BUSINESS DATA MINING (IDS 572)

HOMEWORK 5
DUE DATE: FRIDAY, APRIL 29 AT 11:59 PM

- You need to complete this assignment in Rmarkdown.
- Please explain your findings and thoughts in detail.
- Submit an electronic pdf file along with your Rmd file in blackboard.
- Please include the names of all team-members in your write up and in the name of the file.
- One submission is sufficient for the entire group.

Your HW5 is the case study "Champo Carpets: Improving business-to-business sales using machine learning algorithms". You can purchase this case from HBP.
You can work with the suggested samples sets for different tasks. However, you should use the raw data to add additional features to your models (if needed). After cleaning data, feel free to take any suitable approaches to answer the following questions.

(1) With the help of data visualization, provide key insights using exploratory data analysis.

(2) What kind of analytics and machine learning algorithms (e.g. classification, regression, clustering, recommender systems and etc) can be used by Champo Carpets to solve their problems, and in general for value creation? Justify your choices.

(3) Develop ML models (e.g. logistic regression, decision trees, random forest, neural network, and boosting) to help identify features that contribute toward conversion (or non-conversion) of samples sent to customers. Hint: For each model, discuss how you select features, and how you tune different parameters. How do you evaluate the performance of each model? How do you select the best model(s). Run all your models on both balanced and imbalanced data and check the difference.

(4) Discuss the data strategy for building customer segmentation using clustering. What are the benefits Champo Carpets can expect from clustering? Hint: Data strategy should clearly identify the data that should be used and how it should be used, including any feature engineering that may be performed before the model building.

(5) Discuss clustering algorithms that can be used for segmenting Champo Carpets's customers. Please justify your choices. Discuss what distance and similarity measures is suitable in this case.

(6) Develop customer segmentation using k-means clustering. Discuss the optimal number of clusters., significant variables, and cluster characteristics.

(7) Write your own collaborative filtering function as recommender system. Hint: Collaborative filtering technique is based on an aggregation of customer purchase history. For each customer, you can use various measures such as Pearson correlation, Euclidean distance, or cosine similarity to find the nearest neighbors. You can then use the nearest neighbors to recommend products. For example, suppose using cosine similarity, you find out that the closet customer to customer H-2 is customer T-5. Customer T-5 has purchased carpet type double black and gray color, which are not purchased by H-2. Hence these products can be recommended.

(8) What will be your final recommendation to Champo Carpets?