Birla Institute of Technology and Science, Pilani
Data Mining CS C415 / IS C415
I Semester 2017-18
Assignment: Data Challenge

Submission Deadline: 11th Nov, 2017 **Maximum Marks: 30**

- This assignment will be done in a group of two students.
- The weightage of the assignment is 15%.
- Regarding any queries, you may contact TAs (Multimedia & HCI Lab 6014) or IC.

DATASET

The dataset comprises of all transactions that have happened at the All Night Canteen, BITS Pilani for months from August to December (odd semester). The ID number of every student has been given a new random 3-digit ID to prevent any breach of privacy, however, it still preserves the program (First degree, Higher Degree and PhD) and year of study for a student. Also, year of transactions has been fictitiously set to 2020.

METADATA

The .zip package contains the following files:

1. **monthwisePriceList.csv**
   Contains the prices of each item available at ANC, at least for one month in Aug - Dec.
   - Number of Attributes : 7, Number of items : 122
   - Attributes Information
     a. *ItemID [Number]* : unique ID number for each item
     b. *ItemName [String]* : name of the item
     c. SellingPriceAug, SellingPriceSep, Selling Price Oct, SellingPriceNov, SellingPriceDec

2. **augSales.csv, sepSales.csv, octSales.csv, novSales.csv, decSales.csv**
   Each file contains the set of transactions that have happened in that particular month.
   - Number of Attributes : 7
   - Attribute Information
     a. *BillNo [Number]* : sequential numbering, starts from 1 each day
     b. *TransactionID [Number]* : sequential numbering for different items on the same bill
     c. *ItemID [Number]* : unique ID number of the item
     d. *SellingPrice [Real]* : selling price of the item. *(except in decSales.csv)*
     e. *Quantity [Number]* : how many of the item is purchased
     f. *StudentID [String]* : Fabricated ID of the student who makes the purchase

Value “F0000” is used to represent cash payment. For payment through ID card, 5-character ID is stored with the following convention. 1st character represents ”F” (first degree), or ”H” (higher degree), or ”P” (PhD). 2nd character shows year of study (1st, 2nd, 3rd, 4th, or 5th year). 3rd to 5th characters form student’s unique ID in year wise degree program.

g. SellingDate [Time] : Timestamp of the sale in ”DD/MON/YYYY HH:MM:SS” format
h. final_rating [Number] : rating given by the purchaser on the item in range 0 to 5.

- Number of Instances
  a. augSales.csv : 31773 entries

b. sepSales.csv : 77958 entries
c. octSales.csv : 77685 entries
d. novSales.csv : 56048 entries
e. decSales.csv : 37619 entries (15 days data, to be used for testing)

SUBMISSION

1. [PDF] One page executive summary for each of three problems, clearly mentioning the data mining techniques used and important parameters/results.
2. [PDF] Detailed report on the all the steps taken to derive all inferences backed with proper data-driven justification.
3. [Source Code/SPSS Streams] Well documented source code or IBM SPSS modeler stream(s) used to solve the problem.
4. **[For Problem 1 : CSV]** A price filelist will be provided named *newPrices.csv*. This is a template of prices of all items for all operational hours.
    - Attribute Information
        i. *itemID [Number]* : unique ID number for items
        ii. *timeSlot [Number]* : sequential numbering of hours from 1600 hrs to 0200 hours mapped as 16 to 23 then 0 to 2. eg. '*timeSlot* 16' means 1600 hrs to 1659.59 hrs
        iii. *sellingPrice [Real]* : selling price of the item during that *timeSlot.*

    All the proposed changes have to be given in this file by modifying '*sellingPrice*' of the required *itemID* on the required '*timeSlot*'. The modified *newPrices.csv* will be cross-verified on *decSales.csv*. [Please note that no entry has to be deleted or added, otherwise it will give spurious results.]
5. **[For Problem 2 : CSV]** A file containing each proposed combo meal in a new row with following columns:
    a. *newItemId [Number] :* new unique ID for the combo meal
    b. *oldPriceSum:* Sum of old prices of items in combo meal
    c. *newPrice:* Price of combo meal
    d. *items:* Comma separated list itemIDs
6. **[For Problem 3 : PDF]** A file containing details of questions formed along with their solutions.

EVALUATION

Following should be addressed for each problem separately.

1. There is a fixed pool of marks dedicated to each step of data mining, from preprocessing to pattern interpretation.
2. Discuss how well a data mining technique worked on these datasets. What combination of parameters yielded particularly good results?
3. Overall project conclusion: strength & weaknesses
4. Make a list of tasks that will work for any dataset and another list of tasks specific to given dataset
5. Certain constraint values in both questions are variables (**X** in Problem 1 and **Y**, **Z** & **M** in Problem 2). These will be announced later on Nalanda.
6. No marks for inferences drawn without using any proper data mining technique or without proper justification. Statistical methods can be used to study the dataset but will not be considered as a valid explanation to inferences.
7. To keep the assignment interesting, students are encouraged to regularly post their ***best achieved results*** on Nalanda, without discussing their techniques.
8. Any submission reporting results not aligned with the constraints, will not be evaluated.
9. Marks distribution will be relative keeping the best performance as benchmark.

Problem 1

The governing body of ANC needs to increase their revenue to cater the rising wages of their employees. However, before enforcing new prices they need to get the modifications sanctioned from the student body which is not in favor of the increase; though, they have given their consent for implementing a dynamic pricing scheme on items.

According to the dynamic pricing scheme, the price of an item can be increased for particular hours of sale only (like, 8pm to 10pm) and for a particular segment of students (like, 'F5 : First degree 5th yr' or 'P1 : PhD 1st yr'). For instance, the price of "Butter Chicken" can increase for ' F3 : First degree 3rd yr' from 8pm to 9pm and 12am to 2am. However, there is a penalty associated with each item (for increased price), given by the following formula -

$$p_i = changeInPrice_i * hourWeight * segmentWeight$$

where pi is the penalty associated with item i, $changeInPrice_i$ is the difference between the new price and the old price of that item. $hourWeight$ depends upon the number of hours the new price is in effect on a particular day, as given by the table :

No of Hours	1	2	3	4	5	6	6+
hourWeight	1	4	9	16	25	36	50

$segmentWeight$ depends upon the group for which the price increase is enforced, given by the table:

Segment	F1	F2	F3	F4	F5	H1	H2	others
segmentWeight	12	32	30	20	3	2	2	1

The objective is to propose a new pricing scheme (according to the concept of dynamic pricing) so as to achieve the maximum $\%\_increase\_on\_total\_revenue$ (for month of December), with the constraint that $total\_penalty$ is minimal.

$$\%\_increase\_on\_total\_revenue = \left(\frac{total\ revenue\ by\ new\ price\ scheme\ in\ Dec}{total\ revenue\ by\ old\ price\ scheme\ in\ Dec} - 1\right) * 100$$

$$total_{penanlty} = \sum_{i:\, items} p_i$$

There are a few constraints which needs to be fulfilled:

1. The minimum % increase on revenue that needs to be achieved is $\mathbf{X}$%. There is no upper bound.
2. The price of an item can only be increased to a max of 10% or Rs.10 (whichever is minimum).
3. There can be only one 'new price' associated with an item, i.e. an item original selling for Rs.50 cannot be sold at 52 at one instance and 55 at other instance.

In order to keep the penalty minimal, prices needs to be altered only for the targeted segment of students and for a targeted bracket of time. Prices ***can*** be increased for whole duration of sales and for all segments of students, however, this is discouraged as evident by the distribution of *hour_weight* and *segment_weight* - since it would increase the penalty considerably. Teams are advised to apply data mining techniques to find a target segment of

students and target bracket of time for which price increase on items can be most effective. For example, if an inference says that 'First degree 2yr' students eat 'Veg burger' predominantly between 10pm – 12am, then a price increase on such segment can increase revenue insuring a minimum penalty.

Problem 2

The students are asked to rate items, after the purchase, on scale of 0 to 5 with 0 being for least satisfied and 5 being most satisfied response. By making use of the individual ratings on items, the average rating for an item can be computed which are indicative of the items' popularity among students. The ANC governing body fears that the sales of less popular items (having low ratings) will fall in coming months. To tackle such a scenario, they have decided to offer `combo meals`, with an intention to maintain (or even ramp up) the sales of such items.

Combo meals is the concept in which pair of K items are put up on sale for a price that is (**Y**/K)% less than the sum of prices that would have been if the items were bought separately. For example, if price of 'Coke' is Rs. 35, 'plain maggi' is Rs. 20, **Y** is 20% and K is 2 then they can be offered as combo for [(100-(20/2) % of (35 + 20) → Rs. 49.50.

The governing body wants to form 'combo meals' in such a way to tempt the students in buying less frequent/popular products along with popular ones. As a matter of fact, all items in the combo should complement each other in a *sensible* way. For example, the combo of 'Coke' and 'Pulpy Orange' is not suitable and will not sell. Another strategy to form combo meals is to club few less frequently bought items which are always bought in group i.e. they occur together whenever they occur. The governing body has put a **Z** % cap on maximum loss of profit due to combo meals.
The objective is to find maximum number (at least **M**) of as large as possible combo meal groups such that loss in profit is minimal.

Problem 3

Figure 2 or more specific and interesting questions **about the domain** that you want to answer with the data mining techniques. Try to choose questions that are about the domain not about the techniques or the experimental parameters.

Performance Metrics

Explain what performance metric(s)/formulae will be used and why to evaluate the models that you have constructed in order to answer questions identified.

Model Interpretation

Describe the resulting model (e.g., size of the model, readability, accuracy, etc.). Summarize the model in your own words focusing on the most interesting/relevant patterns. Elaborate on if and how the model answers the objectives identified.