

**Certificate Program in Business Analytics and Intelligence – Batch 9**

**Assignment 5: Advanced Analytics (Total Points – 105)**

**Instructions**

1. This is a **take-home** assignment.You are free to discuss the assignment questions with your classmates. However, you are not allowed to copy the answers from other students.
2. Answer all questions.
3. **Show all work and give adequate explanations to get credit.**
4. Encircle or underline your final answer for each part.
5. **Completed assignment MUST BE uploaded on Moodle by 31th March 2019, 23.55 Hours.**
6. **Assignments will not be accepted after 31th March 2019, the grades of such students will be marked Incomplete (I) in the grade sheet.**

**Question 1 (50 Points)**

Read the case, “Markdown optimization for an Apparel Retailer”. Carryout the following activities (data sets are uploaded in the Moodle).

**PART A – CLUSTERING**

1. List and derive the metrics that can be used in ‘‘hierarchical clustering’’ and ‘‘partition around medoids’’ clustering algorithms. (**Note: Use the data in sheet “Clustering\_Raw\_data.xlsx”**).
2. Do you find outliers in the derived data from Q1? If yes, how can the same be treated for use in cluster modeling?
3. Develop a hierarchical clustering model with the modified data from Q2. How many clusters seem appropriate? Justify.
4. Develop a partition around medoids clustering model with the modified data from Q2. What are the advantages of using partitioning around medoids (PAM) over *K*-means? How do you decide on the appropriate number of clusters in this scenario?
5. Validate the goodness of resulting clusters from hierarchical and PAM models obtained in Q3 and Q4. Which is a better model as per validation measures?
6. Having selected the most appropriate clustering model from Q5, perform cluster profiling and list the unique characteristics of each cluster.

**PART B – TIME SERIES FORECASTING**

1. Conduct exploratory data analysis on “Cluster=1, Brand=CRESCENT SET, Brick=CKD” combination to identify the following relationships. Give a short description about the relationships observed.
   1. Relationship between Sales units(sales\_units) & Discount % (discount\_per)
   2. Relationship between Sales units & Net Price (per\_unit\_netprice)
   3. Relationship between Sales units & Age (age)

Note: 1) **Use data in the following csv files “1. CRESCENT SET.CKD.csv”** to answer this question.

2) Variables names to be used are mentioned in brackets.

1. What is over-fitting and under-fitting in the context of regression models? What are the consequences of over-fitting?
2. Explain why we have to partition the time series data before building the forecasting model. Use data for “Cluster=2, Brand=BLINK, Brick=HAREMS” and partition this time series data as explained below.
   * 1. Consider all weeks until 51st week (including) of 2014 as training data.
     2. Consider weeks from 52nd week of 2014 to 3rd week of 2015 as test data.

Note:

* These 4 weeks are winter EOSS weeks.
* Use data in “**2.BLINK.HAREMS.csv**” only for answering questions from now onwards.

1. Develop a time series forecast model using regression on the training data to forecast sales units for “Cluster=2, Brand=BLINK, Brick=HAREMS” combination using the below variables as predictors.
   1. Lag 1 (i.e. immediate previous week) of sales units
   2. Discount %
   3. Lag 1 (i.e. immediate previous week) of discount %
   4. Promotion week flag
   5. Age

Apply appropriate transformations and evaluate the model fit.

1. Perform checks to ensure that the model is valid, and assumptions of regression are met. Conduct appropriate statistical test and back the findings by visual examination of relevant plots.
2. Based on the model result, explain the following:
   1. How is this forecasting model able to account for trend and seasonality?
   2. What is price elasticity and determine the price elasticity value (a proxy representing price elasticity is enough) from the model output?
   3. How do you interpret the coefficient of promotion week flag variable?
3. How do you check if the forecasting model is able to explain most of the important features of the time series? Explain white noise in the context of time series.
4. Using the forecast model built, generate sales units forecast for test period (52nd week of 2014 to 3rd week of 2015).
   1. Assess the forecast model accuracy on the test time period which is not used for modeling by calculating MAPE for the test period.

Note: In **“Forecast\_test\_week\_predictor\_input.xlsx”** sheet “Input data for forecasting” contains the data required for forecasting sales and sheet “Actual Sales” contains the actual sales generated.

**PART C – OPTIMIZATION**

1. Formulate an optimization model and solve it to determine the optimal discount % to be given for “Cluster=2, Brand=BLINK, Brick=HAREMS” combination for each of the 4 weeks of EOSS.
   1. Objective function: Maximize the total revenue generated during EOSS.
      1. Total revenue is defined as the revenue generated during 4 weeks of EOSS and revenue from the left-over inventory after EOSS. Assume the residual inventory left over even after EOSS is liquidated by giving a flat discount of 60%.
   2. Constraints
      1. Sales units in a week should be less than equal to starting inventory of the same week.
      2. Relationship between demand (sales unit) for a week and predictors should follow the relationship identified in the forecasting regression model.
      3. Discount for a week should be greater than equal to previous week’s discount and discount should vary between 10% and 60% only.
      4. Sales & inventory values for a week should be ≥ 0.
      5. Inventory should be updated on basis of the sales for a week.
   3. Inputs required
      1. Inventory at the beginning of EOSS = 2,476
      2. Age of the brick at the beginning of EOSS = 96 weeks
      3. Previous week discount = 57.9%
      4. Previous week sale = 48
      5. MRP of the brick = INR 606
2. What are the weekly forecasted sales units if the optimal discounts identified are implemented for the EOSS?

We know that actual revenue realized by the retailer for “Cluster=2, Brand=BLINK, Brick=HAREMS” combination during the 4 weeks of EOSS is INR[[1]](#footnote-1) 41,320. Then, what is the incremental lift in revenue the retailer would have achieved in these 4 weeks if he/she implemented our analytics solution instead?

**Question 2 (50 Points)**

Read the case, “Enhancing visitor experience at ISKCON using Text Analytics”, and answer to the following questions.

*(Case and Dataset: “ISKCON Case” and “ISKCON Data” is uploaded on moodle)*

* Question 1: What are the issues with text data and discuss what type of pre-processing would be required for text data?
* Question 2: Discuss about few exploratory data analysis (EDA) and trend analysis that can be carried out on text data.
* Question 3: How to perform feature extraction and selection from text data for sentiment analysis
* Question 4: Discuss the sampling strategy you would be using for the given data set.  What will be your strategy for dealing with the imbalanced data set?
* Question 5: Develop an appropriate supervised machine learning model to predict sentiment of a review
* Question 6: Discuss about lexicon-based approaches to predict the sentiment of a review
* Question 7: Develop an appropriate model to predict emotion of a review
* Question 8:What are the overall insights from this text mining process?
* Question 9:What should be the strategy for deployment of the model?

**Question 3 (5 points)**

Comments on a hotel at “Trip Advisor” is provided below:

Positive Comments:

1. Service was very good. Excellent breakfast in beautiful restaurant included in price. I was happy there and extended my stay for extra two days.
2. Really helpful staff, the room was clean, beds really comfortable. Great roof top restaurant with yummy food and very friendly staff.
3. Good location. The Cleanliness part was superb.
4. I stayed for two days in deluxe A/C room (Room no. 404). I think it is renovated recently. Staff behaviour, room cleanliness all are fine.

Negative Comments

1. The room and public spaces were infested with mosquitoes. I killed a dozen or so in my room prior to sleeping but still woke up covered in bites.
2. Unfriendly staff with no care for guests.
3. Very worst and bad experience, Service I got from the hotel reception is too worst and typical.

Use the following vocabulary set:

V = {Beautiful, Good Service, Good Location, Superb, Cleanliness, Mosquitoes, Unfriendly, bad experience}

Use Naïve Bayes’s algorithm and calculate the probability of positive sentiment for the following comment: “**Good location but the staff was unfriendly”.**

--------------------------------------------------END OF ASSIGNMENT ----------------------------------

1. 1 USD = INR 65.8 in May 2016 [↑](#footnote-ref-1)