FINAL ASSIGNMENT- GROUP

BUS 4023- WINTER 2021

35 MARKS

COVERS MATERIAL FROM CHAPTERS (1-6,9,10,15)
TOPICS COVERED :
DATA EXPLORATION, LOGISTIC REGRESSION, DECISION TREES AND CLUSTERING

CASE STUDY (PAGE 537 OF TEXTBOOK)

21.6 SEGMENTING CONSUMERS OF BATH SOAP

#1) Perform a cluster analysis using k-means. Remember to normalize the data and try a few different cluster values i.e., k=3, k=4 etc. Before coming up with your final cluster set characterize the clusters i.e., analyze the descriptive statistics, comment on the characteristics (demographic, brand loyalty, basis of purchase etc.) of these clusters and create personas for each of the clusters. Come up with some creative persona names for each of the clusters that summarize their characteristics well e.g., 'loyal price conscious'. Think of a creative visual to show case the difference in the clusters and present them

Further Instructions/Hints:

- *With your group, go through the dataset properly. Read the case and variable information from the textbook to get a good domain knowledge related to the issue. You need a good understanding of the variables.*
- *Group the variables according to demographic indicators, Brand Indicators, Behaviour Indicators, Purchase Indicators etc. Use your judgement for grouping the variable and take information from the case. (Hint: for brandwise purchase create a variable 'maxbrandpurchase' to show which brand is purchased most by the households).*
- *Perform cluster analysis using K=2 first. (Do not use demographic data/demographic indicators for clustering).*
- *Comment on the characteristics (demographic, brand loyalty, basis of purchase etc.) of these clusters and create personas for each of the clusters. Come up with some creative persona names for each of the clusters that summarize their characteristics well e.g., 'loyal price conscious'. Think of a creative visual to show case the difference in the clusters and present them. (Hint: Use code: bathSoap_df[demographicIndicators].groupby(clusters.labels_).mean() to group clusters by demographic variables)*
- *Perform cluster analysis using K=3 and 4 (not more than 4) and characterise these as well.*
- *Groups which come up with well-defined personas, characteristics of the clusters and visualization will get better marks.*

#2) Develop a predictive model to classify clients as value conscious or not. Use a binary logistic model. Use the segments/clusters to identify who is value conscious i.e., 1 vs. 0. The model will be used to mail these clients a promotion. ~~If the cost to mail was $5 and total expected profit from value conscious clients who received the mailing was $250. What would be the estimated profit gained from mailing the value conscious clients based on the model results versus choosing a random sample of clients to mail?~~ (Not required for the assignment) Use data visualizations to present the business value of your final model.

Hints:

- *Use the K=2 cluster and define one group as successful group or Y=1, the other group as Y=0. Partition the dataset and fit the model. Use confusion matrix to determine accuracy. Show some lift curves, ROC to provide more details of the modelling.*
- *You can use decision trees to look at the bins or find other ways to transform the data if need be.*

Present your findings in power point format in terms of steps taken within the Data Mining framework and the final results. Why did you select the final model over other model you might have come up with? Please submit Python code and output as well

Tell a high-level story of steps taken to get to the end result. Start with the framework i.e., objective, exploration, output results of the analysis, and final recommendations.
Keep it simple, engaging, creative and have fun!

DELIVERABLES FOR ASSIGNMENT INCLUDES

- You will be preparing a video presentation having no more than 15 slides (PowerPoint has the option for recording audio and video with the slides. Try to add the video as well to make the presentation engaging)
- Maximum presentation time 15 minutes per group
- Every team member must present to get awarded the points. Each team member needs to present on one of the topics in the presentation i.e., methodology, background, results, reporting, recommendations/conclusions. Just introducing the group in the presentation does not count.
- Tell a high-level story of steps taken to get to the end result.
- Use visualizations. You can use any other applications such as Tableau, PowerBI, Excel as well if you want.
- Keep it simple, engaging, creative and have fun! Remember your audience (Business users on the Marketing team, who are not highly technical)

Submit the following in BB:

- Upload the presentation in OneDrive and share the link
- The presentation file
- The Python code (1 single ipynb file)

High-level breakdown of the Points awarded are as follows:

- Data Mining Methodology, output and final results – **15 points** (team)
- In-class presentation team– **10 points** (team) (does the story flow well and is it easy to follow, pay attention to aesthetics and visuals)
- In-class presentation individual part: style, presentation, content etc.– **10** points (individual).

Details:

1. **Data Mining Methodology/Framework– 5 points** (Are all the steps laid out in a clear, step by step manner, does the overall data mining approach for the project align with the framework)
2. **Output of Data Mining tasks –** (Python code from each of the steps that align with the data mining framework e.g., data exploration, variable reduction, transformation efforts like Decision Trees for finding bins; dummy variables, data partitioning, running variations of models, cluster variations i.e., k=4, k=5; elbow chart) . *Please make sure to add comments to each of the sections in your code to call out the steps taken at each point.*
3. **Final results –** (Python Code and output detailing : Final cluster results, variables contributing to defining the final clusters, model comparisons, final model selection for promotion including results from training and validation, lift and gain charts and finally net profit)
4. **Presentation –** (Are the results from the above 3 sections reflected in the final presentation? Do the results add up and is it easy to find the output used in the presentation in one of the 3 sections above? Is the story easy to interpret, is based

on a sound approach and assumptions? Visuals are easy to interpret, and final results align with the objectives and business goals).

The assignment will be marked according to ranking. The best presentation will get the best mark and all other presentations will be judged according to the best one.