Assignment 1

Question 1. What dataset will you use for your course project? (describe your dataset, and either include a link to where I can find it online or submit it as a spreadsheet (.csv) or data file along with your report).

The name of the dataset is **Bank Marketing Campaign.** This dataset, as the name suggest is about a campaign launched by a bank in-order to encourage people to open a term deposit in their bank. This survey took in consideration many factors/variables, including demographic, financial, and social-economic information about the bank's clients, along with details of the marketing interactions.

Link : https://www.kaggle.com/code/janiobachmann/bank-marketing-campaign-opening-a-term-deposit

Question 2. Describe the dataset. What kind of data does it contain? What data sets do you plan to use?

The main purpose of this campaign was to persuade clients to open a term deposit in their bank. It includes data points like:

- Client Information: Features like **age, job, marital status and education**.
- Banking information: Features like **Balance, credit status, previous outcome of the** marketing
- Campaign Information: **Client contact information, and type of contact, and if they opened a term deposit or not.**

The target variable:

The target variable has a low cardinality. This is a Binary Categorical feature, which only has the value "yes" and "no". That means this feature will require Binary encoding.

Question 3. Is there anything about your data you do not understand? (i.e. what do column headers mean)? How will you find this out?

There is a feature that

Upon close observation, we decided to drop the "poutcomes" feature from the dataset because it is 85% missing in values, and even if we filled in the values by any sorts of imputation, this values are synthetic, which might lead to overfitting. The thing is decision tree is very prone to overfitting so we avoid any noise.

4.Why this dataset? What questions we are trying to answer?

The purpose of this dataset is to predict the behaviour of the customer to optimize future marketing campaigns.

- Identify the key factors: By analysing the demographics, financial profiles and the past interactions, we can understand which factors were most influential.
- What are the strongest factor that influence the target variable the most.

- How does the subscription rate vary across different demographics.

- We can feed this data into decision tree and it will split the dataset into features creating branches that leads to prediction. The decision tree can interpret which factors are influential in predicting a clients likelihood of opening a bank account. It will classify clients into categories, “likely to open a bank account” and “not likely to open a bank account”.

- We can feed this dataset into the k means clustering and find clusters based on the similarities in the feature. This clusters can be used for segmentation. This will allow us to understand the clients and tailor marketing strategies for each segment.

5. What is the main Problem ?

This is the historical data that the banks have now from the previous marketing campaign. Now if the bank uses this dataset to train the ML model, they can understand which group of people are more likely to open a term deposit. That way , if they are hoping to run a new campaign after few years of the first campaign, they know the demographics of the people they will focus on. This way they can efficiently allocate their resources. We can use this dataset to predict which group of people are more likely to open a bank deposit.

Question 6. If you must do significant work to get the data or convert it into the proper format, describe the process and approximate effort required.

There is lots of cleaning and data preprocessing that needs to be done with this dataset. Though we are the early steps of the cleaning, we already have planned out our task. First we looked for if there is any duplicate observations or not, then any inconsistencies with the data. Then we had to look for any missing values. There were quite a few, in 4 features to be specific. We used conditional imputation to handles the missing values. Then we will look for outliers, and to our surprise, every feature except the day feature had outliers less than 2%. We have to have a carefull approach on how to deal with them. Then we have to standardize and normalize the data to feed it into the ML model so that there is no overfitting.

Question 7: What relevant references will you read or examine?

a. Moro, S., Cortez, P., & Rita, P. (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. b. Han, J., Kamber, M., & Pei, J. (2011). **Data Mining: Concepts and Techniques** (for understanding algorithms). c. Bank Marketing Dataset Documentation on UCI

Q8. How will you formulate the problem as a data mining problem ?

- The ML model that I am going to use is K means clustering, which is a classification algorithm, which will give me certain groups of different types that will give me a better understanding of the customers as a whole, Q9. What exactly are you trying to predict, and how will you evaluate your results?

- To predict that in future which customers are going to open a term deposit and who are not.
- I will use accuracy, precision , f1 score, recall to assess the performance of the ML algorithm ie the quality of the prediction. I use the cross validation to ensure the models generalizibility.

10. What can you compare them to ?

- Beside using the two must algorithm, I will use logistic regression as a baseline classifier. For clustering, I will compare the quality using silhouette scores.

11. What data mining techniques do you plan to use .

- Decision tree : To create a rule that will predict whether a client will subscribe or not based on their profile.

-K means Clustering: To make groups of people and identify patterns that will allow us to spot a pattern or trend to improve the campaign

12. How are you going to evaluate your results ?

- **Qualitative Evaluation**: I will use heatmaps, bar charts, and decision tree visualizations to illustrate insights and decision pathways.

- **Quantitative Evaluation**: Performance metrics like accuracy, F1-score, and confusion matrices will measure classification performance. For clustering, I'll use the **silhouette score** to assess cluster quality and validate the segmentation approach