Direct Marketing Campaigns of Portuguese Banking Clients

**DSC 423 Project Proposal**

**Morgan Cho, Crystal Contreras, Todd Lehky, Keerthi Surendranath Rao Mirajkar, & Roshen Samuel**

**Problem Description - Todd**

Highlights of ideas, methods and data description here.

**Methodology – Keerthi & Morgan**

1. Define Problem Statement and Import Data

a. Import Dataset:

1. Data Cleaning & Exploratory Data Analysis
   1. Randomly select a subset of 600 to 3,000 records from our dataset to analyze.
   2. Deal with null values by either removing or inserting mean or median value, depending on feature.
   3. Removing duplicate values.
   4. Checking Values for Categorical attributes & Numerical Values
      1. Exploring Categorical Values
      2. Exploring Numerical Values
      3. Correlation Matrix
   5. Encoding Categorical values to numerical codes.
   6. Dropping less meaningful columns
   7. Removing Outliers
   8. Exploratory Data Analysis
   9. Creating Dummy Variables and splitting data in Training and Testing set
2. Data Modeling & Evaluation
   1. Define Model data including dummy variable, X, Y, Training, Test Dataset
   2. Function to build and asses the model which can be used generally for all the

Classification model

* 1. Building Model
     1. Logistic Regression
     2. Decision trees
     3. Random Forest
  2. Model Evaluation
     1. ROC Curve for all models
     2. Confusion matrix

1. Exporting Final Cleaned data, Training data, testing data and Prediction generated by the model on Testing Data
2. CONCLUSION (INSIGHTS):
3. Appendix/Reference

**Data Schema and Size – Crystal**

Dataset name: bank-additional.csv

UCI Dataset URL: <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>

Brief Description: This dataset contains client information from a Portueguese banking institution. The data contains a mix of numerical and categorical variables collected to look for patterns in clients that opted ‘yes’ to making a bank term deposit from a phone marketing campaign.

Number of observations: 4,119

Number of Dependent Variables: 1

Description of Dependent Variable: A categorical, binary attribute determining whether a client subscribed to a bank deposit or not.

Number of Independent Variables: 20

Description of Independent Variables:

**Numerical:**

* Age
* Duration (last contact duration, in seconds)
  + Important note from the authors regarding this attribute: “this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.” (S. Moro)
* Campaign
  + “Number of contacts performed during this campaign and for this client” (S. Moro).
  + Last contact is included.
* pdays
  + Days passed since previous campaign’s contact.
  + A client that had not been contacted previously is represented with the value 999.
* Previous
  + Amount of times the client had been contacted before the current campaign.
* emp.var.rate
  + The employment variation rate, indicated quarterly.
* cons.price.idx
  + The monthly consumer price index
* cons.conf.idx
  + The monthly consumer confidence index
* euribor3m
  + The euribor 3 month rate, indicated daily
* nr.employed
  + The number of employees, indicated quarterly

**Categorical:**

* job
  + admin
  + blue-collar
  + Entrepreneur
  + Housemaid
  + Management
  + retired
  + self-employed
  + Services
  + Student
  + Technician
  + Unemployed
  + Unknown
* marital status
  + divorced || widowed
  + married
  + single
  + unknown
* education
  + basic.4y
  + basic.6y
  + basic.9y
  + high.school
  + illiterate
  + professional.course
  + university.degree
  + unknown
* default (has credit in default?)
  + no
  + yes
  + unknown
* housing (has housing loan?)
  + no
  + yes
  + unknown
* loan (has personal loan?)
  + no
  + yes
  + unknown
* contact (contact communication type)
  + cellular
  + telephone
* month (last contact month of year)
  + [jan, feb, mar, ..., nov, dec]
* day\_of\_week (last contact day of the week)
  + [mon, tue, wed, thu, fri]
* Poutcome (previous campaign’s outcome)
  + failure
  + nonexistent
  + success

**Plan Work Distribution Timeline – Roshen**

Wednesday, May 5th Proposal meeting

Saturday, May 8th Submit rough draft of proposal

Sunday, May 9th 5:15pm Follow up

Monday, May 10th Submit final draft of proposal

Monday, May 17th Data cleaned & explored

Monday, May 24th Analysis; Begin writing report; Begin looking for research references

Monday, May 31st Finish rough draft of report; Add each person’s research reference; Begin writing future work;

Monday, June 7th Submit all deliverables & team evaluations