Assignment Specification

Continuous Assessment 2
Class Group: TU059 / TU060

Worth: 50% of the overall mark for the module	**Due date:** 04/01/2021, 11:59pm

Penalty for late submission: 3^d% where d is the number of days late. No submission allowed after 4 days.

Objective: To demonstrate what you have learned so far in terms of data warehouse modelling, data analysis and machine learning through the use of Oracle SQL.

Description of Tasks

Section A: Data Warehouse Modelling (40%)

A telecommunications company is starting a data project to improve their customer analytics capabilities. The primary purpose of this project is to help customer service agents build a better picture of the customers they speak to on the phone. During informal interviews with the stakeholders, the following were suggested as possible features for the system:

- identify how valuable a customer is to the company relative to other customers
- build up a picture of their customers' profiles
- determine whether a customer's behaviour patterns have changed recently
- identify the call plans which bring in the most revenue

The stakeholders are open to additional insights not covered above. Using the data provided as a starting point, design and build a data warehouse to facilitate data analysis for the proposed system. Your submission should contain the following.

a) A brief description (including screenshots) of how you imported the data

b) A short report (which should be pitched at stakeholders who are familiar with data analysis concepts, but not SQL experts) in which you discuss the decision-making process behind the design of the fact-table(s) and dimensional models, including the types of queries you anticipate the fact table answering.

c) The SQL script used to transform the data from the imported tables into your final data-warehouse format.

Ensure you can run this script without error on the Oracle database. Data in the import tables should be kept in place as this script will be executed by the examiner correcting your assignment.

Section B: Data Analysis and Queries Using SQL (30%)

Using the tables you have created in part (a), carry out an analysis of the data using SQL.

Use the data warehouse to implement queries suitable for the data project outlined in section A

Add a section to your report detailing the data analysis you have carried out and a discussion of the queries you have implemented.

Section C: Machine Learning using SQL (30%)

For this part of the assignment you will create two machine learning models using in-database machine learning features. You have been asked to produce a churn model which will predict customers who are likely to churn this month.

We want to predict for any given customer and any given month, will the customer churn that month? Create a fact table for use with this model, think about the most suitable grain.

You may need to create additional tables, views, and use sampling to prepare the training and test data sets. You may need to create a view that contains the predicted values for the testing data set.

Complete this process by evaluating the accuracy of two separate models.

Write a PL/SQL program to combine the accuracy measures from the various models and to present them to the user. For example, you can use DBMS_OUTPUT function to display the results

Add a section to your report explaining the parameters values you have selected and why, and evaluating the accuracy of your chosen models.

Deliverable: an SQL script which

- creates and populates the fact table
- trains the models on training data
- evaluates their accuracy on test data and
- outputs the results to the user.
- a screenshot showing the output.

Submission

You should submit a zip file with your report and any other necessary files. The code should work without issue when run on the lecturer's machine.

Plagiarism

Code sourced from elsewhere must be clearly cited (bear in mind your Python and/or R skills are also being examined). Plagiarism will result in a zero mark (0%). You should make yourself familiar with the plagiarism policy of Technological University Dublin.

Grading Rubric (Question 1)

	Import process (10)	Dimension models (10)	Fact table (10)	Report (10)
Expectations Far Surpassed (100%)	Data imported without error. Import process clearly described and well documented with screenshots.	Dimension tables are correctly structured according to the star schema model. All relevant dimensions are captured in the star schema in exhaustive detail.	Fact table is correctly structured (star schema). Design decisions are clearly justified. Allows for varied analytical queries and is ideally suited to these queries.	Report meets all requirements. Report is exceptionally well formatted. Information is communicated very clearly and concisely , and pitched correctly at the target audience. Text is enhanced with figures where appropriate. Demonstrates a mastery of the module content.
Expectations Exceeded (75%)	Data import without error. Import process is generally well documented	Dimension tables are correctly structured according to the star schema model. All significant relevant dimensions are captured in the star schema in detail.	Fact table is correctly structured (star schema). Design decisions are justified. Allows for varied analytical queries.	Report meets all requirements. Report is well formatted. Information is communicated clearly. Text is enhanced with figures where appropriate. Demonstrates comfort with the module content.
Expectations Met (50%)	Data import largely error-free. Import process is documented to some extent	Includes most important dimensions in sufficient detail	Demonstrates some awareness of how to design a fact table. Shows evidence of design decisions taken. Demonstrates some consideration of the types of queries which may be run.	Report broadly meets the requirements, communicates the required information and demonstrates knowledge of the module content.
Expectations Not Met (25%)	Data import process contains substantial issues, or is undocumented	Many dimensions missing or substantial detail lacking from dimension tables	Fact table is not designed correctly (star schema), little evidence of design decisions or queries for which the fact table will be used.	Report does not meet the requirements and does not communicate the required information, or fails to demonstrate awareness of the module content
Not Done (0%)	Not convincingly attempted	Not convincingly attempted	Not convincingly attempted	Not convincingly attempted

Grading Rubric (Question 2)

	Data analysis (10)	Queries (15)	Report (5)
Expectations Far Surpassed (100%)	Data analysis is flawless, extensive and appropriate, leading to discovery of new and interesting insights in the data. The output of the data analysis is clearly displayed to the end user and well-formatted. Demonstrates a mastery of SQL for data analysis.	Queries are appropriate, varied and complex, yielding new and interesting insights into the data. Queries are correctly interpreted and would meet a perceived business need. Demonstrates extensive research extending beyond the module material. Demonstrates significant original thinking. No inaccuracies	Report meets all requirements. Report is exceptionally well formatted. Information is communicated very clearly and concisely , and pitched correctly at the target audience. Text is enhanced with figures where appropriate. Demonstrates a mastery of the module content. Report constitutes and outstanding piece of writing
Expectations Exceeded (75%)	Data analysis is extensive and appropriate, conveying useful information about the distribution etc. of the data. The output of the data analysis is clearly displayed to the end user and well-formatted. Demonstrates a strong ability to use SQL for data analysis	Queries are appropriate, varied and complex, yielding interesting insights into the data. Queries are correctly interpreted and would meet a perceived business need. Demonstrates some research extending beyond the module material. Demonstrates original thinking. Few if any inaccuracies	Report meets all requirements. Report is well formatted. Information is communicated clearly. Text is enhanced with figures where appropriate. Demonstrates comfort with the module content.
Expectations Met (50%)	Data analysis is sufficient and somewhat appropriate. The output of the data analysis is clearly displayed to the end user and well-formatted. Demonstrates a strong ability to use SQL for data analysis	Queries are appropriate, yielding some insight into the data. Demonstrates familiarity with module material.	Report broadly meets the requirements, communicates the required information and demonstrates knowledge of the module content..
Expectations Not Met (25%)	Data analysis is insufficient. It is unclear how to produce or where to find the output of the data analysis. Does not demonstrate an ability to use SQL for data analysis	Queries fail to yield some insight into the data. Does not demonstrate familiarity with module material.	Report does not meet the requirements and does not communicate the required information, or fails to demonstrate awareness of the module content
Not Done (0%)	Not convincingly attempted	Not convincingly attempted	Not convincingly attempted

Grading Rubric (Question 3)

	Fact table (5)	Models (15)	PL/SQL Code (5)	Report (5)
Expectations Far Surpassed (100%)	Fact table is correctly structured (star schema). Fact table is optimally designed for the problem at hand. Fact table is populated correctly.	All 2 models implemented. Choice of models and rationale demonstrate a deep understanding of the techniques used. Parameters are chosen effectively using empirical techniques. Model output is clear. Results are reproducible on multiple script runs	PL/SQL code is optimal in terms of efficiency and presentation. Code is well-commented and clear, variable names are well chosen, error handling is incorporated.	Report meets all requirements. Report is exceptionally well formatted. Information is communicated very clearly and concisely , and pitched correctly at the target audience. Text is enhanced with figures where appropriate. Demonstrates a mastery of the module content. Report constitutes and outstanding piece of writing
Expectations Exceeded (75%)	Fact table is correctly structured (star schema). Fact table is well designed for the problem at hand. Fact table is populated correctly.	All 2 models implemented. Choice of models and rationale show evidence of extensive research. Demonstrates thoughtfulness in selecting model parameters. Model output is clear. Results are reproducible on multiple script runs	PL/SQL code is efficient and well presented. Code is clear, variable names are well chosen, error handling is incorporated.	Report meets all requirements. Report is well formatted. Information is communicated clearly. Text is enhanced with figures where appropriate. Demonstrates comfort with the module content.
Expectations Met (50%)	Demonstrates some awareness of how to design a fact table. Design is of some relevance to the problem at hand.	Demonstrates capability to utilize ML in the database. Model parameters are explained to some extent. At least one model is generally correctly implemented,	PL/SQL code is broadly correct. The student demonstrates some ability to use PL/SQL in the database.	Report broadly meets the requirements, communicates the required information and demonstrates knowledge of the module content..
Expectations Not Met (25%)	Fact table is not designed correctly (star schema), little evidence of design decisions or queries for which the fact table will be used	Does not demonstrate capability to utilize ML in the database. Model parameters are not explained at all.	PL/SQL code is extremely unclear or has many substantial errors. Student does not demonstrate capability of using PL/SQL code in the database.	Report does not meet the requirements and does not communicate the required information, or fails to demonstrate awareness of the module content
Not Done (0%)	Not convincingly attempted	Not convincingly attempted	Not convincingly attempted	Not convincingly attempted