Assignment 1: Decision Tree analysis -- German Credit Data
Due: Feb 14th

The German Credit dataset has data on 1000 past credit applicants, described by 30 variables. Each applicant is also rated as "Good" or "Bad" credit (encoded as 1 and 0 respectively in the Response variable). The GermanCredit.xls file contains the variable descriptions and the data. New applicants for credit can be evaluated on these 30 variables. We would like to develop a credit scoring rule that can be used to help determine whether a new applicant presents a good or bad credit risk. Here, we will attempt to obtain a decision tree based model to determine if new applicants present a good or bad credit risk.

The original data has been transformed to ease analysis in this initial assignment. It is informative to compare the original and transformed data descriptions to see how different variables have been transformed. The original variables are given below.

Assignment questions:

1. Explore the data: What is the proportion of "Good" to "Bad" cases? Are there any missing values – how do you handle these? Obtain descriptions of the predictor (independent) variables – mean, standard deviations, etc. for real-values attributes, frequencies of different category values. Examine variable plots.
   Do you notice 'bad' credit cases to be more prevalent in certain value-ranges of specific variables, and is this what one might expect (or is it more of a surprise)?
   What are certain interesting variables and relationships (why 'interesting')?
   From the data exploration, which variables do you think will be most relevant for the outcome of interest, and why?

2. We will first focus on a descriptive model – i.e. assume we are not interested in prediction.
   (a) Develop a decision tree on the full data. What decision tree node parameters do you use to get a good model (and why?)
   (b)Which variables are important to differentiate "good" from "bad" cases – and how do you determine these? Does this match your expectations (from the your response in Question 1)?
   (c)What levels of accuracy/error are obtained? What is the accuracy on the "good" and "bad" cases? Obtain and interpret the lift chart.
   Do you think this is a reliable (robust?) description, and why.

   We next consider developing a model for prediction. For this, we should divide the data into Training and Validation sets. Consider a partition of the data into 50% for Training and 50% for Test
   (a) Develop decision trees using the rpart package. What model performance do you obtain? Consider performance based on overall accuracy/error and on the 'good' and 'bad' credit cases – explain which performance measures, like recall, precision, sensitivity, etc. you use and why. Also consider lift, ROC and AUC.
   Is the model reliable (why or why not)?

   In developing the models above, change decision tree options as you find reasonable (for example, complexity parameter (cp), the minimum number of cases for split and at a leaf

node, the split criteria, etc.) - explain which parameters you experiment with and why. Report on if and how different parameters affect performance.
Describe the pruning method used here. How do you examine the effect of different values of cp, and how do you select the best pruned tree.
Which decision tree parameter values do you find to be useful for developing a good model.

(b) Consider another type of decision tree – C5.0 – experiment with the parameters till you get a 'good' model. Summarize the parameters and performance you obtain.
Also develop a set of rules from the decision tree, and compare performance.

Does performance differ across different types of decision tree learners? Compare models using accuracy, sensitivity, precision, recall, etc (as you find reasonable – you answer to Questions (a) above should clarify which performance measures you use and why). Also compare performance on lift, ROC curves and AUC.
How do the models obtained from these decision tree learners differ?

(c) Decision tree models are referred to as 'unstable' – in the sense that small differences in training data can give very different models. Examine the models and performance for different samples of the training/test data (by changing the random seed). Do you find your models to be unstable -- explain?

(d) Which variables are important for separating 'Good' from 'Bad' credit? Determine variable importance from the different 'best' trees. Are there similarities, differences?

(d) Consider partitions of the data into 70% for Training and 30% for Test, and 80% for Training and 20% for Test and report on model and performance comparisons (for the decision tree learners considered above).
In the earlier question, you had determined a set of decision tree parameters to work well. Do the same parameters give 'best' models across the 50-50, 70-30, 80-20 training-test splits? Are there similarities among the different models ….in, say, the upper part of the tree – and what does this indicate?
Is there any specific model you would prefer for implementation?

3. Consider the net profit (on average) of credit decisions as:
Accept applicant decision for an Actual "Good" case: 100DM, and
Accept applicant decision for an Actual "Bad" case: -500DM
This information can be used to determine the following costs for misclassification:

		Predicted	
Actual		Good	Bad
	Good	0	100DM
	Bad	500DM	0

(a) Use the misclassification costs to assess performance of a chosen model from Q 2 above. Compare model performance.
Examine how different cutoff values for classification threshold make a difference. Use the ROC curve to choose a classification threshold which you think will be better than the default 0.5. What is the best performance you find?
(b) Calculate and apply the 'theoretical' threshold and assess performance – what do you notice, and how does this relate to the answer from (a) above.

(c) Use misclassification costs to develop the tree models (rpart and C5.0) – are the trees here different than ones obtained earlier? Compare performance of these two new models with those obtained earlier (in part 3a, b above).

4. Let's examine your 'best' decision tree model obtained. What is the tree depth? And how many nodes does it have? What are the important variables for classifying "Good' vs 'Bad' credit?
   Identify two relatively pure leaf nodes. What are the 'probabilities for 'Good' and 'Bad' in these nodes?

5. The predicted probabilities can be used to determine how the model may be implemented. We can sort the data from high to low on predicted probability of "good" credit risk. Then, going down the cases from high to low probabilities, one may be able to determine an appropriate cutoff probability – values above this can be considered acceptable credit risk. The use of cost figures given above can help in this analysis.

   For this, first sort the validation data on predicted probability. Then, for each validation case, calculate the actual cost/benefit of extending credit. Add a separate column for the cumulative net cost/benefit.

   How far into the validation data would you go to get maximum net benefit? In using this model to score future credit applicants, what cutoff value for predicted probability would you recommend? Provide appropriate performance values to back up your recommendation.

B. Data description

Var. #	Variable Name	Description	Variable Type	Description
1.	OBS#	Observation No.	Categorical	
2.	CHK_ACCT	Checking account status	Categorical	0 : < 0 DM 1: 0 < ...< 200 DM 2 : => 200 DM 3: no checking account
3.	DURATION	Duration of credit in months	Numerical	
4.	HISTORY	Credit history	Categorical	0: no credits taken 1: all credits at this bank paid back duly 2: existing credits paid back duly till now 3: delay in paying off in the past 4: critical account
5.	NEW_CAR	Purpose of credit	Binary	car (new) 0: No, 1: Yes
6.	USED_CAR	Purpose of credit	Binary	car (used) 0: No, 1: Yes
7.	FURNITURE	Purpose of credit	Binary	furniture/equipment 0: No, 1: Yes
8.	RADIO/TV	Purpose of credit	Binary	radio/television 0: No, 1: Yes
9.	EDUCATION	Purpose of credit	Binary	education 0: No, 1: Yes
10.	RETRAINING	Purpose of credit	Binary	retraining 0: No, 1: Yes
11.	AMOUNT	Credit amount	Numerical	
12.	SAV_ACCT	Average balance in savings account	Categorical	0 : < 100 DM 1 : 100<= ... < 500 DM 2 : 500<= ... < 1000 DM 3 : =>1000 DM 4 : unknown/ no savings account
13.	EMPLOYMENT	Present employment since	Categorical	0 : unemployed 1: < 1 year 2 : 1 <= ... < 4 years 3 : 4 <=... < 7 years 4 : >= 7 years
14.	INSTALL_RATE	Installment rate as % of disposable income	Numerical	
15.	MALE_DIV	Applicant is male and divorced	Binary	0: No, 1: Yes
16.	MALE_SINGLE	Applicant is male and single	Binary	0: No, 1: Yes
17.	MALE_MAR_WID	Applicant is male and married or a widower	Binary	0: No, 1: Yes
18.	CO-APPLICANT	Application has a co-applicant	Binary	0: No, 1: Yes
19.	GUARANTOR	Applicant has a guarantor	Binary	0: No, 1: Yes
20.	PRESENT_RESIDENT	Present resident since - years	Categorical	0: <= 1 year 1<…<=2 years 2<…<=3 years 3:>4years
21.	REAL_ESTATE	Applicant owns real estate	Binary	0: No, 1: Yes
22.	PROP_UNKN_NONE	Applicant owns no property (or unknown)	Binary	0: No, 1: Yes
23.	AGE	Age in years	Numerical	
24.	OTHER_INSTALL	Applicant has other installment plan	Binary	0: No, 1: Yes

		credit		
25.	RENT	Applicant rents	Binary	0: No, 1: Yes
26.	OWN_RES	Applicant owns residence	Binary	0: No, 1: Yes
27.	NUM_CREDITS	Number of existing credits at this bank	Numerical	
28.	JOB	Nature of job	Categorical	0 : unemployed/ unskilled - non-resident 1 : unskilled - resident 2 : skilled employee / official 3 : management/ self-employed/highly qualified employee/ officer
29.	NUM_DEPENDENTS	Number of people for whom liable to provide maintenance	Numerical	
30.	TELEPHONE	Applicant has phone in his or her name	Binary	0: No, 1: Yes
31.	FOREIGN	Foreign worker	Binary	0: No, 1: Yes
32	RESPONSE	Credit rating is good	Binary	0: No, 1: Yes

****

C. ORIGINAL DATA DESCRIPTION

Description of the German credit dataset.

1. Title: German Credit data

2. Source Information

Professor Dr. Hans Hofmann
Institut f"ur Statistik und "Okonometrie
Universit"at Hamburg
FB Wirtschaftswissenschaften
Von-Melle-Park 5
2000 Hamburg 13

3. Number of Instances: 1000

Two datasets are provided. the original dataset, in the form provided by Prof. Hofmann, contains categorical/symbolic attributes and is in the file "german.data".

For algorithms that need numerical attributes, Strathclyde University produced the file "german.data-numeric". This file has been edited and several indicator variables added to make it suitable for algorithms which cannot cope with categorical variables. Several attributes that are ordered categorical (such as attribute 17) have been coded as integer. This was the form used by StatLog.

6. Number of Attributes german: 20 (7 numerical, 13 categorical)
   Number of Attributes german.numer: 24 (24 numerical)

7. Attribute description for german

Attribute 1: (qualitative)
Status of existing checking account
A11 : ... < 0 DM
A12 : 0 <= ... < 200 DM
A13 : ... >= 200 DM /
salary assignments for at least 1 year
A14 : no checking account

Attribute 2: (numerical)
Duration in month

Attribute 3: (qualitative)
Credit history
A30 : no credits taken/
all credits paid back duly
A31 : all credits at this bank paid back duly
A32 : existing credits paid back duly till now
A33 : delay in paying off in the past
A34 : critical account/
other credits existing (not at this bank)

```
Attribute 4:  (qualitative)
            Purpose
            A40 : car (new)
            A41 : car (used)
            A42 : furniture/equipment
            A43 : radio/television
            A44 : domestic appliances
            A45 : repairs
            A46 : education
            A47 : (vacation - does not exist?)
            A48 : retraining
            A49 : business
            A410 : others

Attribute 5:  (numerical)
            Credit amount

Attibute 6:  (qualitative)
            Savings account/bonds
            A61 :          ... <  100 DM
            A62 :   100 <= ... <  500 DM
            A63 :   500 <= ... < 1000 DM
            A64 :          .. >= 1000 DM
              A65 :   unknown/ no savings account

Attribute 7:  (qualitative)
            Present employment since
            A71 : unemployed
            A72 :       ... < 1 year
            A73 : 1  <= ... < 4 years
            A74 : 4  <= ... < 7 years
            A75 :       .. >= 7 years

Attribute 8:  (numerical)
            Installment rate in percentage of disposable income

Attribute 9:  (qualitative)
            Personal status and sex
            A91 : male   : divorced/separated
            A92 : female : divorced/separated/married
              A93 : male   : single
            A94 : male   : married/widowed
            A95 : female : single

Attribute 10: (qualitative)
            Other debtors / guarantors
            A101 : none
            A102 : co-applicant
            A103 : guarantor

Attribute 11: (numerical)
            Present residence since

Attribute 12: (qualitative)
            Property
            A121 : real estate
```

```
            A122 : if not A121 : building society savings agreement/
                           life insurance
            A123 : if not A121/A122 : car or other, not in attribute 6
            A124 : unknown / no property

Attribute 13: (numerical)
            Age in years

Attribute 14: (qualitative)
            Other installment plans
            A141 : bank
            A142 : stores
            A143 : none

Attribute 15: (qualitative)
            Housing
            A151 : rent
            A152 : own
            A153 : for free

Attribute 16: (numerical)
              Number of existing credits at this bank

Attribute 17: (qualitative)
            Job
            A171 : unemployed/ unskilled  - non-resident
            A172 : unskilled - resident
            A173 : skilled employee / official
            A174 : management/ self-employed/
                 highly qualified employee/ officer

Attribute 18: (numerical)
            Number of people being liable to provide maintenance for

Attribute 19: (qualitative)
            Telephone
            A191 : none
            A192 : yes, registered under the customers name

Attribute 20: (qualitative)
            foreign worker
            A201 : yes
            A202 : no

**********************************************************************
```