**Cover sheet for submission of**

**work for assessment**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **UNIT DETAILS** | | | | | | | | | | |
| Unit name | | Data Science Principles | | | | | | Class day/time | THU12:30 | Office use only |
| Unit code | | COS10022 | | | Assignment no. | | 2 | Due date | 14/5/2023 |  |
| Name of lecturer/teacher | | | | Dr. Pei-Wei Tsai | | | | | |  |
| Tutor/marker’s name | | | Teja Gowda | | | | | | | Faculty or school date stamp |
| **STUDENT(S)** | | | | | | | | | | |
| Family Name(s) | | | | | | Given Name(s) | | | | Student ID Number(s) |
| (1) | Nguyen | | | | | Nam Tung | | | | 103181157 |
| (2) |  | | | | |  | | | |  |
| (3) |  | | | | |  | | | |  |
| (4) |  | | | | |  | | | |  |
| (5) |  | | | | |  | | | |  |
| (6) |  | | | | |  | | | |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **DECLARATION AND STATEMENT OF AUTHORSHIP** | | | | | | |
| 1. I/we have not impersonated, or allowed myself/ourselves to be impersonated by any person for the purposes of this assessment. 2. This assessment is my/our original work and no part of it has been copied from any other source except where due acknowledgement is made. 3. No part of this assessment has been written for me/us by any other person except where such collaboration has been authorised by the lecturer/teacher concerned. 4. I/we have not previously submitted this work for this or any other course/unit. 5. I/we give permission for my/our assessment response to be reproduced, communicated, compared and archived for plagiarism detection, benchmarking or educational purposes.   I/we understand that:   1. Plagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a form of cheating and is a very serious academic offence that may lead to exclusion from the University. Plagiarised material can be drawn from, and presented in, written, graphic and visual form, including electronic data and oral presentations. Plagiarism occurs when the origin of the material used is not appropriately cited. | | | | |
| **Student signature/s**  I/we declare that I/we have read and understood the declaration and statement of authorship. | | | | |
| (1) | Tung | (4) |  | |
| (2) |  | (5) |  | |
| (3) |  | (6) |  | |

Further information relating to the penalties for plagiarism, which range from a formal caution to expulsion from the University is contained on the Current Students website at **www.swin.edu.au/student/**

Copies of this form can be downloaded from the Student Forms web page at **www.swinburne.edu.au/studentforms/** PAGE 1 OF 1

# http://www.swinburne.edu.my/images/2013/logo.jpgSwinburne University of Technology Hawthorn Campus Dept. of Computer Science and Software Engineering

**COS10022 Data Science Principles**

Assignment 2 - *Semester 1, 2023*

**Assessment Title**: Data Cleaning and Analytics

## Assessment Weighting: 30%

**Due Date**: Saturday, 14th May 2023 at 11.59 pm (AEDT)

**Assessable Item:**

* One (1) piece of a written report no more than 10-page long with the signed Assignment Cover Sheet.
* A unit peer must review your submission before it can be marked.

The submitted report should answer all questions listed in the assignment task section in sequence.

You must include a digitally signed Assignment Cover Sheet with your submission.

1. Follow the instructions to clean the data and answer questions. If any of the nodes you used in the workflow has a random seed, set **3122** to the seed to fix the random state. **[70 marks in total]**
   1. Our goal is to predict the credit score from the given data. There is/are one (or multiple) attribute(s) which is/are significantly irrelevant to the goal. Exclude the attribute(s) and give a persuasive rationale for that. The excluded attribute(s) is(are) , and the reason(s) for removing it(them) is(are) . **[5 marks]**

Ans:

The excluded attribute is(are)  **name**  because  **The user ‘s name does not create any impact on the credit score. Credit score should be impacted by some factor such as payment, tax, … not name. Removing unimportant attributes such as name can also reduce the risk of overfitting of the model. Last but not least, name is a personal identifier of the user and it should not be included in the dataset to prevent the privacy and security-related problems.** .

* 1. After removing the selected attribute(s), let’s start to remove tuples containing missing values. Remove tuples only if any of the attributes listed below have missing values: “Month,” “Age,” “Occupation,” “Annual\_Income,” “Num\_Bank\_Accounts,” “Num\_Credit\_Card,” “Interest\_Rate,” “Num\_of\_Loan,” “Delay\_from\_due\_date,” “Changed\_Credit\_Limit,” “Credit\_Mix,” “Outstanding\_debt,” “Credit\_Utilization\_Ratio,” “Credit\_History\_Age,” “Payment\_of\_Min\_Amount,” “Total\_EMI\_per\_month,” “Amount\_invested\_monthly,” and “Payment\_Behaviour.” Moreover, some tuples with infeasible values in the attributes, such as “Monthly\_Inhand\_Salary” < 0, “Num\_Bank\_Accounts” < 0, “Num\_Credit\_Card” < 0, and “Changed\_Credit\_Limit” contains “\_”, should also be removed. List the node(s) (in sequence) and the corresponding command(s) used in this process. **[5 marks]**

Ans:

The used node is  **Missing Value and Rule-based row filter .** The usedcommends are listed as follows:

# Put Your Commands Here #

$Monthly\_Inhand\_Salary$ < 0 => TRUE

$Num\_Bank\_Accounts$ < 0 => TRUE

$Num\_Credit\_Card$ < 0 => TRUE

$Changed\_Credit\_Limit$ LIKE "\_" => TRUE

Then exclude the TRUE value

* 1. Check for the “Age” attribute to eliminate symbols that are not numbers to recover the data into the usual number format. Moreover, drop the tuples whose “Age” value is lower than or equal to 0 or greater than 120. List the node(s) (in sequence) and the corresponding command(s) used in this process. **[5 marks]**

Ans:

|  |  |  |
| --- | --- | --- |
| **Sequence** | **Node** | **Command** |
| 1 | String Manipulation | toInt(regexReplace($Age$, "[^0-9-]", "")) |
| 2 | Rule-based row filter | $Age$ <= 0 OR $Age$ > 120 => TRUE |
| 3 |  |  |
|  |  | # Extend the table if you need more spaces. # |

* 1. Remove the non-numerical symbol in the “Annual\_Income” column and convert it to the double format. List the node(s) (in sequence) and the corresponding command(s) used in this process. **[5 marks]**

Ans:

|  |  |  |
| --- | --- | --- |
| **Sequence** | **Node** | **Command** |
| 1 | String Manipulation | toDouble(regexReplace($Annual\_Income$, "[^0-9.]", "")) |
| 2 |  |  |
| 3 |  |  |
|  |  | # Extend the table if you need more spaces. # |

* 1. Convert the “\_\_\_\_\_\_\_” in the “Occupation” attribute to Null. Please note that Null is different from an empty string. Remove the non-numerical symbol in “Num\_of\_Loan” and convert it to integer data type. Take absolute values of attributes “Num\_Bank\_Accounts” and “Num\_Credit\_Card.” Set values to 0 for the “Num\_of\_Loan” attribute if the original values are negative. Remove the non-numerical symbol in “Num\_of\_Delayed\_payment” and convert it into integer format. Set the “Credit\_Mix” value to “Unknow” if the original value is “\_”. Remove the non-numerical symbol in “Outstanding\_Debt” and convert it into the double format. List the node(s) (in sequence) and the corresponding command(s) used in this process. **[10 marks]**

Ans:

|  |  |  |
| --- | --- | --- |
| **Sequence** | **Node** | **Command** |
| 1 | String Manipulation | toNull(replace($Occupation$,"\_\_\_\_\_\_\_","")) |
| 2 | String Manipulation | toInt(regexReplace($Num\_of\_Loan$, "[^0-9-]", "")) |
| 3 | Math Formula | abs($Num\_Bank\_Accounts$) |
| 4 | Math Formula | abs($Num\_Credit\_Card$) |
| 5 | Rule Engine | $Num\_of\_Loan$ < 0 => “0”  TRUE => $Num\_of\_Loan$ |
| 6 | String Manipulation | toInt(regexReplace($Num\_of\_Delayed\_Payment$,"[^0-9-]", "")) |
| 7 | Rule Engine | $Credit\_Mix$ = "\_" => "Unknow"  TRUE => $Credit\_Mix$ |
| 8 | String Manipulation | toDouble(regexReplace($Outstanding\_Debt$,"[^0-9.-]" ,"" )) |

* 1. Convert the “Credit\_History\_Age” to the count of months and store it in the integer format. For example, if the original value from a tuple is “22 Years and 1 Months”, the value will be 265 after the conversion (22 \* 12 + 1 = 265). Store the converted result in a new attribute called “Total\_CHA.” List the node(s) (in sequence) and the corresponding command(s) used in this process. **[10 marks]**

Ans:

|  |  |  |
| --- | --- | --- |
| **Sequence** | **Node** | **Command** |
| 1 | String Manipulation | toInt(regexReplace($Credit\_History\_Age$, "^(\\d+).\*$", "$1")) |
| 2 | String Manipulation | toInt(regexReplace($Credit\_History\_Age$, "^.\* (\\d+).\*$", "$1")) |
| 3 | Math Formula | $Year$ \* 12 + $MonthCount$ |
| 4 | Column filter | Filter out the 2 columns used to calculate the Total\_CHA: Year and MonthCount |

* 1. Remove the non-numerical symbol in “Amount\_invested\_monthly” and convert it to the double format. Set the value to “Unknow” if the original value in “Payment\_Behaviour” attribute starts with “!@”. Remove the non-numerical symbol in “Monthly\_Balance” and convert it to the double format. Convert “Changed\_Credit\_Limit” into the double format. List the node(s) (in sequence) and the corresponding command(s) used in this process. **[5 marks]**

Ans:

|  |  |  |
| --- | --- | --- |
| **Sequence** | **Node** | **Command** |
| 1 | String Manipulation | toDouble(regexReplace($Amount\_invested\_monthly$, "[^0-9.-]", "")) |
| 2 | Rule Engine | $Payment\_Behaviour$ LIKE "!@\*" => "Unknow"  TRUE => $Payment\_Behaviour$ |
| 3 | String Manipulation | toDouble(regexReplace($Monthly\_Balance$, "[^0-9.-]", "")) |
| 4 | String Manipulation | toDouble($Changed\_Credit\_Limit$) |

* 1. Use the “Missing Value” node and use the “Next Value\*” to replace missing values in all string type attributes. Use the “Previous Value\*” in the same node to replace missing values in any numerical format. If the value of “Monthly\_Balance” is negative, replace the value with 0. Screenshot the pop-up window with the correct settings. **[5 marks]**

Ans: (next page)

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, màn hình

Description automatically generated

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, màn hình

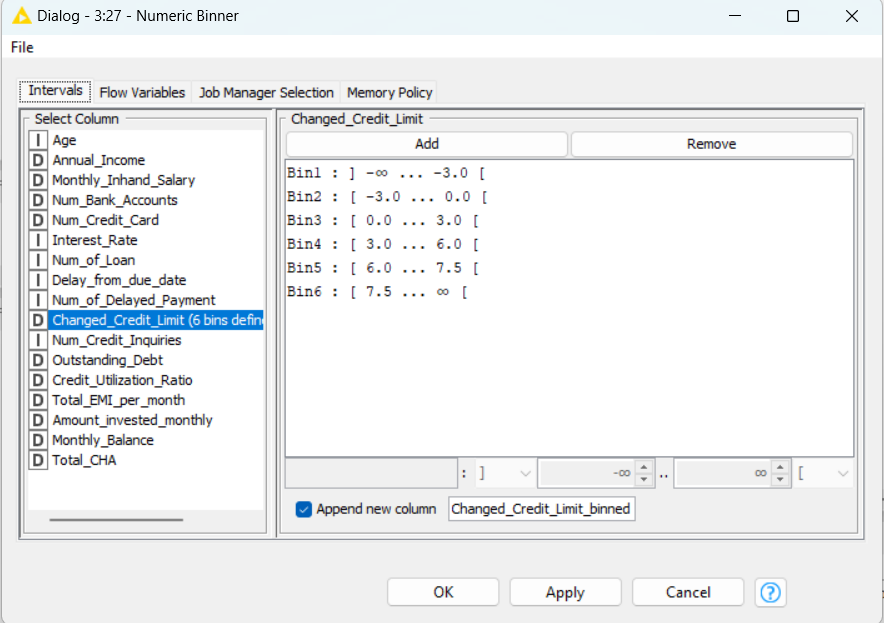
Description automatically generated

* 1. Simplify the “Type\_of\_Loan” attribute. If the original content has more than one type separated by a comma, keep only the first part. Otherwise, keep the full description if there is no comma included. For example, “Auto Loan, Credit-Builder Loan, Personal Loan, and Home Equity Loan” will become “Auto Loan”, “Credit-Builder Loan” will still be “Credit-Builder Loan”, and “Not Specified, Auto Loan, and Student Loan” will become “Not Specified” after the process. List the node(s) (in sequence) and the corresponding command(s) used in this process. **[10 marks]**

Ans:

|  |  |  |
| --- | --- | --- |
| **Sequence** | **Node** | **Command** |
| 1 | String Manipulation | regexReplace($Type\_of\_Loan$, ",.\*$", "") |
| 2 |  |  |
|  |  | # Extend the table if you need more spaces. # |

* 1. Bin the “Changed\_Credit\_Limit” attribute with six bins of ranges: , , , , , and and put the result into a new attribute called “Changed\_Credit\_Limit\_binned”. Screenshot the pop-up window with the correct settings of your binner. **[5 marks]**

Ans:

# Put the screenshot here #

* 1. Remove all temporarily created or useless attributes. Use the “Feature Selection Loop Start (1:1)” node to select the feature. The class label should be excluded from the features in the feature selection node. The Genetic Algorithm is specified to be the feature selection strategy with default population size and the maximum number of generations. Again, **3122** should be used as the static random seed. After selecting features, shuffle the data with seed **3122**. The data should be partitioned by “Linear sampling”, with 75% data in the training set and 25% in the test set. How many tuples and attributes (excluding the class label) are in the training set at the end? **[5 marks]**

Ans:

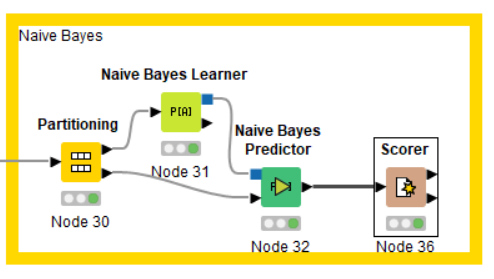
Static Column: Credit\_Score

There are  **68196 tuples and 13 attributes** in training set at the end.

1. Build a Naïve Bayes classifier using thetraining and test sets created in the previous task. Answer the following questions after completing the model training and test. **[15 marks in total]**
2. Give a screenshot of the Naïve Bayes classifier in the KNIME workflow. You can take the screenshot starting from the portioning node output to the end of the Naïve Bayes classifier part scorer. **[2.5 marks]**

Ans:

# Put the screenshot here #



1. The default probability should be 0.0001, the minimum standard deviation is 0.0001, the threshold standard deviation is 0, and the maximum number of unique nominal values per attribute should be set to 600 in the classifier. Screenshot the setting dialogue of your Naïve Bayes Learner. **[2.5 marks]**

Ans:

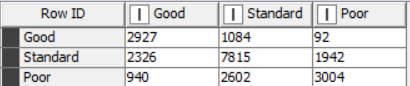
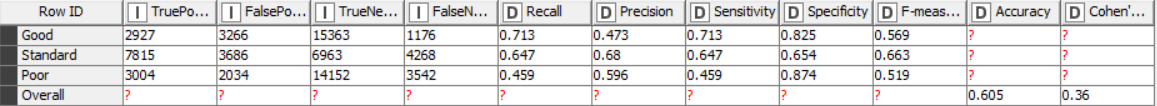
Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm, màn hình

Description automatically generated

1. Screenshot the confusion matrix and the Accuracy statistics of the test result. If the bank wants to minimise the risk of lending money to customers, the “Good” in “Credit\_Score” should be the major target. Based on the current result, does the classifier perform satisfactorily? **[5 marks]**

Ans:

# Put the screenshots here # (next page)



Rationales: In my point of view, the classifier does not perform satisfactorily. The reason will be given in the next answer.

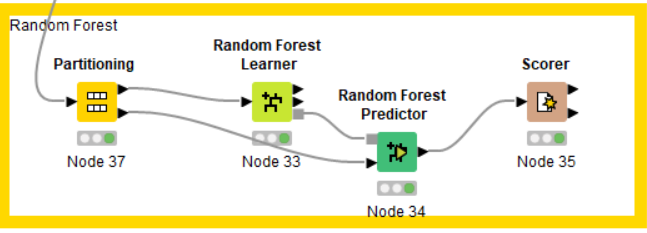
1. Which measurement should we look at to interpret your conclusion in this case? **[5 marks]**

Ans: We should look at the precision and recall rate. The recall rate is high (0.713) but the precision rate is relatively row. This indicates that only 47.3% of the “Good” predicted credit score is corrected. It does mean that more than half of the customers are being misclassified when they do not have a good credit score, leading to financial risk. In my point of view, for a model to be effective, both recall and precision rate should be high.

1. Build a random forest classifier using thetraining and test sets created in the previous task. Answer the following questions after completing the model training and test. Use the information gain ratio as the split criterion and **3122** as the static random seed to build the random forest model. **[15 marks in total]**
2. Give a screenshot of the random forest classifier in the KNIME workflow. You can take the screenshot starting from the portioning node output to the end of the Naïve Bayes classifier part scorer. **[2.5 marks]**

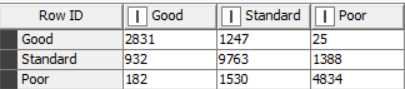
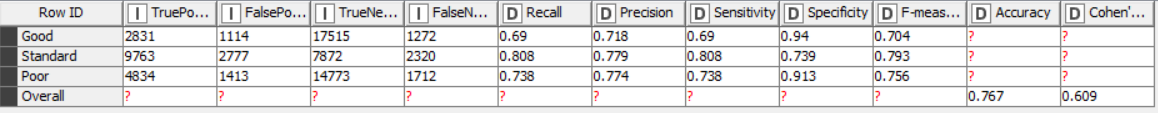
Ans:

# Put the screenshot here #



1. Screenshot the confusion matrix and the Accuracy statistics of the test result. **[2.5 marks]**

Ans:



1. If the bank wants to minimise the risk of lending money to customers, the “Good” in “Credit\_Score” should be the major target. Compare the measurements between random forest results and Naïve Bayes results. Which model presents a more suitable result? Which measure should be used to make the comparison? **[5 marks]**

Ans:

Here is a table demonstrating the comparison of the Good class in Naïve Bayes and Random Forest

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Overall accuracy | Precision | Recall | F-measure |
| Naïve Bayes | 0.605 | 0.473 | 0.713 | 0.569 |
| Random Forest | 0.767 | 0.718 | 0.69 | 0.704 |

The random forest presents a more suitable result for the prediction, according to the figures. The measurement that should be used to compare the model is the accuracy, recall, precision, and F-measure.

1. Which class does the built random forest model perform the best? What measurement(s) should we look at to find the answer? **[5 marks]**

Ans: The class that perform the best is the Standard class with very high recall, precision, and F-measure (0.808, 0.779, 0,793 respectively). The measurement that we should look at is recall, precision and F-measure.

---------------------------------------------------- End of Submission ----------------------------------------------------