Data Mining Assignment 1: Classification

Toon Calders
toon.calders@uantwerpen.be

Ngoc Quang Luong
ngoc.quang.luong@imec.be

Deadline: April 16th, 2023

1 The Assignment

Download the datasets **existing-customers.xlsx** and **potential-customers.xlsx** from BlackBoard.

Imagine the following scenario. You are the head of the data analytics team of an investment company and you have been asked by your manager to provide her with a list of people that should be targeted for a special promotion in order to attract as new customers. She suggests you to first go and talk to marketing to get a better idea of the problem.

While discussing the problem with your contacts at marketing, you learn the following observations:

- People with higher incomes (more than 50K) are more likely to react positively to the special promotion; 10% of them is likely to accept the offer, whereas only 5% of the lower income people reacts positively.
- The average return of a new customer highly depends on his or her income. Higher income people tend to generate more revenue; on average the profit for a high-income client is 980 Euro. Low income clients on the other hand, cost money on average; a low income customer will cost you on average 310 Euro.
- The cost of producing and mailing the promotional package is 10 Euro.

Furthermore, the following interesting datasets are available to support your analysis task:

- Recently the company bought demographic information that was obtained the data through a survey. Unfortunately, the data does not contain the income of people. Only for those people that are already clients of our company, the income is available. As a result you get the following two datasets: **potential-customers.xlsx** contains the demographic information of all people that are not yet clients, while **existing-customers.xlsx** contains the demographic information *and* whether or not the income of the person exceeds 50K for all clients of the company.

Solve this problem (provide the liust of people to send the promotion to) and give an estimate of the profit you expect when sending the promotion to the people you selected. The goal is to maximize revenue.

2 Deliverables

You are supposed to submit the following documents with your assignment:

- A short document describing how you solved the problem. Include the preprocessing steps you performed (e.g. what did you do with missing values, did you normalize the data, etc.), the classification methods tried out, how you evaluated your solution. The report should contain an estimation of the expected gain for the company for your selection (see next point). **Clearly indicate your estimate.** Present **a single estimate**, and explain it.
- A text file containing all rowIDs of the potential customers to whom you would send the promotion.
- Your code. The easiest way to share your code is for instance by adding a link to a public github repository.

3 Tools

You are free to use any tool you prefer. Recommended tools are: Knime[1] if you prefer a graphical tool, of sklearn[2] if you prefer a scripting-based solution. Both tools are sufficiently versatile to allow for perfectly solving the assignment. Both tools have excellent documentation.

A Data Set Description

The dataset contains the following attributes:

1. age: the age of an individual
   Integer greater than 0
2. workclass: a general term to represent the employment status of an individual
   Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. education: the highest level of education achieved by an individual.
   Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
4. education-num: the highest level of education achieved in numerical form.
   Integer greater than 0
5. marital-status: marital status of an individual. Married-civ-spouse corresponds to a civilian spouse while Married-AF-spouse is a spouse in the Armed Forces.
   Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

[1] `https://www.knime.com/`
[2] `https://scikit-learn.org/stable/`

6. occupation: the general type of occupation of an individual
   Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

7. relationship: represents what this individual is relative to others. For example an individual could be a Husband. Each entry only has one relationship attribute and is somewhat redundant with marital status.
   Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

8. race: Descriptions of an individual's race
   White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

9. sex: the biological sex of the individual
   Male, female

10. capital-gain: capital gains for an individual
    Integer greater than or equal to 0

11. capital-loss: capital loss for an individual
    Integer greater than or equal to 0

12. hours-per-week: the hours an individual has reported to work per week
    continuous

13. native-country: country of origin for an individual
    United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, &Tobago, Peru, Hong, Holand-Netherlands.

14. the label: whether or not an individual makes more than $50,000 annually.
    <= 50K, >50K