Topics in Data Analytics S23
Assignment 1

May 4, 2023

1 Submission

You need to submit your analysis as an executable Python Jupyter Notebook file, to onq. This Jupyter Notebook file should be named "1234-Assn1.ipynb", where 1234 stands for the last 4 digits of your Queen's student ID.

You should use Markdown cells in Jupyter Notebook to describe the motivation of questions and findings. Make sure the TA can find your answer to each question easily.

An "I uploaded the wrong file" excuse will result in a mark of zero.

2 Background

The data is a bank card transaction and user (card) loyalty analysis dataset provided by a bank in Brazil. This assignment aims to practice on topics covered in lectures 1-3, i.e., data quality analysis, statistical analysis, and regression analysis given the dataset. The overall purpose of the analysis is to predict a loyalty score for each card_id represented in userscore.csv.

The dataset contains four files.

userscore.csv contain card_ids and information about the card itself - the first month the card was active. It also contains the predict/analysis target, i.e., score, which is a score calculated by the bank, indicating the loyalty of each card owner. Three features are provided, all of which are anonymized card categorical features.

merchants.csv contains aggregate information for each merchant_id represented in the data set. merchants can be joined with the transaction sets to provide additional merchant-level information.

The *historical_transactions.csv* and *new_merchant_transactions.csv* files contain information about each card's transactions. historical_transactions.csv contains up to 3 months' worth of transactions for every card at any of the provided merchant_ids. new_merchant_transactions.csv contains the transactions at new merchants (merchant_ids that this particular card_id has not yet visited) over two months.

historical_transactions.csv and *new_merchant_transactions.csv* are designed to be joined with userscore.csv and merchants.csv. They contain information about transactions for each card, as described above.

Given the dataset, answer the following questions. Your submission will be judged based on your answer's relevance and your regression model's performance.

3 Data Exploration

Q1 (10 points) Describe how you want to make use of *merchants.csv*, *historical_transactions.csv*, and *new_merchant_transactions.csv*, for user loyalty prediction. Note, this is a very important question, you may want to update your whole pipeline multiple times to find the best usage of three files.

Q2 (20 points) Discuss the quality of the dataset, considering missing values, missing value patterns, missing value mechanism, and noise (e.g., consistency). Note, you do not need to discuss each attribute in each table. Focus on the ones you believe are the most important for user (card) loyalty analysis and prediction.

Q3 (10 points) Perform necessary data cleaning based on your answer to Q2.

4 Statistics and Hypothesis Test

Q4 (10 points) Report important statistics in preprocessed data created in Q3. The target user loyalty score must be covered in the statistical analysis.

Q5 (20 points) Propose two hypothesis tests exploring information related to user loyalty score. For each hypothesis test, you must describe the motivation (why this hypothesis is interesting and important to test), null hypothesis, and select a proper statistical test to report the test result.

5 Regression Analysis

Randomly split your selected dataset into training and testing sets based on the results and preprocessing data from answers to above questions. Answer the following questions:

Q6 (10 points) Create a regression model for user loyalty score prediction based on the above analysis. You need to determine what features to use and which regression model to use.

Q7 (10 Points) Detect if multicollinearity exists in selected features (used in Q6).

Q8 (10 points) Build one regression model and report the performance of your model on train and test.