Executive Summary

Overview

During the last campaign, bank contacted 5,000 customers to offer the personal loan. However, the last campaign showed only 9.6% success rate, in which only 480 customers accepted the personal loan offered by the bank. This paper is a study in the application of data mining to build strong classification models to identify the important attributes of customers and to predict whether or not the customers will accept the personal loan offered by the bank.

Objective

The goal of this project is to identify the important characteristics of bank's customers who are likely to accept the personal loans offer and to develop strong predictive models to predict campaign outcomes which determine whether or not the customers will accept the loan offer.

Problem Description

The study shows that the last campaign had a low success rate. Thus, bank needs to identify important attributes of potential clients and focus their marketing efforts on those specific customers. We analyzed data to address the following questions: Which types of customers are more likely to accept the loan offer? Which attributes have high influence on target attribute? What are predicted campaign outcomes?

Dataset

The "Bank Personal Loan" dataset contains 5,000 records, retrieved from Kaggle. The dependent variable is a binary variable (Did client accept the personal loan offer?). The 13 independent variables include 7 numeric attributes, 2 categorical attributes, and 4 binary attributes.

Methodology

The models were developed using classification algorithms: Decision Tree, Naive Bayes (Gaussian) classifier, Linear Discriminant Analysis, and K-Nearest Neighbors Classification. Several Python libraries such as Pandas, Matplotlib, NumPy, sklearn, and imblearn were utilized in this study. The following steps of KDD process were implemented for this project:

- Explored the distribution and basic statistics and visualized data of all attributes
- Performed correlation analysis and preprocessed data by dropping some attributes, normalizing numeric attributes, creating dummy variables for categorical attributes, and performing PCA.
- Divided data into a training set (80%) for fitting models and a test set (20%) for model evaluation
- Developed models and performed model evaluation using a balanced accuracy score and index of balanced accuracy (for an imbalanced class) to select final model for making a prediction

Model	Precision	Recall	Specificity	f 1	Balanced Accuracy Score	Index of Balanced Accuracy
Tree 1	98.0%	98.0%	90.0%	98.0%	94.0%	88.0%
Tree 2	97.0%	97.0%	83.0%	97.0%	90.0%	82.0%
Tree 3	98.0%	98.0%	87.0%	98.0%	92.0%	86.0%
Naive Bayes	89.0%	87.0%	64.0%	88.0%	74.0%	56.0%
LDA	94.0%	94.0%	68.0%	94.0%	79.0%	64.0%
KNN 1	95.0%	95.0%	67.0%	95.0%	79.0%	64.0%
KNN 2	91.0%	92.0%	57.0%	91.0%	71.0%	52.0%

Summary

The final model is Decision Tree #3, which generates 11 classification rules. The final model exhibits high accuracy for prediction. We use final model to predict the campaign outcomes for new records and to identify important attributes of client's characteristics. The most important attributes are Income, Education Level, and Family Size. The least important attributes are Online and Area.