Project: Data Mining

Data Mining for Improved Retail and Healthcare
Due date: October 30, 2022
Weighting: 40%

Introduction

In this assessment, you will use Python libraries and code to analyse three scenarios and associated datasets to identify useful insights to benefit the retail and healthcare industry. You should mostly use the code and libraries introduced in the subject. If you use a major library outside of the subject scope, explain why it was needed.

The tasks set in this assessment allow you to display your knowledge and understanding of Descriptive and Predictive Data Mining. You will build several data mining models including association mining, clustering, decision tree, logistics regression, and neural network, suiting each scenario. You will be able to display your technical competence gained from the practicals and the knowledge gained from the lectures and your readings.

Instructions

1. We take academic integrity and plagiarism very seriously in this unit. You are **NOT allowed to consult with any other groups and/or take help outside for assessment tasks**. There may be serious consequences. Read the Assessment Policies on Blackboard or the QUT Website.
2. This is a group assignment. You can continue with the same group as in Assessment 1 or can form a new group. If you change the group, contact the teaching team to remove you from the old team – send all the information via email.
3. The assignment report is due on October 30, 2022. A single report per group should be submitted via the Blackboard Assignment submission link.
4. The dataset required for this assignment can be found on Blackboard with the file named **Project datasets.zip**.
5. The assignment will be **marked as two components.** The first component, **(30% worth), the group work**, will be assessed via the report that your group will submit on Blackboard. The second component, **(10% worth), the individual work**, will be assessed via the quiz. **A short quiz (~15 min) will be held online on 31st October at 9 am.** You will be provided with a link to finish this quiz near the date. The quiz will contain multiple-choice questions related to the project tasks and will test your understanding of hands-on data modelling capability. There is no extension allowed on the quiz part of the assessment.
6. A single report per group should be submitted via online submission, answering each question of the three scenarios and modelling tasks. There is no need for including an introduction, summary, conclusion, or references in the report. The report should just include responses to the questions set in the scenario. Some answers may require screenshots. Use them as needed, but you may include a table detailing those results. While you may like to go into extreme detail, you will not have the space to do so. Rather, write down the important points and attach the important screen dumps, to show that you have

thought the matter through. The report is expected to be about 25-30 pages long, including screenshots.

Name the project report as **project.docx**. The word file should include a cover page with the Student ID number, full name (as in QUT-Virtual), and **the marks appraisal table** for all students, along with the group name. Combine this file with your Jupyter notebook. Name the compressed file as **project_*.zip, replacing * with your group number.** Submit this file on **Blackboard (under the Assessment Item 2 link)**.

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

A retail store records the purchase of consumer goods by 'scanning' the bar codes of individual products. The "Scanner" dataset consists of 131706 instances where each instance presents 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**'.

1. get data set organize,
according category,
only use var for association
2.

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

find association in 187 products

Instead of individual products, the store owners would like to know what categories of products are bought together. 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 what are the common product categories that customers purchase. The task is to conduct Association mining on this data set.

Answer the following concerning this data and analysis. (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? Rationalize why these values were chosen. choose larger values and see and choose them
    confidence: give the rules
    b. Report the top-5 rules (as per Lift values) and interpret them.
3. Identify the top-5 common product categories that customers bought with the product category '01F'.
   java => do sequence
4. Can you perform sequence analysis on this dataset? If yes, present your results. If not, rationalize why. need to track all category bought by customer => if can you can do analysis
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 diabetic patient hospital admission information 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.

Suppose you are working as a data analyst in eHealth. Your task is to (1) conduct clustering on this data set to understand the common characteristics of the patients and (2) describe the minimum number of effective clusters identified.

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	String
19	repaglinide		String

20	nateglinide	values of these variables indicate whether the drug was prescribed or there was a change in the dosage. 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
21	chlorpropamide		String
22	glimepiride		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

The Agency is interested in profiling the diabetic patients described 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. Answer the following concerning this data and clustering analysis (add screenshots as appropriate).

choose best variable and do clustering

1. What pre-processing was required on the dataset (D2.csv) before building the clustering model on the chosen attributes?
2. Build a clustering model to profile the characteristics of diabetic patients defined by the chosen attributes. Answer the followings:
   a. What clustering algorithm is used in the analysis?
   b. What is the optimal number of clusters identified? How did you reach this optimal number?
   c. Report the cluster centroids (each vector and its interpretation).
   d. Did you normalise the variables? What was its effect on the model – Does the variable normalization 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: use cluster distribution.
4. Now, build another clustering model by including the attribute ‘age’.
   Use the best setting (e.g., variable standardisation, optimal K, etc) obtained in the previous steps. Answer the followings:
   a. What clustering algorithm have you used?
   b. Report the cluster centroids (each vector and its interpretation).
   c. What difference do you see in this clustering interpretation when compared to the previous one (task 3)?
5. How can the outcome of this study be used by the relevant decision-makers?

Project (c): Building and Evaluating Predictive models

As used in Project (b), the Diabetes dataset consists of diabetic patient hospital admission information represented by 30 variables including patient information, treatments, and medications. This dataset (**D2.csv**) consists of 20,000 unique patients and has been pre-processed to deal with some common errors such as missing values, etc.

Use this processed data 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	String
19	repaglinide		String
20	nateglinide		String

21	chlorpropamide	the dosage. 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
22	glimepiride		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

<u>Tasks</u>

Your task is to build various predictive models including decision tree, regression model, and neural network on this data set, and compare them to find the best predictive model that can be used for prediction purposes in future. Answer the following (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 have you 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 have been used in building 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.

performance wise (accuracy): compare between default and gridsearch CV. Explain why overfitting in default and not in gridsearch

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. From the better model, can you identify the general characteristics of patients that could potentially be "readmitted"? If yes, describe those characteristics. If it is difficult (or even infeasible) to comprehend, discuss why.

Predictive modelling using Regression

1. What pre-processing was required on the dataset before regression modelling? What distribution split between training and test datasets have you used?
2. Build a regression model using the default regression method with all inputs. 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 to identify a good feature set for building an accurate model?
   b. What is the classification accuracy of training and test datasets?
   c. Report any sign of overfitting.
   d. Report the top 3 important variables (in the order) in the model.

4. What differences do you observe between these two regression models (with and without feature selection)? How do they compare performance-wise? Produce the ROC curve for all different regression models.
5. Using the best regression model, can you identify the general characteristics of patients that could potentially be "readmitted"? If yes, describe those characteristics. If it is difficult (or even infeasible) to comprehend, discuss why.

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. Refine this network by 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 of 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?

3. 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 that you have

built in the previous modelling task). Tune the model with GridSearchCV to find the best parameters setting. Answer the followings:

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

4. What differences do you observe between these two neural network models (with and without feature selection)? How do they compare performance-wise? Produce the ROC curve for all different NNs.
5. Using the best neural network model, can you identify the general characteristics of patients that could potentially be "readmitted"? If yes, describe those characteristics. If it is difficult (or even infeasible) to comprehend, discuss why. Look like dt is the best model for patient characteristic

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. Can you summarise the positives and negatives of each predictive modelling method based on this analysis?

Marks Distribution (Total 40 marks)

This assignment will be marked in two components.

1. **Group marks (30)** based on a written report submitted by the group (due on 30th October).
2. **Individual marks (10)** based on a **quiz (~15 min) held online on 31st October at 9 am** on the project-related questions to test your understanding of hands-on data modelling.

The marks for a group report are distributed as follows.

Assignment Report Components	**Marks (30)**		
Association Mining	5	Pre-processing	1
		Rule Mining	3.5
		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	4.5	Pre-processing + Model 1	1.5
		Fine Tuned Model	1
		Models Comparison & Patients' Characteristics	2
Predictive Mining: Regression Models	6.5	Pre-processing + Model 1	2.5
		Model 2 with RFE	2
		Models Comparison & Patients' Characteristics	2
Predictive Mining: Neural Network Models	5.5	Pre-processing + Model 1	2.5
		Model with FS	2
		Models Comparison & Patients' Characteristics	1
Predictive Mining Final Remarks: Comparison	1.5		
Poor Report Presentation	A penalty of 3 marks.		

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 project report (word file) should be correctly formatted and should be easy to navigate through answers provided for all questions defined for all tasks. There will be a penalty if the report is not well presented. The project report should contain all necessary information. We will only check the Jupyter Notebook if any doubt about the answers provided.

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.	9-12

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.	10-11
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.	12-14
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.	15-16
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.	17-18