Project: Data Mining

Data Mining for Improved Retail and Healthcare
Due date: October 25, 2024
Weighting: 40%

Introduction

For this assessment, you will utilise Python libraries and code to analyse three case studies and their corresponding datasets. The analysis aims to uncover valuable insights that can benefit the respective sectors. You are required to rely primarily on the code and libraries that were covered in the practical sessions. If a library or code outside the course scope is used, it is important to explain its use.

This assessment presents an opportunity for you to demonstrate your proficiency in Descriptive and Predictive Data Mining by performing a range of tasks. These tasks include constructing various data mining models, such as association mining, clustering, decision trees, logistics regression, and neural networks, tailored to each case study. Through this, you can showcase your technical aptitude acquired from the practical sessions and your knowledge attained from the lectures and readings.

Instructions

1. This unit treats Academic integrity and plagiarism with the utmost seriousness. You are **NOT allowed to collaborate with other groups or seek external assistance for assessment tasks such as those generated by an AI, obtained from another person, or sourced from the internet**. In all cases, the submission is not an original piece of work created by the student. Any violation of this policy may result in severe consequences. Read the Assessment Policies on Canvas or the QUT Website for further details.
2. The dataset required for this assignment is available on Canvas with the file named **Project Datasets.zip**.
3. The assignment will be **evaluated based on two components**. The first component, worth 30%, will be assessed based on the report submitted by the group on Canvas. The second component, worth 10%, will be assessed based on an individual quiz. Note that **NO extension will be allowed on the quiz part**.
4. The deadline for the **assignment report submission is 25 October 2024**. Each group is required to submit only one final report.
   The report must include responses to all questions specified in the tasks and should be properly formatted to ensure easy navigation through the answers. You do not need to include an introduction, summary, conclusion, or references in the report. Some answers may require screenshots, which should be included in a table. While it may be tempting to provide excessive detail, it is crucial to focus on essential points and attach relevant screenshots to demonstrate thorough consideration of the matter. The report is anticipated to be around 25-30 pages long.
   The Jupyter notebook outlining all the code must also be submitted. However, the primary document for marking will be the group report, and the code will only be used in case of

uncertainty. Therefore, the report must be self-explanatory and not rely on the Jupyter Notebook attachment.

Name the final report as **assessment2-report.doc or assessment2-report.pdf**. The Word file should include a cover page with the Student ID number and full name (as in QUT-Virtual) for all students, along with the group number. **Combine the report with your team agreement and Jupyter notebook**. Name the compressed file **‘assessment2.zip’**. Submit this file on **Canvas.**

The submission process involves accessing the "Assignments" section on Canvas and submitting the report under the Assignment 2 section by selecting the "IFN509 Assignment 2 Submission" link.

5. The **Quiz** will be approximately 10-15 minutes long and **will be conducted online on October 28 at 11 am**. The quiz will consist of multiple-choice questions testing each student's comprehension of the project tasks and assess their ability to perform hands-on data modelling. Note **this is an individual assessment, and you should not seek support from your group members**. This part of the assessment cannot be extended.

6. You are required to **work in a group** for the report submission. You can either continue working with the same group as for Assessment 1 or form a new one. Regardless of the choice made, it is mandatory to register the team on the Canvas site for this assessment by Week 8. You are to choose an available group designated for Assignment 2.

   You are to manage the group. As in real life, the performance of the individuals in the team is judged by the performance of the team together, so choose your partners carefully. Once you form the group, group movements are NOT allowed.

   To ensure that everyone agrees to their responsibilities in the team, we have asked the team to complete a Team Agreement form. The team agreement template is in the Assessment module, under the Assignment 2 section. Once the team is formed, complete the team agreement form and register the team on Canvas. It should clearly state if there is an unequal distribution of marks between the team members. This should be agreed upon by all members, as shown by their signatures.

Project (a): Association mining to find common purchase patterns in a retail store sales data

Consider a retail store that records consumer goods purchases by ‘scanning’ the bar codes of individual products. The ‘Scanner’ dataset consists of 131706 instances, each presenting the sales of a product, quantity, price, and date. There are a total of 22625 unique customers, 5242 unique products, and 187 unique product categories. The table below details the attributes in the dataset ‘**D1.csv**’.

Attribute	Description	Data type
Date	Date of purchase	Datetime
Customer_ID	Unique identifier of the customer	ID
Sales_ID	Unique identifier of the purchase	ID
SKU	Stock keeping unit (SKU) is a unique identifier of a product	String
SKU_Categories	Unique identifier of SKU categories	String
Quantity	Quantity of products sold	Numeric
Sales_amount	Product price times quantity	Numeric

Tasks

Store owners would like to know what categories of products are bought together instead of individually. Using ‘**D1.csv**’, prepare a transactional dataset where each transaction represents the category of products along with other details. Build an association mining model on this dataset to identify the common product categories that customers purchase.

The task is to build an association mining model and, based on the data and analysis, answer the subsequent questions. (Add screenshots as appropriate)

1. What pre-processing was required on the dataset before building the association mining model? What variables did you include in the analysis? Justify your choice.
2. Conduct association mining and answer the following:
   a. What ‘min_support’ and `min_confidence’ thresholds were set for this mining exercise? Rationalise why these values were chosen.
   b. Report and interpret the top-5 rules (as per Lift values).
3. List the five most common product categories that customers bought with the product category ‘01F’.
4. Can you perform sequence analysis on this dataset? If yes, present your results. If not, rationalise why.
5. How can the outcome of this study be used by the relevant decision-makers?

Project (b): Clustering Diabetes data

The Diabetes dataset consists of hospital admission information for diabetic patients represented by 30 variables, including patient information, treatments, and medications. This is a subset of data provided to you in Assessment 1. This dataset (**D2.csv**) consists of 20,000 unique patients and has been pre-processed to deal with some common errors, e.g., missing values etc. Your task is to perform clustering on D2.csv and explain the least number of useful clusters discovered.

The description of the dataset is as follows:

No.	Variable Name	Description	DataType
1	race	Race of the patient	String
2	gender	Gender of the patient	String
3	age	Grouped ages	String
4	admission_type_id	Integer identifier corresponding to 9 distinct admission types (see file IDs_mapping for details)	Numeric
5	discharge_disposition_id	Integer identifier corresponding to 29 distinct values (see file IDs_mapping for details)	Numeric
6	admission_source_id	Integer identifier corresponding to 26 distinct values (see file IDs_mapping for details)	Numeric
7	time_in_hospital	Integer number of days between admission and discharge	Numeric
8	medical_specialty	Categories of medical specialties	String
9	num_lab_procedures	Number of lab tests performed during the encounter	Numeric
10	num_procedures	Number of procedures (other than lab tests) performed during the encounter	Numeric
11	num_medications	Number of distinct generic names administered during the encounter	Numeric
12	number_outpatient	Number of outpatient visits of the patient in the year preceding the encounter	Numeric
13	number_emergency	Number of emergency visits of the patient in the year preceding the encounter	Numeric
14	number_inpatient	Number of inpatient visits of the patient in the year preceding the encounter	Numeric
15	number_diagnoses	Number of diagnoses entered into the system	Numeric
16	max_glu_serum	Indicates the range of the result or if the test was not taken. Values include “>200”, “>300”, “normal” and “none” if not measured	String
17	A1Cresult	Values include: ‘>8’ if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.	String
18	metformin	These are 10 variables for medications. The values of these variables indicate whether the drug was prescribed or there was a change in the dosage.	String
19	repaglinide		String
20	nateglinide		String
21	chlorpropamide		String

22	glimepiride	Values include: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed.	String
23	acetohexamide		String
24	glipizide		String
25	glyburide		String
26	tolbutamide		String
27	insulin		String
28	change	Indicates if there was a change in diabetic medications (either dosage or generic name). Values include: “True” if there was a change and “False” if there was “no change”	Boolean
29	diabetesMed	Indicates if there was any diabetic medication prescribed. Values include: “TRUE” and “FALSE”	Boolean
30	readmitted	Inpatient readmission. Values include 0 for no record of readmission and 1 if the patient was readmitted.	Boolean

Tasks

Suppose you are interested in profiling diabetic patients by the number of lab procedures, the number of outpatient visits, the number of inpatient visits, the number of medications, and the time spent in the hospital. The task is to perform clustering and answer the subsequent questions based on the data and analysis. (Add screenshots as appropriate)

1. What pre-processing was required on the dataset (D2.csv) before building the clustering model and why?
2. Build a clustering model to profile the characteristics of diabetic patients defined by the chosen attributes. Answer the following:
   a. What clustering algorithm have you used and why?
   b. List the attributes used in this analysis.
   c. What is the optimal number of clusters identified? How did you reach this optimal number?
   d. Report the cluster centroids (each vector and its interpretation).
   e. Did you normalise/standardise the variables? What was its effect on the model – Does the variable normalisation/standardisation process enable a better clustering solution?
3. For the model with the optimal number of clusters, answer the following.
   a. Visualize the clusters using ‘pairplot’ and interpret the visualization.
   b. Characterize the nature of each cluster by giving it a descriptive label and a brief description. Hint: visualize cluster distribution.
4. Build another clustering model by including the variable ‘age’.
   Use the best setting (e.g., variable standardisation, optimal K, etc) obtained in the previous steps. Answer the following:
   a. What clustering algorithm have you used and why?
   b. What difference do you see in this clustering interpretation compared to the previous one (task 3)?
5. How can the relevant decision-makers use the outcome of this study?

Project (c): Building and Evaluating Predictive models

Utilize the Diabetes dataset (D2.csv), as in Project(b), for predictive modelling with various techniques. Based on past observations, the objective is to perform predictive modelling with different methods namely Decision tree, Logistic regression, and Neural network. The objective is to classify if a patient will be readmitted (reported as 1) or not readmitted (reported as 0) based on the variables representing patient information, treatments, and medications as in the table below.

No.	Variable Name	Description	DataType
1	race	Race of the patient	String
2	gender	Gender of the patient	String
3	age	Grouped ages	String
4	admission_type_id	Integer identifier corresponding to 9 distinct admission types (see file IDs_mapping for details)	Numeric
5	discharge_disposition_id	Integer identifier corresponding to 29 distinct values (see file IDs_mapping for details)	Numeric
6	admission_source_id	Integer identifier corresponding to 26 distinct values (see file IDs_mapping for details)	Numeric
7	time_in_hospital	Integer number of days between admission and discharge	Numeric
8	medical_specialty	Categories of medical specialties	String
9	num_lab_procedures	Number of lab tests performed during the encounter	Numeric
10	num_procedures	Number of procedures (other than lab tests) performed during the encounter	Numeric
11	num_medications	Number of distinct generic names administered during the encounter	Numeric
12	number_outpatient	Number of outpatient visits of the patient in the year preceding the encounter	Numeric
13	number_emergency	Number of emergency visits of the patient in the year preceding the encounter	Numeric
14	number_inpatient	Number of inpatient visits of the patient in the year preceding the encounter	Numeric
15	number_diagnoses	Number of diagnoses entered into the system	Numeric
16	max_glu_serum	Indicates the range of the result or if the test was not taken. Values include “>200”, “>300”, “normal” and “none” if not measured	String
17	A1Cresult	Values include: ‘>8’ if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.	String
18	metformin	These are 10 variables for medications. The values of these variables indicate whether the drug was prescribed or there was a change in the dosage.	String
19	repaglinide		String
20	nateglinide		String
21	chlorpropamide		String
22	glimepiride		String

23	acetohexamide	Values include: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed.	String
24	glipizide		String
25	glyburide		String
26	tolbutamide		String
27	insulin		String
28	change	Indicates if there was a change in diabetic medications (either dosage or generic name). Values include: “True” if there was a change and “False” if there was “no change”	Boolean
29	diabetesMed	Indicates if there was any diabetic medication prescribed. Values include: “TRUE” and “FALSE”	Boolean
30	readmitted	Inpatient readmission. Values include 0 for no record of readmission, and 1 if the patient was readmitted.	Boolean

Tasks

Build different predictive models on this data set, such as decision trees, regression models, and neural networks, and compare their performances. The task is to perform classification mining and address the following questions based on the data and analysis. (add screenshots as appropriate)

Predictive modelling using Decision Tree

1. What pre-processing was required on the dataset (D2.csv) before decision tree modelling? What distribution split between training and test datasets was used?

2. Build a decision tree using the default setting. Answer the followings:
   a. What is the classification accuracy of training and test datasets?
   b. What is the size of the tree (number of nodes and rules)?
   c. Which variable is used for the first split?
   d. What are the 5 important variables (in the order) in building the tree?
   e. What parameters were used to build the tree? Detail them.

3. Build another decision tree tuned with GridSearchCV. Answer the followings:
   a. What is the classification accuracy of training and test datasets?
   b. What is the size of the tree (i.e. number of nodes and rules)?
   c. Which variable is used for the first split?
   d. What are the 5 important variables (in the order) in building the tree?
   e. Report if you see any evidence of model overfitting.

4. What differences do you observe between these two decision tree models (with and without fine-tuning)? How do they compare performance-wise? Produce the ROC curve for both DTs. Explain why those changes may have happened.
5. Using the better model, can you identify which patients could potentially be "readmitted"? Also, provide the general characteristics of those individuals.

Predictive modelling using Regression

1. What pre-processing was required on the dataset before regression modelling? What distribution split between training and test datasets was used?
2. Build a regression model with all inputs using the default regression method. Build another regression model tuned with GridSearchCV. Now, choose a better model to answer the followings:
   a. Explain why you chose that model.
   b. Name the regression function used.
   c. Did you apply standardisation of variables? Why would you standardise the variables for regression mining?
   d. Report the variables included in the regression model.
   e. Report the top 5 important variables (in the order) in the model.
   f. What is the classification accuracy of training and test datasets?
   g. Report any sign of overfitting in this model.

3. Build another regression model on the reduced variables set. Perform dimensionality reduction with Recursive feature elimination. Tune the model with GridSearchCV to find the best parameter setting. Answer the followings:
   a. Was dimensionality reduction useful in identifying a good feature set for building an accurate model? How these two models differ?
   b. What is the classification accuracy of training and test datasets?
   c. Report any sign of overfitting.
   d. Report the model's top 3 important variables (in the order).

4. Produce the ROC curve for all different regression models. Using the best regression model, can you identify which patients could potentially be "readmitted"? Provide the general characteristics of those individuals.

Predictive modelling using Neural Networks

1. What pre-processing was required on the dataset before neural network modelling? What distribution split between training and test datasets have you used?
2. Build a Neural Network model using the default setting. Answer the following:
   a. Explain the parameters used in building this model, e.g., network architecture, iterations, activation function, etc.
   b. What is the classification accuracy on training and test datasets?
   c. Did the training process converge and result in the best model?
3. Refine this network by fine-tuning it with GridSearchCV. Report the trained model.
   a. Explain the parameters used in building this model, e.g., network architecture, iterations, activation function, etc.
   b. What is the classification accuracy on training and test datasets?
   c. Did the training process converge and result in the best model?
   d. Do you see any sign of over-fitting?
4. Let us see if feature selection helps in improving the model. Build another neural network model with a reduced feature set. Perform dimensionality reduction by

selecting variables with a decision tree (use the best decision tree model you built in the previous modelling task). Fine-tune the model with GridSearchCV to find the best parameter setting. Answer the followings:

e. Did feature selection favor the outcome? Any change in network architecture? What inputs are being used as the network input?
f. What is the classification accuracy of training and test datasets?
g. How many iterations are now needed to train this network?
h. Do you see any sign of over-fitting? Did the training process converge and result in the best model?

5. Produce the ROC curve for all different neural networks. Now, using the best neural network model, provide characteristics of patients that could potentially be "readmitted”. If it is difficult (or even infeasible) to comprehend, discuss why.

Final remarks: Decision making

1. Finally, based on all models and analysis, is there a model you will use in decision-making? Justify your choice. Draw a ROC chart and accuracy table to support your findings.
2. Based on this analysis, Can you summarise the positives and negatives of each predictive modelling method?

Marks Distribution (Total 40 marks)

The assessment will be marked in two parts.

The first part, (30% worth), the group work, will be assessed via the final report that you will submit on Canvas. Note that in data analytics, there is hardly ever a single solution. The solution depends upon various settings such as the variable's role and measurements and the selected method parameters. You may find that your project partner may have a different solution than yours. Your group should decide on a single project that you would like to be marked. Submit **a single project report per group** discussing the final project components.

The second part, (10% worth), the individual work, will be assessed via an online **Quiz** that **will be held online on 28th October at 11 am.** This will test your understanding of the tasks. Note **the quiz is an individual (non-extendible) assessment and you should not seek support from your group members or anyone else during the quiz**.

The marks for the group report are distributed as follows.

Assignment Report Components	**Marks (30)**		
Association Mining	4	Pre-processing	1
		Rule Mining	2.5
		Decision Making	0.5
Clustering	7	Pre-processing	0.5
		Clustering Model 1	2
		Clustering Model 2	2
		Interpretation	2.5
Predictive Mining: Decision Tree Models	6	Pre-processing + Model 1	2
		Fine Tuned Model	2
		Models Comparison	1
		Patients' Characteristics	1
Predictive Mining: Regression Models	6	Pre-processing + Model 1	3
		Model 2 with RFE	2
		Models Comparison & Patients' Characteristics	1
Predictive Mining: Neural Network Models	6	Pre-processing + Model 1	1.5
		Fine Tuned Model	1.5
		Model with FS	2
		Models Comparison & Patients' Characteristics	1
Predictive Mining Final Remarks: Comparison	1		
Poor Report Presentation	A penalty of 3 marks.		

Assessment 2 – Report Criteria Sheet

Descriptive Mining: Projects (a &b)

Criteria	Comments and scoring
Non-Submission of all components/ evidence of plagiarism	0
Has demonstrated a task with a working model and demonstrated the ability to run the Python program and add some components. Questions were poorly answered.	1-2
Has implemented models for both components with at least one being substantially correct. Shows some understanding of concepts with some success applying knowledge in basic questions	3-4
Has the fundamentally correct implementation of both parts i.e. selection of correct variables in both datasets, correct utilisation, understanding, and explanation of clusters, findings association rules. Many questions have been reasonably answered. Demonstrate a good understanding of the methods and terms used in clustering and association mining, during written and verbal analyses. Some minor errors are allowed. A written report is required to be of a reasonable standard. Response to questions shows basic knowledge of the topic.	5-6
Has implemented all the requirements above with very few errors. A strong focus on the application of tools and evaluation and interpretation of results is evident. Response to questions shows an in-depth knowledge of the topic.	7-8
All the criteria above are met; extensive model generation and analysis have been conducted to produce quality outcomes and have applied principles learned in lectures to enhance the results. Response to questions shows extensive knowledge of the topic.	9-11

Predictive Mining: Project (c)

Criteria	Comments and scoring
Non-Submission of all components/ evidence of plagiarism	0
Has demonstrated a task with a working model and demonstrated the ability to run the Python program and add some components. Questions were poorly answered.	1-4
Has demonstrated a task with a working model and the process flow with the substantial but incorrect implementation of at least one of the three components. Questions were poorly answered.	5-6
Has implemented models for all three data mining algorithms with at least one being substantially correct. Shows some understanding of concepts with some success applying knowledge in basic questions	7-9
Has implemented models for all three algorithms with pre-processing. Two of the three tasks are fundamentally correct, with a substantially correct process flow that may contain minor errors. Response to questions shows a fundamental understanding of terms and concepts.	10-12
Has the fundamentally correct implementation of all three algorithms i.e. correct role and type assigning of variables, the effective building of models and their comparison/ assessment. Includes a demonstration of the competent application of tools. Almost all questions have been reasonably answered. Demonstrate a strong understanding of the methods and terms including predictive mining, accuracy, ROC chart, during written analyses. Some minor errors are allowed. The written report is required to be of a reasonable standard.	13-15
Has implemented all the requirements above with very few errors. A strong focus on the creative application of tools and evaluation and interpretation of results is evident.	16-17
All the criteria above are met; extensive model generation and analysis have been conducted to produce quality outcomes and have applied principles learned in lectures to enhance the results.	18-19