Week 6
Classification

Examination Assignments

Regarding the Assignments

This document outlines the **individual** examination assignments for this week. We strongly recommend that you work through the training assignments for this week before progressing to the examination assignments.

Week 6 – Examination Assignment

You will be working with the dataset: "**bankloan.csv".** This dataset contains data from a customer survey from a bank regarding personal loans. Appendix 1 gives an overview of the variables

Introduction

You are an employee of a large bank and just started the third year of your employment. One morning your boss knocks at your office door. She asks you "Have you got a few minutes?" to which you reply "Yes". The boss continues.

"You might have heard that we have been struggling lately with our approval rate for personal loans to our customers. To be frank, we have no accurate model for when to approve a loan application or not. During last month we collected information from 5000 customers who either got a loan approved or got rejected. If I remember correctly, you took a course in data analysis?"

You reply.

"Yes! During my education I took a course K0021N called Applied Data Analysis for Engineers.

Boss.

"Wonderful! Can you do a logistic regression model on the data we collected and suggest a better cutoff for loan approval?"

You.

"Yes, I get right on it!

Examination assignment

You should complete the following tasks and/or answer the following questions:

Task 1:

Do a logistic regression analysis on the bankloan data using the factor "*Personal.Loan*" as the response. In your analysis you must do the following:

- Data cleansing
- Descriptive analysis
- Cross-validation (using the 70/30% split)
- Initial model
- Stepwise regression
- Model adequacy checking
- Present odds-ratio for the variables (document the meaning of the values)
- Initial confusion matrix
- ROC and AUC
- Select cutoff (adequately motivated)
- Use the selected cutoff for a new confusion matrix

For everything in the list above, do not just include a table/graph and move to the next task. You are expected to comment on every table and graph included in your analysis!

Task 2:

After your model is finalized, 15 new customers submit a loan application. Use your model and cutoff to approve or reject these 15 applicants.

ID	Age	Experience	Income	ZIP.Code	Family	CCAvg	Education	Mortgage	Securities.Account	CD.Account	Online	CreditCard
1	38	12	48	95617	4	0.2	3	0	0	0	1	0
2	58	32	73	94523	2	0.7	2	0	0	0	1	1
3	39	14	155	94577	2	3.9	1	0	0	0	1	0
4	64	37	138	94709	2	2.8	2	0	0	0	1	0
5	33	6	78	90250	4	2.0	2	119	1	0	1	0
6	24	-2	150	94720	2	2.0	1	0	0	0	1	0
7	46	20	91	92521	4	2.6	3	0	0	0	0	0
8	52	28	178	92647	3	5.4	3	147	0	0	1	0
9	44	19	74	90041	4	1.9	3	0	0	0	0	0
10	56	30	111	93106	4	0.3	1	372	1	1	1	0
11	53	29	118	94066	2	0.3	1	0	0	0	1	0
12	60	35	48	94538	3	1.5	1	0	0	0	1	0
13	46	22	125	94536	2	4.7	3	0	0	0	1	0
14	43	19	83	92691	4	2.0	3	0	0	0	1	0
15	61	35	74	91320	2	0.7	2	0	0	0	1	1

The new assignments at the examination session? If you have solved the above tasks, documented your solutions, and are able to run the code fast in the classroom you should be well-prepared to solve the additional assignment(s) that you will get at the examination session. In preparation for the examination assignments, it is also important that you and have worked through the training assignments. If you have done all this, **do not worry!**

Appendix 1 – Variables

- ID
  - Just an ID number of the applicant. Not to be used in the regression model
- Age
  - The age of the applicant
- Experience
  - The amount of experience the applicant has with the bank
  - There are a few negative numbers which (in my opinion) are difficult to interpret. Perhaps a negative experience represents a "bad" customer that the bank does not want to do business with. You are free to interpret the negative experience as you wish.
- Income
  - Yearly income of the applicant in k$
- Zip Code
  - The area where the applicant lives
- Family
  - Number of people in applicants' family (spouse and/or children)
- CC average
  - The average interest rate for credit cards for the applicant
- Education
  - A factor variable ranging from 1-3. 1=no high school diploma, 2=high school diploma and 3=higher education diploma.
- Mortgage
  - The amount in mortgage the applicant has in k$
- Personal loan
  - The response, 1=approved, 0=rejected
- Securities account
  - If the applicant was a securities account or not
- CD account
  - If the applicant has a Certificate of Deposit account or not
- Online
  - If the application was submitted online or not
- Credit Card
  - If the applicant has one (or more) credit card or not