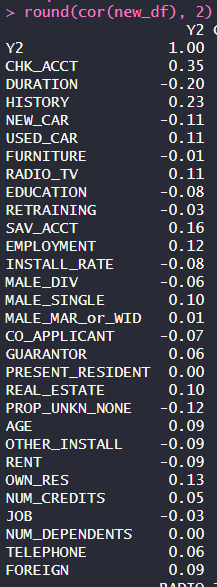
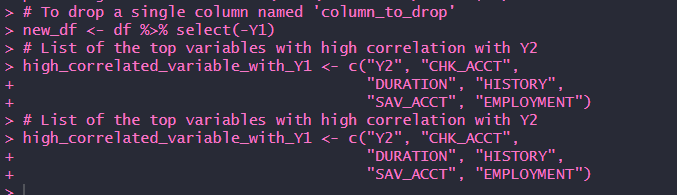
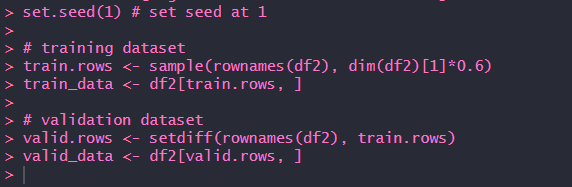
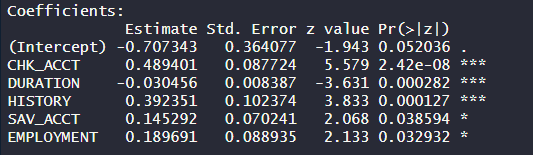
**1. Explain what data are you using**  
I have loaded the German Credit dataset into a data frame called 'df'. This data frame includes **912** observations, with each observation corresponding to a row, and **30** variables, with each variable corresponding to a column.   
  
Number of observations, number of variables  
  
  
What is Y2  
Y2 is the binary dependent variable in a logistic regression model, indicating an individual's credit rating: '0' for bad and '1' for good.  
  
What percent of Y2 is =1.  
  

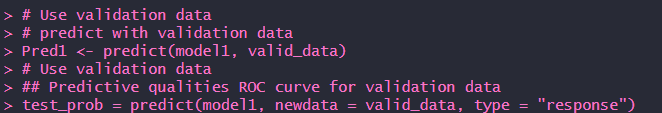
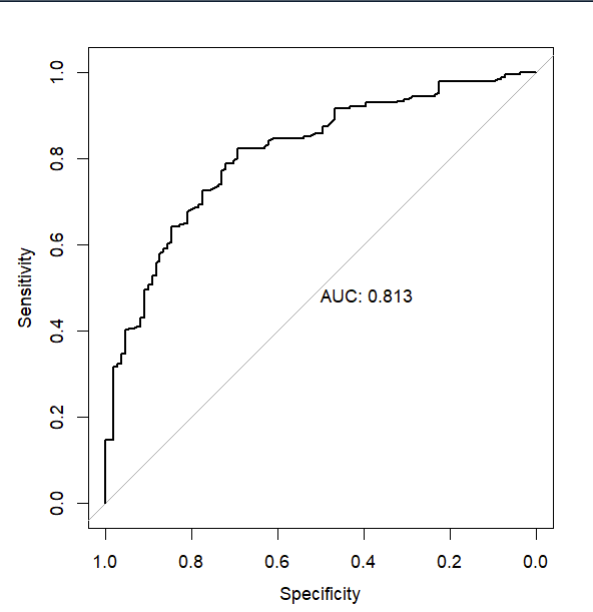

2. Present correlation matrix with Y2 and the rest of the variables (continuous or dummy).   
If you have categorical variables - create dummy variables instead. Drop Y1 variable.  
List the top variables with high correlation with Y2 (positive or negative). About 5-10 should be good.  
  
Give the selected variables labels/explanation of what they are.   


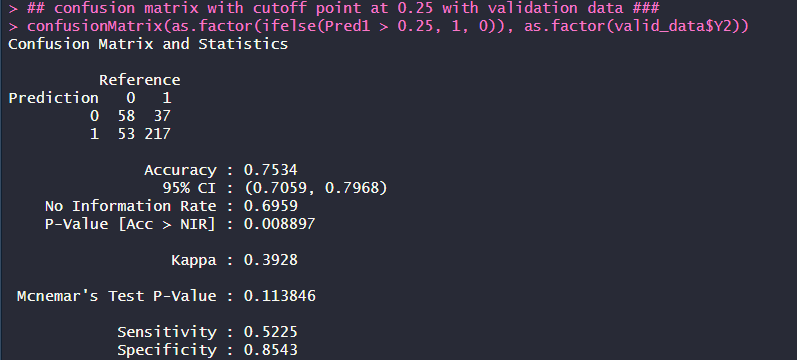
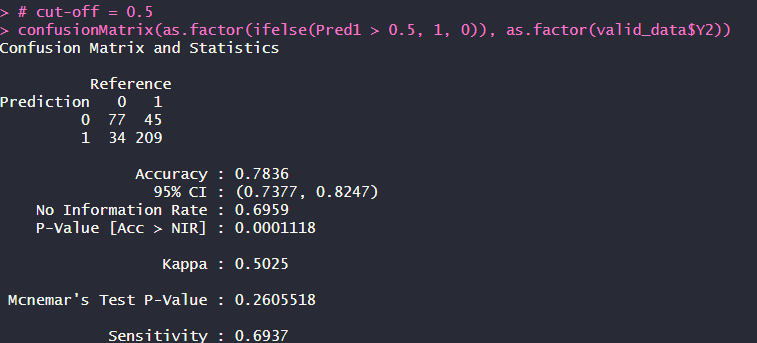
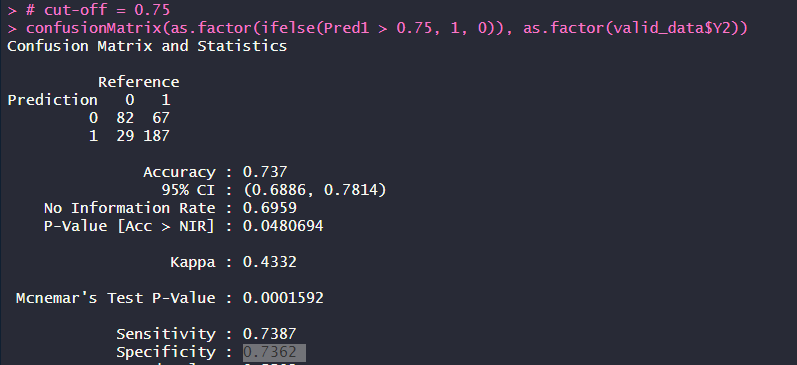
**3. Create a new dataset with Y2 and the top variables with high correlation with Y2 (positive or negative).**

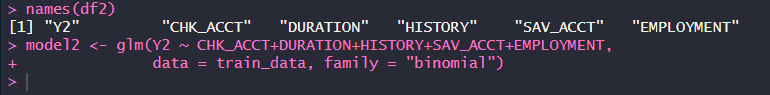
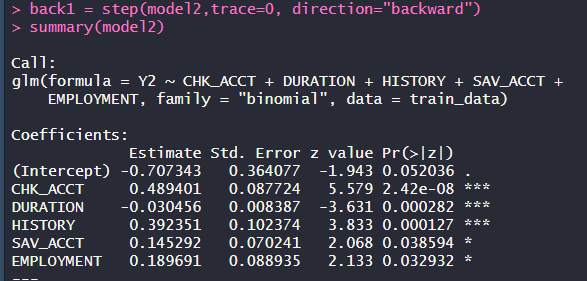
  
  
  

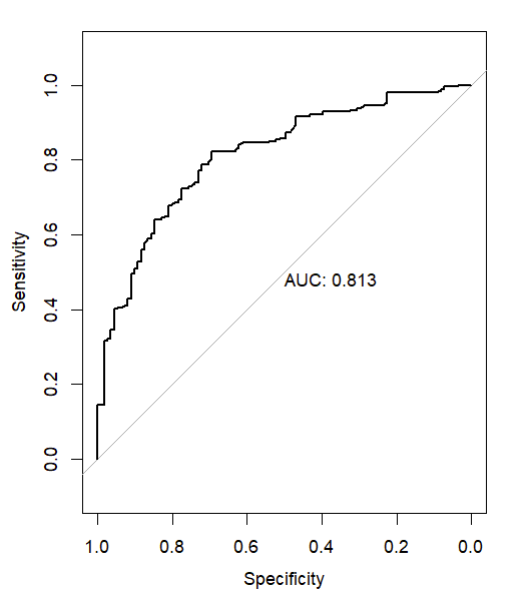
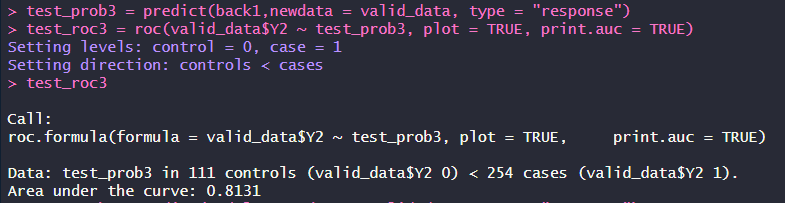

**4**. **Partition the data in 60% training and 40% validation (use seed).**  
  


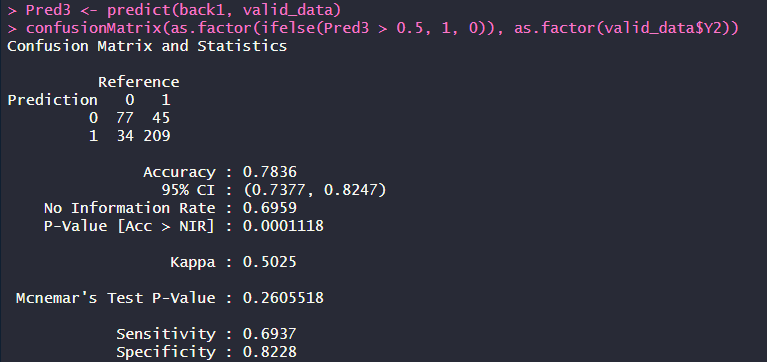
5. **Model 1.** Estimate logistic regression model with Y2 as dependent variable and all available variables with **training** data.   
  
  
  
**Comment of the results.**  
The output indicates that all variables except the intercept are statistically significant in predicting the response variable. Lower p-values (\*\*\* for p<0.001, \*\* for p<0.01, \* for p<0.05) suggest strong evidence against the null hypothesis, indicating that these predictors have a statistically significant association with the outcome.

**6. Evaluate the accuracy of Model 1 (report area under ROC and ROC graph) with validation data.**   
  
  
  
  
  
**Comment of the results.**  
The takeaway from the ROC curve image is that the model has good predictive power. With an AUC of 0.813, the model is much better than random guessing at distinguishing between the two classes in the binary classification task. This suggests that the model could be useful in practical applications where such predictions are needed.

**7. Present separate confusion matrices, overall accuracy, sensitivity and specificity for 3 cut-off points for Model 1:**  
0.25, 0.5, and 0.75.   
  
The model correctly predicts 75% of outcomes and is better at identifying true negatives (85%) than true positives (52%)  
  
  
With 78% overall accuracy, this model is more effective at detecting true positives (69%) than model a, with fairly good identification of true negatives (82%).  
  
  
This model has a balanced accuracy of 73.7%, with similar capability in identifying true positives (73.87%) and true negatives (73.62%).

**8. Model 2. Estimate logistic regression model with Y2 as dependent variable and a stepwise selection of the variables with training data.**   
  
  
  
  
  
  
  
  
**Comment of the results.**  
Variables with lower p-values indicate a stronger statistical significance. Variables CHK\_ACCT, DURATION, HISTORY, SAV\_ACCT, and EMPLOYMENT all have asterisks, signifying that they have a statistically significant relationship with the dependent variable Y2. The more asterisks next to the coefficients, the more confident we can be that the relationship is not due to random chance, with three asterisks representing very strong evidence (p < 0.001).

**9. Evaluate the accuracy of Model 2 (report area under ROC and ROC graph) with validation data. Present the confusion matrix**  
  
  


**Overall accuracy, sensitivity and specificity for cut-off point of 0.5.**   
  
  
**Comment of the results**  
This confusion matrix shows that the model's overall accuracy is 78.36%, meaning it correctly predicts the outcome in about 78 out of 100 cases. The sensitivity, or the model's ability to identify true positives, is 69.37%, indicating it correctly identifies approximately 69 positive cases out of 100. The specificity is 82.28%, reflecting that the model correctly recognizes about 82 negative cases out of 100, avoiding false alarms.

**10. Compare the two models by Area Under ROC.**An Area Under the Curve (AUC) of 0.8131 for both Models means that both models have a good ability to differentiate between the positive and negative classes in the dataset.