



Swinburne University of Technology Hawthorn Campus
Dept. of Computer Science and Software Engineering

COS10022 Data Science Principles
Assignment 2 - Semester 1, 2023

Assessment Title: Predictive Model Creation and Evaluation

Assessment Weighting: 30%

Due Date: Saturday, 26th March 2023 at 11.59 pm (GMT+7)

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.

Purpose of Assignment

This assignment aims at evaluating students' achievement of the following unit learning outcomes:

1. **Explain the key concepts, techniques, and tools for cleaning the data and creating prediction models.**
2. **Work on feature and model selection with a bit of discovery of the prebuilt tool and implementation in a data science project.**

This is an individual assignment that requires peer review and communication with colleagues. Refer to the Unit Outline for the late submission penalty policy. You can ignore the high similarity on the cover page and the template wording, but not in your report content. You must ensure your submitted report has a similarity lower than 12% in total and less than 6% from a single source. Otherwise, your report will not be marked.

Key Lessons:

You are asked to clean the dataset based on the given instruction, divide the dataset, and then build a Naïve Bayes and random forest models in the KNIME analytic platform.

Introduction

The dataset contains 100,000 tuples of 3 different financial credit score classes. There are 24 attributes included in the source data. We have two goals in this assignment: the first is cleaning and preparing the data for later use, and the second is building two predictive models to predict the "Credit_Score" class. You are expected to follow the instructions for making your predictive model and answer questions.

Assignment Goal

This assignment aims to build experiences for students to clean the dataset, split the data into training and test sets, train usable predictive models, and explain the outputs. A small part of the discovery and research component is included in the assignment to expand the students' skill set.

Assignment Task

The dataset contains messy values as the dataset is collected from the real world. Your tasks are to clean the data and create the predictive models according to the instructions for answering the questions listed below. The source file is "**data_2023.csv**". The report should be prepared with the template and answer the questions. A table of content is not required.

Data Cleaning (70%)

You must follow the instructions to clean and split the given data set into training and test sets. Remember, a well-split dataset is the foundation of support for the model training and test. It is estimated that you will need to use around 30 nodes for data cleaning and partitioning before sending the partitioned data into the predictive models. Suggested nodes to be used include "File Reader," "Column Filter," "Rule-based Row Filter," "String Manipulation," "Math Formula," "Math Formula (Multi Column)," "Rule Engine," "Missing Value," "Shuffle," "Numerical Binner," "Feature Selection Loop Start (1:1)," and "Partitioning." You may see a warning sign on the "Missing Value" node stating, "The current settings use missing value handling methods that cannot be represented in PMML 4.2." It is normal; you can ignore it because we are not using PMML in the assignment.

Naïve Bayes Model (15%)

After partitioning the cleaned data into training and test sets, build a Naïve Bayes classifier to predict the "Credit_Score."

Random Forest Model (15%)

After partitioning the cleaned data into training and test sets, build a random forest classifier to predict the "Credit_Score."


Important Note

You must use the seed value specified in the instructions. Otherwise, you will get different results than the correct answer in almost all questions.

There are 100 marks on this assignment. Your proposal must address the following tasks.

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]**
 - 2) 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]**
- 3) 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]**
 - 4) 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]**
 - 5) 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]**
 - 6) 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]**
 - 7) 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]**
 - 8) 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]**
 - 9) 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]**
 - 10) Bin the “Changed_Credit_Limit” attribute with six bins of ranges: $[-\infty, -3.0)$, $[-3.0, 0)$, $[0, 3.0)$, $[3.0, 6.0)$, $[6.0, 7.5)$, and $[7.5, \infty)$ 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]**
 - 11) 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]**
2. Build a Naïve Bayes classifier using the training and test sets created in the previous task. Answer the following questions after completing the model training and test. **[15 marks in total]**
- 1) 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]**

- 2) 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]**
-  3) 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]**
- 4) Which measurement should we look at to interpret your conclusion in this case? **[5 marks]**
3. Build a random forest classifier using the training 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]**
 - 1) 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]**
 - 2) Screenshot the confusion matrix and the Accuracy statistics of the test result. **[2.5 marks]**
 - 3) 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]**
 - 4) Which class does the built random forest model perform the best? What measurement(s) should we look at to find the answer? **[5 marks]**

Submission Requirement

To fulfil the requirement of this assignment, the submission should be prepared in MS Word or PDF format, named **COS10022_[Student_ID]_Assignment_2** and submitted. Replace the **[Student_ID]** with your student ID number.

Failure to adhere to the submission requirements will immediately result in losing marks for this assignment.

----- End of Assignment -----