Project - Data Mining

Please find below the Data Mining Project instructions:

* You have to submit 2 files :   
  1. **Answer Report**: In this, you need to submit all the answers to all the questions in a sequential manner. **It should include the detailed explanation of the approach used, insights, inferences, all outputs of codes like graphs, tables etc.** Your report should **not** be filled with codes. You will be evaluated based on the business report.
  2. **Jupyter Notebook file**: This is a must and will be used for reference while evaluating
  3. **PowerPoint Presentation**: A brief about the project, findings, insights and recommendations for the business. DON’T include codes in the PPT and only include relevant graphics.

**Problem 1: Clustering**

A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. Please note that it is a summarized data that contains the average values in all the columns considering all the months, and not for any particular month. You are given the task to identify the segments based on credit card usage.

1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).

1.2 Do you think scaling is necessary for clustering in this case? Justify

1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them

1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.

1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.

**Data Dictionary:**

1. spending: Amount spent by the customer using the credit card per month (in 1000s). For example, if the spending is 19.94, then the customer has actually spent (19.94 \* 1000 = 19940) 19940 Rs per month on an average.
2. advance\_payments: Amount paid by the customer in advance by cash even before the credit card bill got generated for any particular month (in 100s). For example, if the advance\_payments is 16.92, then the customer has paid (16.92\*100 = 1692) 1692 Rs on an average per month.
3. probability\_of\_full\_payment: Probability of the credit card payment done in full by the customer to the bank. If it is 0.8752, then it means that the customer has a chance of 87.52% to pay the entire credit card bill on an average per month.
4. current\_balance: The balance amount left in the credit card account to make the future purchases (in 1000s). For example, if the current\_balance is 6.675, then it means that the customer is left out with a credit card balance of (6.675\*1000 = 6675) 6675 Rs which he can use for the future purchases.
5. credit\_limit: Limit of the amount in credit card (10000s) sanctioned by the bank to the customer. For example, if the credit\_limit is 3.763, it means that the customer has been sanctioned a credit card limit of (3.763\*10000 = 37,630) 37630 Rs.
6. min\_payment\_amt : The average minimum amount paid by the customer while making payments for the credit card bill purchases made monthly (in 100s). For example, if the min\_payment\_amt is 3.252, it means that the customer has paid only (3.252\*100 = 325.2) 325.2 Rs as the minimum payment instead of paying the entire credit card bill amount on an average per month.
7. max\_spent\_in\_single\_shopping: Maximum amount spent by the customer for a single transaction using the credit card (in 1000s). For example, if the max\_spent\_in\_single\_shopping is 6.55, it means that the customer has spent a maximum of (6.55\*1000=6550) 6550 Rs for a single transaction using credit card on an average per month.

***Dataset for Problem 1: bank\_marketing\_part1\_Data.csv***

­­\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Problem 2: Decision Tree, Random Forest & Artificial Neural Network**

An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the models' performances in train and test sets.

2.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).

2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network

2.3 Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score, classification reports for each model.

2.4 Final Model: Compare all the models and write an inference which model is best/optimized.

2.5 Inference: Based on the whole Analysis, what are the business insights and recommendations

**Data Dictionary:**

1. Target: Claim Status (Claimed)

2. Code of tour firm (Agency\_Code)

3. Type of tour insurance firms (Type)

4. Distribution channel of tour insurance agencies (Channel)

5. Name of the tour insurance products (Product)

6. Duration of the tour (Duration in days)

7. Destination of the tour (Destination)

8. Amount worth of sales per customer in procuring tour insurance policies in rupees (in 100’s)

9. The commission received for tour insurance firm (Commission is in percentage of sales)

10.Age of insured (Age)

***Dataset for Problem 2: insurance\_part2\_data-1.csv***