SET09120 Data Analytics 2020/21

Coursework II

DATA DESCRIPTION

The file credits.xlsx contains historic observations on 10 variables (attributes) for 1000 past applications for credit. Each applicant was given a rate of "good" (700 cases) or "bad" (300 cases) credit. Based on the applicant's profile, a bank can make reasonable decisions about whether or not to award a loan.

The table below shows the metadata about the dataset.

Table: Attributes and Values for the Credit Data

Attribute Name	Value	Code description
Case_no		Case number allocated to each applicant
	numerical	
checking_status		Status of existing current account
	< 0	Less than 0
	0<=X<200	Between 0 (inclusive) and 200
	>=200	Greater or equal to 200
	no checking	No current account in the bank
credit_history		Debt history
	no credits/ all paid	No debt taken or all debts paid back duly
	all paid	All debts at this bank paid back duly
	existing paid	Existing debts paid back duly till now
	delayed previously	Delay in paying off in the past
	critical/other existing credit	Critical account/other debts existing (not at this bank)
purpose		The purpose of a loan
	new car	New car
	used car	Used car
	furniture/equipment	furniture/equipment
	radio/tv	Radio/television
	domestic appliance	Domestic appliance
	Repairs	repairs
	education	education
	Vacation	holiday
	retraining	retraining
	business	business
	other	Other purposes
credit_amount		Debt amount
	numerical	
saving_status		Savings account/bonds
	<100	Less than 100
	100<=X<500	Between 100 (inclusive) and 500
	500<=X<1000	Between 500 (inclusive) and 1000

	>=1000	Greater or equal to 1000
	no known savings	unknown/no savings account
employment		Present employment since
	unemployed	unemployment
	< 1	Less than 1 year
	1<=X<4	Between 1 (inclusive) and 4 years
	4<=X<7	Between 4 (inclusive) and 7 years
	>=7	Greater or equal to 7 years
personal_status		Personal status and gender
	male div/sep	Male: divorced/separated
	female div/dep/mar	Female: divorced/separated/married
	male single	Male: single
	male mar/wid	Male: married/widowed
	female single	Female: single
age		Age in years
	numerical	
job		Job status
	unemp/unskilled non res	Unemployed/ unskilled – non-resident
	Unskilled resident	Unskilled - resident
	skilled	Skilled employee / official
	high qualif/self emp/mgmt	Management/self-employee/officer
class		Decision – good or bad
	good	Safe to provide a loan
	bad	Not safe to provide a loan

Like much of the data that companies store in data warehouses, this is genuine historical data recorded by a Germany bank, and much of the interest lies in trying to discover patterns within it. For example, a bank provides loans to customers who need the monetary resources to meet their individual goals. In exchange for loans the bank charges interest to customers. Repayment of the loan and interest is vital to the lending bank because the loaned money is the "raw materials" of their business, and the interest is the source of profit. How to increase the profit is a big question for the bank. The bank managers have only a vague idea about their customers, who is good (safe, offer a loan) and who is not (risky, don't offer any loan or offer a loan with caution, e.g., charge higher interest). Fortunately, the bank stores data about their customers, the status of existing current account, the saving status, credit history and job status, etc. Bank managers hope to improve their understanding of customers and seek specific actions to increase their profit. An analysis of the data with a discovery tool will be convincing for managers.

The data for this coursework is available at the module's Moodle site.

YOUR TASK

You are asked to use OpenRefine, Weka to conduct an exploratory data mining of this data, and to produce a SHORT report about what you discover from the data.

Tasks and mark allocations are as follows

1. Prepare and clean the data for analysing. At this stage, you are expected to undertake the following procedures: Understand Data and Prepare Data. At the preparation stage, you should clean, and convert the data from the XLSX format into the ARFF format that can be accepted by WEKA. This includes transforming data from one type to another in order to use some particular algorithms. For example, you might need to transform values of a particular attribute from nominal to numerical in order to use a regression algorithm, or from numerical to nominal in order to use association algorithms. Therefore, you are expected to prepare more than one version of data for the analysing.

   [20%]

2. Analyse the data with appropriate techniques/algorithms such as Classification, Regression, Association and Clustering algorithms. At this stage, you should find out some interesting patterns, such as which kind of applicants are likely to be safe to offer loans to, does skilled residents have any advantage? Will applicants' credit history help? It would be fine if algorithms from any three of the four categories are used: Classification, Regression, Association and Clustering. These patterns should be represented by rules (about 6 rules from each of the algorithms used), supported by statistical information, such as accuracy and coverage, if possible.

   (65%)

3. Based on your analysis in 2, summarise the overall findings in the data. Critically discuss and compare the algorithms used and draw an overall conclusion with justification about which techniques are most effective for making discoveries and gaining insights into the data.

   (15%)

Total [100%]

Beware of the fact that some of the algorithms only accept nominal (qualitative) attributes. As with the classic "beer and nappies" story from lectures, the results need some interpretation in the light of common sense and basic knowledge of what the data is actually about. Also remember that the coverage and accuracy of rules generated by each algorithm, if they are available, are important.

THE REPORT

Your report should be no longer than **NINE** pages of A4, using a 12pt font, but you should not paste actual program output into it. Your report should focus on the tasks

set in the above section. A good report should explain how and what you did in your experiment (enough to let an experienced user re-create what you get, e.g., clearly describe the way you performed, the errors identified and the corrections made accordingly when you were processing data cleaning; the way you perform any data transformation at the data preparation stage; at the analysing stage, mention the settings, the attributes involved in each play, mention any change to parameters etc.) and interpret and discuss the results generated. Any screenshots that are necessary can be put into an appendix which is not included in the NINE page limit. Also references for task 3 are expected, which can be listed in appendix.

Note

This coursework contributes 65% to the overall module assessment.

Collaboration and Plagiarism

This is an individual assessment. The work submitted should be entirely your own and will be checked against all other submissions by TurnitinUK.

Deliverables & Submission:-

The report, cleaned, and formatted datasets, including all datasets (in ARFF format) that are ready for the analysing tool WEKA should be zipped into a file called set09120cw_<your matric number> and uploaded to Moodle **on the same day of your submission**, as per instructions. For example, if your matriculation number is 42012345, your zipped file should be named set09120cw_42012345.

Deadline: 15:00, 27th November 2020

Report Template

1. Introduction
   Briefly introduce the aims of this coursework.
2. Data Preparation
   2.1. Data Cleaning
   Describe the way the data is cleaned by the use of OpenRefine, including errors identified and corrected.
   2.2. Data conversion
   Describe how the cleaned data is converted to data sets, which can be analysed by algorithms in Weka.
3. Data Analytics
   3.1. Classification
   3.2. Regression
   3.3. Association
   3.4. Clustering

   Note:
   - Three of the above 4 sections will be fine. For each section, apply at least one of the algorithms in that category. Settings, datasets used, attributes involved for each run should be described. Patterns/rules need to be retrieved, analysed and interpreted. About 6 patterns/rules for each algorithm used are expected.
4. Summary/Conclusion
   References are expected