King Saud University
College of Computer and Information Sciences
Information Systems Department

IS 463
Introduction to Data Mining
Second Semester 1446

Customer Segmentation Based on Marketing Response

Team Members:
Elaf Alshehri, Afnan Alsuliman, Lama Albediwy, Sheikha Alhaqbani, Asma Albahkly, and Shaihanah Alhomidan

Research Area: Classification

Dataset: Bank Marketing Campaign Dataset[1]

Problem Statement and Data Mining Techniques

Predicting customer responses to marketing campaigns is crucial for improving conversion rates in the banking sector. A key challenge is segmenting customers based on their likelihood of subscribing to a term deposit, enabling more targeted and efficient marketing strategies.

In this project, the *Bank Marketing Campaign Dataset*[1] will be used, it contains data from marketing campaigns conducted by a Portuguese banking institution. It includes information about clients, such as their age, job type, marital status, education level, and financial details like account balance, housing loan status, and personal loan status. It also provides information about the marketing campaign, such as the number of previous contacts, the last contact date and duration, and the type of contact (telephone or cellular). The dataset also includes information on past marketing campaign outcomes and whether the client subscribed to a term deposit. It has 45,211 instances, 17 input attributes, and 1 output attribute, the dataset is widely used and perfect for classification tasks, particularly to predict customer responses to marketing efforts and improve targeted marketing strategies.

To prepare the dataset for training, various data mining techniques will be employed to enhance model performance and ensure data quality. First, categorical variables such as job and marital status will be converted into numerical representations using encoding techniques, allowing the data to be effectively processed by machine learning algorithms. Next, to reduce computational constraints, stratified random sampling will be applied on (bank-additional-full.csv) dataset, this technique will ensure that both classes (subscribed vs. not subscribed) are proportionally represented in the sample, preserving the overall class distribution and preventing data imbalance issues. Also, to maintain consistency across features, numerical variables like age, balance, and duration will be normalized, ensuring they are on a similar scale for better model performance. Finally, any missing data will be handled through imputation or removal to prevent inconsistencies that could affect the accuracy of predictions. These steps will collectively improve the dataset's suitability for the classification task.

[1] https://archive.ics.uci.edu/dataset/222/bank+marketing