Data Mining Coursework

Junaid Ali Rasheed

March 25, 2018

Abstract

Contents

1 Introduction

We have been provided with a dataset which contains information about customers who were targets of direct marketing compaigns of a Portugese banking institution. Our task is to develop models on this dataset to determinte whether a customer will make a term deposit or not. We will use equal and unequal costs to develop the model.

2 Data Exploration

The dataset was provided in the ARFF file format which can be used with Weka. Opening the file in a text editor gives us a brief description of each attribute in the dataset. An exploration of the dataset using a text editor and the Weka Explorer interface reveals the following...

- The number of clients (people who were contacted by the banking institution) in the dataset is 36,188.
- 31,981 clients did not make a term deposit, 4188 clients did make a term deposit.
- 88.4% of clients did not make a term deposit. This is the accuracy of the default classifier.
- There are 16 attributes (excluding the output attribute, 'termDeposit').
- Reducing the number of attributes may be beneficial to avoid overfitting.
- There are no missing values. However, some attributes have an 'unknown' value.
  - Job: Only 223 clients had an unknown job.
  - Education: Only 1480 clients had an unknown education.
  - Contact: 10417 clients were contacted by unknown means. The histogram (see Figure 1) shows that clients contacted by a cellular device (23416) were more likely to make a deposit compared to clients contacted by telephone (2336), so it would have been useful to know how all clients were contacted
  - Previous campaign outcome: The value of this is 'unknown' for 29621 clients. This abnormally large number may be due to the fact that 29616 clients were never contacted for previous campagins.

- Viewing the histograms for each variable showed that clients who had defaulted would never make a term deposit
- Some attributes distribution of values were quite imbalanced.
  - Balance: Strongly skewed distribution. Most clients had a low/negative balance.
  - Month: 25800 clients were contacted in May, June, July, and August.
  - Previous days: 29712 clients had a value of -1, meaning they were not previously contacted
  - Previous: 29616 clients were not contacted for previous marketing campagins.
  - Job: There are 13 values but 7 of them include only 7267 clients
- Two-dimensional scatter plots do not show any strong class separation for any of the attributes. From this, we can infer that several attributes will be needed to determine whether a client will make a term deposit or not.

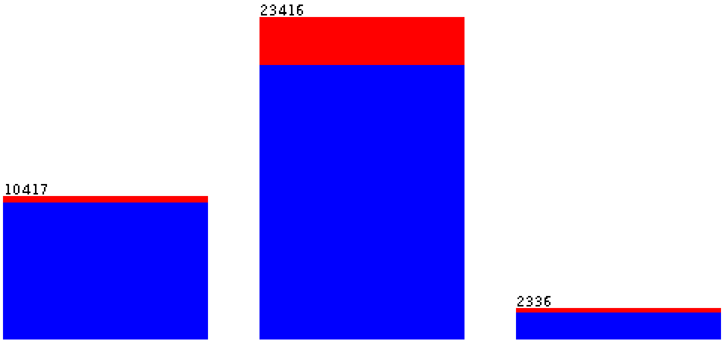

Figure 1: Histogram of the contact attribute

3 Data Preprocessing

4 Classification Modelss

5 Conclusion