Case 3 report

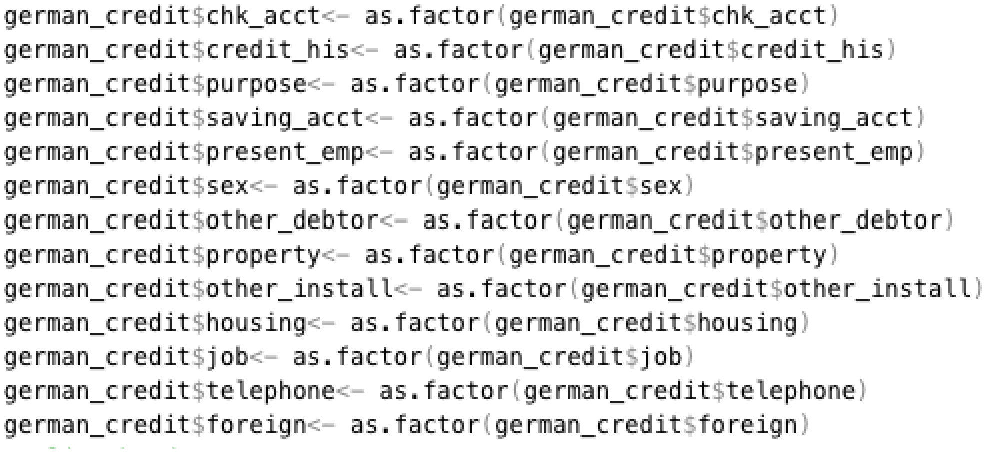
Team 5
Team member names: Branden Ho, Anna Nguyen

#(i) Random sample a training data set that contains 80% of original data points. Start with exploratory data analysis on the training data. Fit a logistic regression model and evaluate the model fitting.

	AIC	**BIC**	**Deviance**
Logistic Regression Model	802.9919	1032.538	704.9919

As we can see here with an AIC of 802 and a BIC of 1032 we can conclude that the regular model has many unnecessary variables to the model and that is the reason that the BIC is so much higher than the AIC.

We also found 13 variables that are considered categorical variables and we converted them to factors.



```
german_credit$chk_acct<- as.factor(german_credit$chk_acct)
german_credit$credit_his<- as.factor(german_credit$credit_his)
german_credit$purpose<- as.factor(german_credit$purpose)
german_credit$saving_acct<- as.factor(german_credit$saving_acct)
german_credit$present_emp<- as.factor(german_credit$present_emp)
german_credit$sex<- as.factor(german_credit$sex)
german_credit$other_debtor<- as.factor(german_credit$other_debtor)
german_credit$property<- as.factor(german_credit$property)
german_credit$other_install<- as.factor(german_credit$other_install)
german_credit$housing<- as.factor(german_credit$housing)
german_credit$job<- as.factor(german_credit$job)
german_credit$telephone<- as.factor(german_credit$telephone)
german_credit$foreign<- as.factor(german_credit$foreign)
```

#(ii) Find a best model using logistic regression with AIC and BIC. Draw ROC curve, report the AUC, and present the misclassification rate table of your final model.

	AIC	**BIC**	**Deviance**

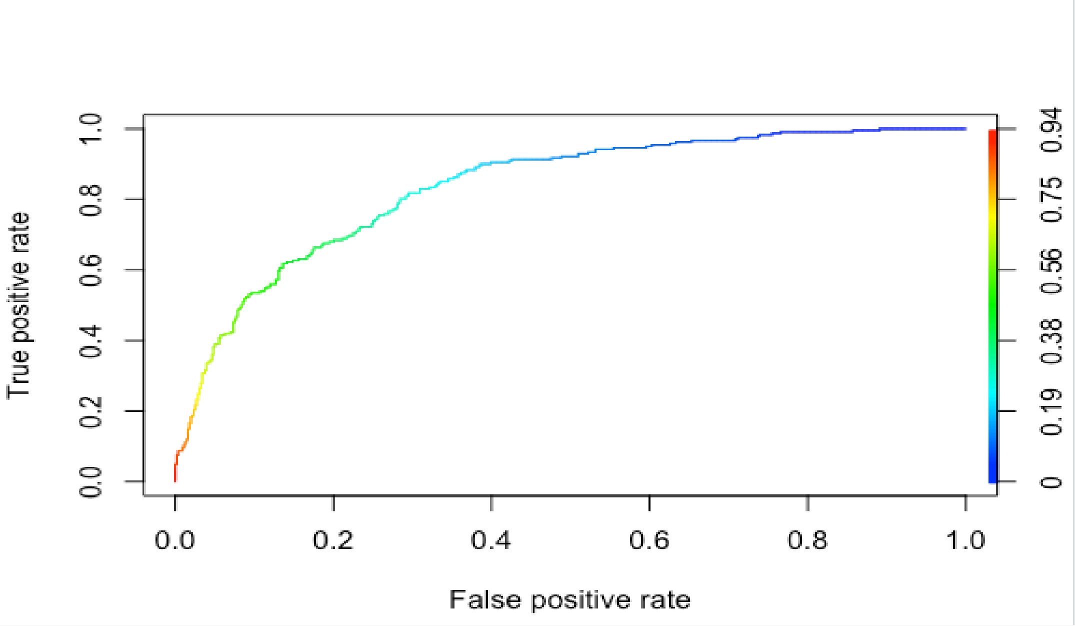
Backwards stepwise AIC Logistic Regression Model	788.5134	952.4748	718.5134

	AIC	**BIC**	**Deviance**
Backwards stepwise BIC Logistic Regression Model	832.522	874.6835	814.522

From the two tables above we can see that the backwards stepwise model built from AIC has better numbers overall. Due to this we have chosen the AIC model as our best fit model.

	AUC
80% Training Data	**0.8341808**

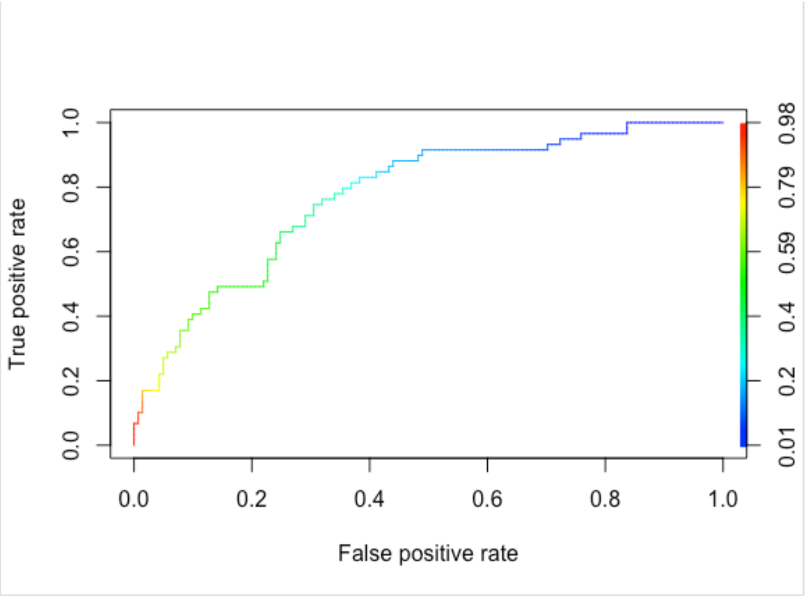
=> Since our AUC is greater than 0.7, this is an acceptable discriminatory power



Pred_response >pcut	Predicted	Predicted
Truth	0	1
0	315	244
1	21	220

This is our ROC curve and misclassification table using asymmetric cost. These are both tools to help us understand how much misclassification is going on with the use of our model and how our cut off point affects that. In the graph we can see the gradual increase in false positives become much steeper once we past 0.56. Also, this is a good ROC curve because it nearly hugs the top left corner.

#(iii) Test the out-of-sample performance. Using a final logistic linear model built from (ii) on the 80% of original data, test with the remaining 20% testing data. (Try predict() function in R.) Report out-of-sample AUC and misclassification rate.



	AUC
20% Testing data	**0.7742517**

=> Since our AUC is greater than 0.7, this is an acceptable discriminatory power

Misclassification rate (our goal is to	**0.33125**

minimize misclassification cost)	
Asymmetric misclassification rate (using asymmetric cost)	**0.43625**

AUC is used to show a summary of the ROC curve (good curve because it nearly hugs the top left corner) this number being 0.77 shows that our model is doing a decent job of classifying our data into the correct predictions, but as we see with the misclassification rate of 0.33, the model is not nearly perfect and still misclassified people 33% of the time. Also, with the asymmetric misclassification rate of 0.44, it appears that 44% of classifications were incorrect as well.

(iv) Cross validation. Use 5-fold cross validation. (Try cv.glm() function in R on the ORIGINAL data.) Does (iv) yield similar answer as (iii)? Make sure that you specify the right cost functions.

Cross-validation delta value	**0.5108**

Asymmetric misclassification rate (using asymmetric cost)	**0.43625 (we use in-sample misclassification matrix table, so this is wrong to compare with delta value)**

=> the cross-validation delta value is the averaged model error over the testing dataset. Hence, the misclassification error appears to be 51%, which is greater than the asymmetric misclassification rate. Therefore, the cross-validation predicted a greater error rate for the testing data.

(v) Now repeat previous steps for another random sample (that is, to draw another training datsa set with 80% of original data, and the rest 20% as testing; or you can try 90% training vs. 10% left as testing). Do you get similar results? What's your conclusion?

AIC	803.6344
BIC	972.2804
Deviance	731.6344

AUC (in-sample)	0.8277141
AUC (out-of-sample)	0.7953869

Pred_response >pcut	Predicted	Predicted
Truth	0	1
0	302	254
1	22	222

Misclassification rate (our goal is to specify asymmetric cost)	**0.476**

Cross-validation delta value	**0.509**

Asymmetric misclassification rate (using asymmetric cost)	**0.455**

=> the AIC and deviance we got this time is higher than the AIC and deviance from previous steps. Also, our in-sample AUC is smaller than the previous in-sample AUC, but our out-of-sample AUC is greater than the previous out-of-sample AUC. Nevertheless, we got a good AUC as they are at an acceptable discriminatory power level and they have done a good job at evaluating our best model that we selected from backward selection process.
=> our cross-validation delta value is still higher than the asymmetric rate, which means that it predicted a greater error rate of 51% for the testing data. Our misclassification and asymmetric misclassification rate reported a misclassification error rate of 48% and 46%, respectively.