Group Project Description

I. Introduction

The course project is an opportunity for student groups to investigate a data mining problem that interests them. The course project should apply data mining techniques to real-world business problems. Data and software (preferably Python) for these projects can be obtained from various internet sites, or developed by students.

You can use whatever means you can find to discover interesting patterns. You do not have to use complex clustering, classification or regression algorithms for this task. Suggested topics and examples are listed as follows (**other topics are also welcome but subject to the approval of teacher before April 13, 2020**):

1) Customer profiling/ segmentation
   a) https://towardsdatascience.com/mall-customers-segmentation-using-machine-learning-274ddf5575d5
   b) https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python
2) Predictive models for customer retention or customer churn
   a) https://www.kdd.org/kdd-cup/view/kdd-cup-2009
   b) https://dianshi.bce.baidu.com/competition/24/rank
   c) https://www.kaggle.com/blastchar/telco-customer-churn
3) Mining of web logs (customer behaviors):
   a) https://www.kdd.org/kdd-cup/view/kdd-cup-2000
   b) https://www.kaggle.com/c/coupon-purchase-prediction
4) Fraud detection:
   a) http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
   b) https://www.kaggle.com/c/ieee-fraud-detection
   c) https://www.kaggle.com/mlg-ulb/creditcardfraud
5) Credit risk/scoring:
   a) https://www.kaggle.com/uciml/german-credit
   b) https://www.kaggle.com/c/GiveMeSomeCredit
6) Product/stock price prediction:
   a) https://tianchi.aliyun.com/competition/entrance/231784/introduction
   b) https://www.kaggle.com/aaron7sun/stocknews

II. Your Role

You need to form a group of 3-4 students to accomplish the project. You will provide a report containing a summary of the data mining process, as well as the conclusion and discussions for suggested actions. Your presentation should summarize your analysis and focus on your contributions by comparing with at least one baseline method.

The technical paper in Tier-1 conference in data mining “I Know You'll Be Back-Interpretable New User Clustering and Churn Prediction on a Mobile Social Application” just provides a good example for describing the approaches proposed to solve the real business problem by decomposing it to two sequential problems (clustering and then prediction).

At a minimum, your presentation and report should include the following:

1. Introduction
Introduce the background of the project, and then highlight the objective and motivation of your proposed project, finally outline your proposed methodology as well as summarize your contributions.

2. Dataset
Describe the dataset your project will work on, A brief introduction on the source, scale, range and sampling method of the dataset. Some preliminary exploratory analysis will also be welcome to analyze the general characteristics and identify potential issues of the data, such as skewness, imbalance, missing data.

3. Assumptions and Data Preprocess
Introduce what kind of assumptions are made to simply to the problem, and the following measures are taken to preprocess based the assumptions or identified issues in the previous step. Possible preprocess includes but not limited to data cleaning, feature encoding/transformation, feature normalization/scaling, feature reduction/selection.

4. Methodology
How do you decompose the big problem into small problems, and also the data mining models that you proposed to solve teach individual problem. In general, you are supposed to implement and compare **at least two** different approaches/models (with at least one baseline for comparison).

5. Performance Evaluation
Describe the framework and metrics you propose to evaluate your proposed methodologies. Usually, you are supposed to demonstrate the advantages of your proposed models/approaches over the baseline models/approaches.

6. Discussions
What conclusions can be drawn from your analysis? What additional actions or analyses would be useful to conduct in the future? What additional questions would be useful to answer that is beyond the scope of your project?

You will be graded on (1) **the relevance** of your project towards identifying and addressing an important business issue, (2) the **novelty** of your approach, (3) **the quality and depth** of your methodology, and (4) **rigorously and straightforward** evaluation and justification of your contribution and/or conclusions.

III. Deliverables and Due Dates

Electronic submission will be required on following materials with specified due time as follows.

Part 1: Report ***(Electronic Copy Due: Sunday May 3, 2020 at 23:59)***

Required (12-point font, single spaced, 10 pages at most including supporting tables, figures, and calculations)

Part 2: Presentation ***(Due: Friday May 8, 2020, tentative)***

Required:

1) 10 minute presentation and 2-minute Q&A
2) In your presentation slides, you need to include one page containing details of the job allocations among group members.
3) Not all team members are required to be active in the presentation. While those who do not present may subject to query during Q&A session.

Part 3: Slides, Cleaned Dataset & Project File ***(Due: Friday May 8, 2020 at 23:59)***

Required: In addition to the presentation slides, you should also include your Python code (or Project files) and cleaned dataset (preferably in Excel or CSV format). You may submit this as a file attachment or a file sharing link if the size of file exceeds the attachment limit (e.g., Baidu Yunpan).

V. Tips

- You can use any of external relevant datasets (please see Appendix for reference) to support your analysis.
- Start exploring your dataset EARLY to identify data issues that will be time consuming. These may include: datasets that require time consuming manual cleaning, datasets that are large and will need to be analyzed in smaller subsets or imported into a database, more data may need to be collected to conduct your analysis, or you decide to learn a new software tool that would greatly improve your analysis.
- If you are working with a large dataset, conduct your analyses over a subset of the data first. Once you have finalized what types of analysis you will conduct, you can then apply your analysis to the full sample. If it is infeasible or very costly (in terms of time or computation) to conduct your analyses over the full sample, then state this and explain whether you believe the results would be similar or different over the full sample.

Appendix: External Relevant Datasets

1. Dataset
   - Business Database: WRDS/CSMAR/Wind
   - Tableau Resources: https://public.tableau.com/en-us/s/resources
   - Kaggle Dataset: https://www.kaggle.com/datasets
   - Data Fountain: https://www.datafountain.cn/datasets
   - Tianchi: https://tianchi.aliyun.com/competition/gameList/activeList
   - UCI: http://archive.ics.uci.edu/ml/index.php
   - AWS Public Datasets: https://aws.amazon.com/public-datasets
   - KDD Cup: https://www.kdd.org/kdd-cup
   - Dataset List : https://github.com/awesomedata/awesome-public-datasets
   - Any publicly accessible data source with the help of web crawler or data API.