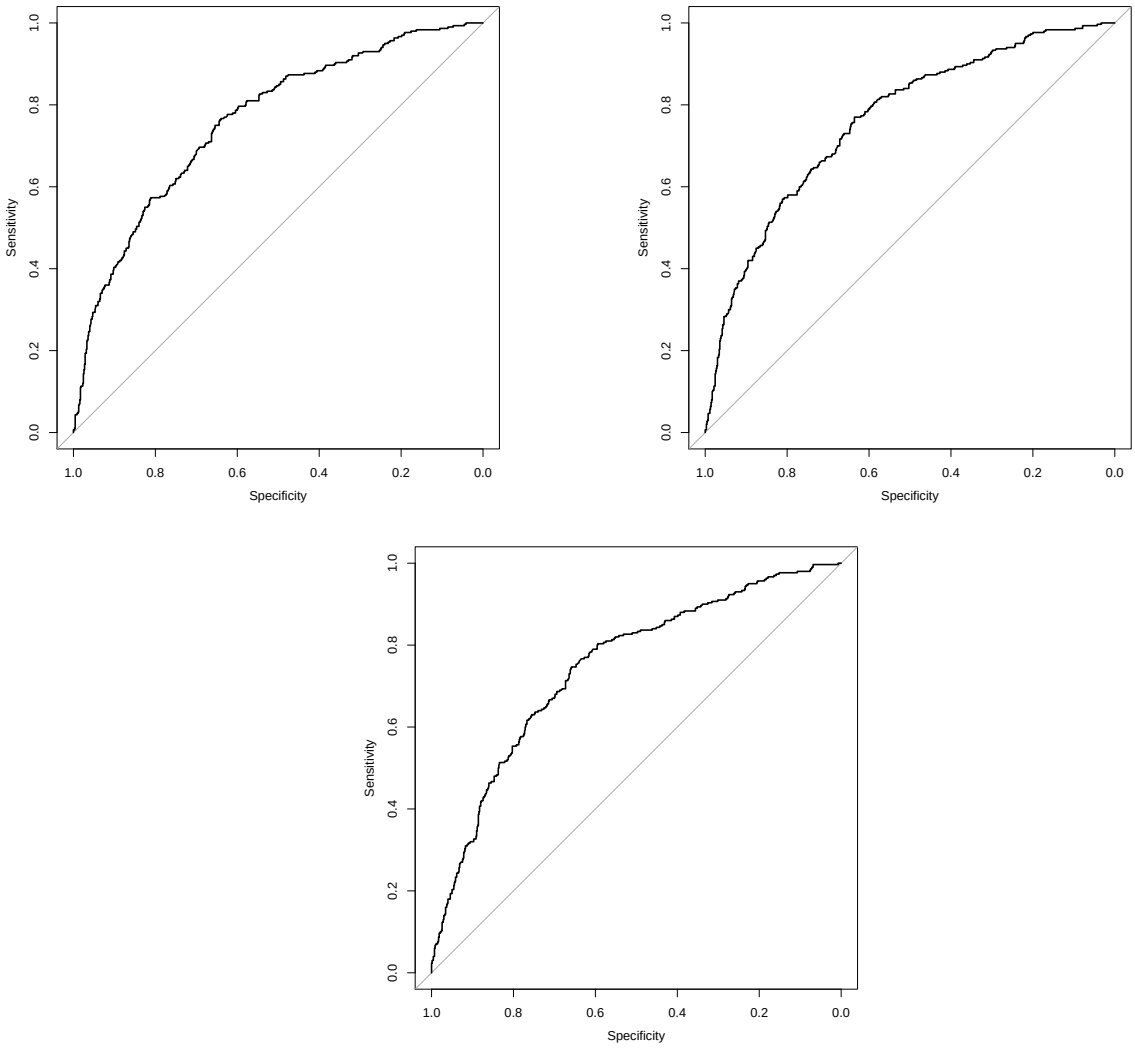
   We reran all three classifiers and got

	0	1	overall	type 1	type 2
logistic					
0	627	159			
1	73	141	0.232	0.1043	0.53
lda	0	1			
0	629	152			
1	71	148	0.223	0.1043	0.5067
qda	0	1			
0	548	93			
1	152	207	0.245	0.2171	0.31

   and the ROC curves have respective AUC of 0.802, 0.8007, 0.7927 and are given in Figure 3.

5. Suppose the credit card issuer assesses that it looses 4 times more money for each "bad credit" that is incorrectly identified (false negative) as it does for a "good credit" that is incorrectly identified (false positive). Based on this, and using your results based on logistic regression with your choice of predictors from the previous question, propose a decision threshold to identify a "bad credit" (via the estimated probability $\hat{P}(Y = \text{bad}|X)$) that minimizes the financial loss due

Figure 2: Logistic (left), lda (right) and qda ROC curve (bottom) with 5 predictors



to wrong classification. Then for all three methods, compute the overall (training) error rate, the type-I error and the type-II error.

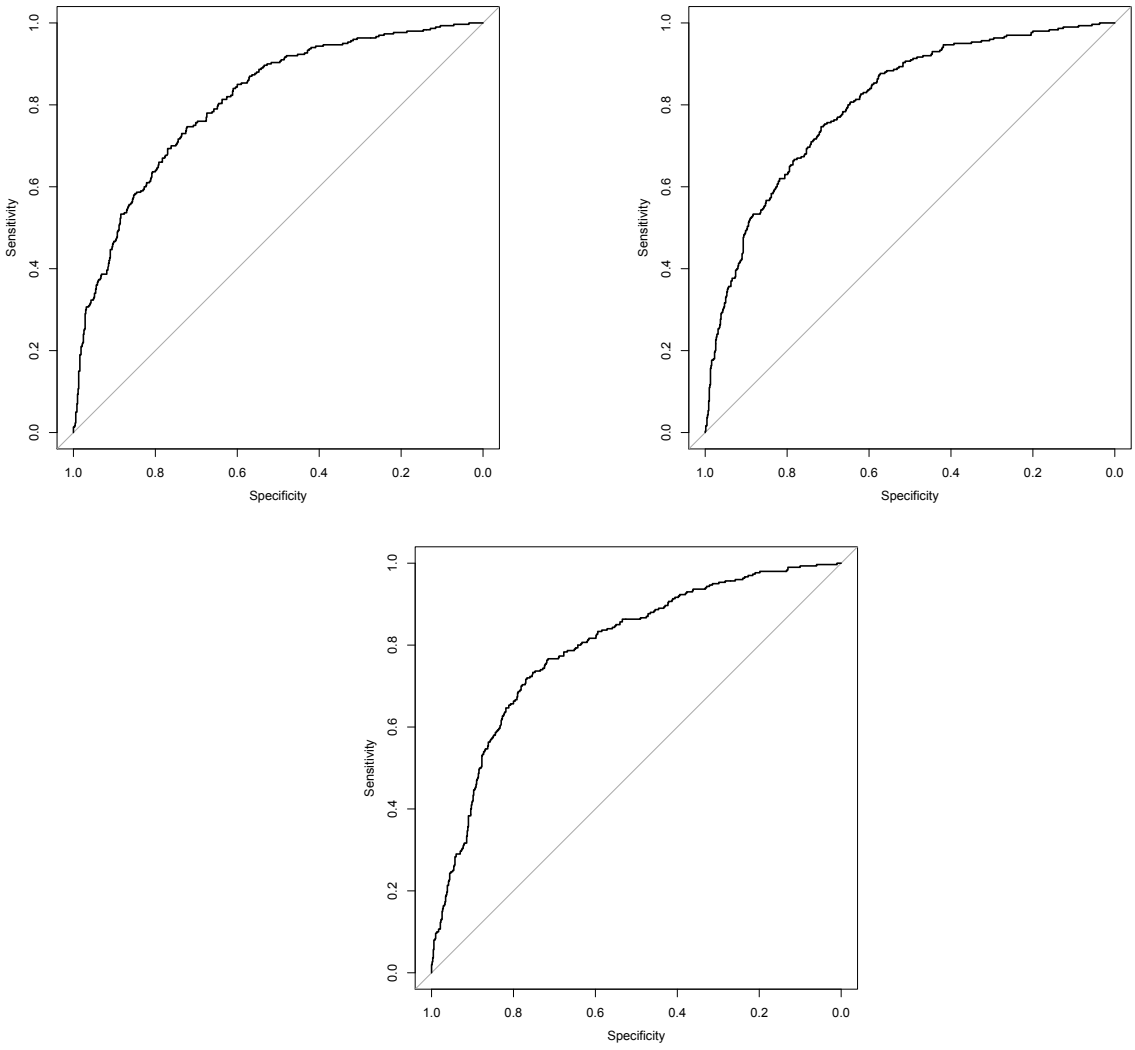
Solution: using logistic regression and the above cost function, we obtain a threshold of 0.167.

We can re-run the prediction obtained from the three methods with this threshold and get

	overall	type-1	type-2
logistic	0.359	0.4686	0.1033
lda	0.356	0.4586	0.1167
qda	0.292	0.3229	0.22

6. As an alternative to the previous question, suppose now that you wish to take a more conservative view point that does not rely on the above cost function, and instead what to choose the threshold that maximize Youden's J statistic (or equivalently, that maximizes the sum of the specificity and sensitivity). As in the previous question, compute the overall (training) error rate, the type-I

Figure 3: Logistic (left), lda (right) and qda ROC curve (bottom) with 7 predictors



error and the type-II error for all three methods using this new threshold. Comment on the results obtained using the three different thresholds used so far.

Solution: We find that the optimal threshold in that case is 0.295. We recompute the errors and get

	overall	type-1	type-2
logistic	0.271	0.2786	0.2533
lda	0.277	0.2814	0.2667
qda	0.268	0.2714	0.26

Overall, we see that using the threshold of 0.5 gives the lowest overall error, but find that for a small increase in the overall error, using the threshold of 0.295 that optimizes Youden's statistic provides a significant decrease in the type-2 error for logistic regression and LDA (going from 0.53 and 0.5067 to 0.253 and 0.267, respectively). The threshold of 0.167 provides the smallest type-2 error but at the expense of a large increase in the type-2 error.

7. Using again what you believe is the best set of predictors from part 4, use 10-fold cross-validation to estimate the overall error rate and type-II error. Is it consistent with the training error rate? Make sure to make your comparisons using the same threshold.

   Solution: we use the threshold of 0.295 to answer this. We use the function cv.glm (from the library boot) for logistic regression, and using the 0.295 threshold we get an overall error rate estimated to be 0.297 using 10-fold cv, and get 0.2776 for type-2 error. For lda we get 0.291 for the overall error rate and 0.2981 for the type-2 error. For qda, if we try all 7 covariates from part 4 we get an error of rank deficiency. Recall that both lda and qda are assuming all predictors are normally distributed, and as such when using these methods on data where some predictors are categorical, we are violating these assumptions and run the risk of having problems such as this one. I realize that some students determined it was not possible to answer this question for that reason and will not penalize them if they did so. Unfortunately I had not foreseen this problem when preparing the assignment. If alerted to this earlier I would have suggested the following approach: remove problematic categorical predictor(s) to avoid the rank deficiency. After a few trials we see that removing "purpose" solves the problem. Using CV we get an estimated overall error of 0.29 and 0.3233. The corresponding error on the training sets are 0.263 and 0.2867.

   Hence in all three cases the training errors are slightly underestimated, as expected.

8. Next you wish to use the K nearest neighbours (KNN) as a classifier for this problem, using the three predictors age, duration and credit_amount. Use a random sample of 750 observations as your training set and for each of $K = 1, 3, 5$, apply the KNN approach. Produce the confusion table, determine the overall error rate, Type-I error and Type-II error. Which choice of K seems the best?

   Solution: we have used the 7 quantitative predictors and got

$K=1$		0	1	overall	type-1	type-2
	0	132	52			
	1	30	36	0.436	0.3631	0.5909
$K=3$		0	1			
	0	133	55			
	1	29	33	0.400	0.2043	0.7045
$K=5$		0	1			
	0	141	61			
	1	21	27	0.364	0.1171	0.7614

   So from these results, choosing $K = 5$ has the smallest prediction and type-1 errors, but $K = 1$ has the smallest type-2 error.

9. Now use 10-fold cross validation to estimate the overall error rate and Type-II error for each of $K = 1, 3, 5$. Do you reach the same conclusion as in the previous question about the choice of K?

   Solution: Using cv we get

	$K=1$	$K=3$	$K=5$
overall	0.385	0.348	0.342
type-2	0.593	0.6767	0.76

   Based on the overall rate, using $K = 5$ still seems the best choice. The error rates are consistent with what we obtained before.

10. What would be your final recommendation for the best classifier to use for this problem?

    Solution: Logistic regression would be my suggestion for this problem. It has good interpretative value compared to KNN, and unlike LDA and QDA, it doesn't rely on assumptions that are violated. The correspondng prediction errors are competitive with other methods.