Evaluated Exercise – Part I: Databases and SQL

MPMD3.2 PÜ Advanced Data Mining Techniques, Databases and Big Data - WiSe2022/23

Points in Parts: Databases SQL: 20% // SciKit-Stack/Cloud: 20% // Big Data: 20% / Natural Language Processing: 20% / Presentation: 20%

Instructions

Report

- Please deliver all commands in your documentation (use a word-document and convert it later into a pdf or an html export of your jupyter-Notebook etc.).
- Also add screenshots of the various steps and of results if necessary.
- Add comments

Software / Platform

- Platforms are
    - MS Azure: SQL
    - (Alternatively, you can use any other database, but I cannot grant any support)
- Instead of a platform, you can run all tasks *on-premise* (on your computer):
    - If you want to perform the SQL task on your computer, just download the MS SQL Express version (which is free to use) and MSSQL Management Studio (which is also free).
    - Please note: If you want to run the tasks *on-premise*, I cannot offer broad support. Cloud based techniques are selected here, because they are state of the art and broadly used in companies today and technical support often not necessary.

Deadline

- Deadline is the day before the presentations at midnight: 18th of December 2022, midnight.
- Please use the drop-off zone in htw eMPMD system (if you experience technical difficulties, inform me!)

Task 1 – Databases / SQL: Databases with MS Azure: Data Preparations for Customer Lifetime Value Analysis and Customer Segmentation

Points: 20%

Task 1: Customer Lifetime Value

Database: NORTHWIND

The financial department wants to estimate **customer lifetime value**. Your job is to provide the data.

Table 1: Get the raw data out of the database.

Example table: simulated data (please add more variables)

customer_id	order_date	order_value	order_qty_articles
16915	2011-08-04	173,7	6
15349	2011-07-04	107,7	77
14794	2011-03-30	-33,9	-2

Table 2: Aggregate the data from table 1. Also add some customer information:

- Customer_id
- Last order_date (Recency)
- Sum of order_value (Monetary)
- Total number of orders (Frequency)

Task 2: Customer Segmentation

Database: AdventureWorksLT2019

The Strategic Management department wants to conduct a customer segmentation. They are contacting you in Data Science department to support this effort.

The idea is to get relevant customer data and to use that a basis to do the necessary segmentation. You should deliver the data.

Two kinds of variables are required:

Descriptors: Variables to describe the identified customer segments, like address data, city, country etc.

Basis variables: Basis for customer segmentation: Revenue, number of items purchased, etc.

1. Check the database for relevant features
2. Start building a dataset which can be used to perform the customer segmentation

Use SQL to perform the job and generate a CSV export.

There is no one valid solution – so please add comments why you have chosen an approach and selected a feature or not. It is possible to delete customers which have no valid data etc.

- Please describe the whole process
- Please describe – based on the database – how a customer segmentation could be performed
- Please describe which Basis variables you could selected in database and why
- Please describe which Descriptors to describe the customer segmentations you selected out of the database