Abstract

This project centers on analyzing a dataset from a Portuguese banking institution's direct marketing campaigns available in the UC Irvine website, primarily conducted via phone calls. The dataset can be found at the following link: http://archive.ics.uci.edu/dataset/222/bank+marketing and comprises two subsets: "bank-full.csv," including examples from May 2008 to November 2010, and "bank.csv," a 10% random sample. The goal is binary classification-predicting if clients will subscribe ('yes') or not ('no') to a term deposit.

In the realm of financial marketing, understanding customer behavior and campaign effectiveness is paramount. This dataset's unique aspect, involving repeated phone contacts and client engagement, makes it an intriguing subject. The main goal is to identify factors influencing term deposit subscriptions and construct predictive models for future marketing strategies.

The aim is to answer three fundamental questions: What drives clients to subscribe to a term deposit in this campaign, can we build accurate predictive models from the data, and how we can improve marketing campaigns?

The dataset contains 45,211 records in "bank-full.csv" and 4521 in "bank.csv." It features 16 input attributes encompassing client data and one output attribute ('y') indicating term deposit subscription. Attributes range from numeric variables like age, balance, and contact duration to categorical descriptors like job type, marital status, education, and communication method. Binary indicators for credit default, housing loans, personal loans, and campaign history, alongside numeric features detailing contact frequency and timing, are also included in the dataset.

The approach involves thorough data preprocessing, addressing missing values, encoding categorical data, and scaling numeric features. Various classification models can be explored for predicting term deposit subscriptions, including logistic regression, decision trees and random forests. Model performance is assessed using accuracy, precision, recall, and the F1-score. Python serves as the primary programming language for data manipulation, analysis, and visualization.

In conclusion, this project aims to unveil the key determinants of term deposit subscriptions in a banking marketing campaign. By comparing different modeling techniques, it seeks to provide actionable insights for refining future marketing strategies and enhancing campaign success