TEAM 8

ISHAL ABHISHEK MUMMIDIVARAPU RAMYA KONDRAKUNTA SAI VENKATA MANOJ VUNGARALA

Project Proposal

Introduction

In this project we use *data* of a Portuguese Bank Institution's marketing campaign through phone calls. Analysing this data, we intend to predict whether a candidate will subscribe to their term deposit or not. The project is a data-driven approach to predict the success of telemarketing.

Data analytics

The dataset contains 20 input variables and 1 output variable. Of the 20 input variables, we believe the following variables are most useful for modeling.

Input Variables

1. age: Age of the candidate approached. It is a numeric variable.
2. job: Candidate's job. It is a categorical variable. Job categories are 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed'
3. marital: Candidate's marital status. It is a categorical variable. Marital status categories are 'divorced', 'married', 'single'. Note that 'divorced' means divorced or widowed
4. education: Candidates' education level. It is a categorical variable. Education level categories include 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree'
5. default: Indicates if the candidate has credit in default or not. It is a categorical variable with categories 'yes' or 'no'.
6. housing: Indicates if the candidate has a housing loan or not. It is a categorical variable with categories 'yes' or 'no'.
7. loan: Indicates if the candidate has a personal loan or not. It is a categorical variable with categories 'yes' or 'no'.

8. month: Month that the candidate was last contacted. It is a categorical variable with categories ‘jan’, ‘feb’, ‘mar’ and so on.
9. duration: Duration of call when the candidate was last contacted. This attribute affects the output significantly. It is a numeric variable.
10. campaign: Number of contacts performed this campaign for this candidate. It is a numeric variable.
11. pdays: Number of days that passed by after the client was last contacted from a previous campaign. It is a numeric variable.
12. previous: Number of contacts performed in the previous campaign for this candidate. It is a numeric variable.
13. poutcome: Outcome of the previous marketing campaign. It is a categorical variable with categories 'failure', 'nonexistent', 'success'.

Output Variable

1. y: Indicates if the candidate has subscribed to the term deposit or not. It is a categorical variable with categories ‘yes’ or ‘no’.

Method

To predict the output variable y, we need to train a classification model. There are multiple classifiers that can be used.
We intend to implement a few classifiers such as:

- Logistic Regression
- Naive Bayes
- Decision Tree
- Random Forest

Upon training them, we’ll perform analysis on which models have higher performance. We shall also examine various boosting algorithms for the model we train to improve accuracy.