Direct Marketing Campaigns of Portuguese Banking Clients

**DSC 423 Project Proposal**

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

**Problem Description - Todd**

The objective of our analysis is to use regression techniques to identify parameters that would impact the success rate of telemarketers selling bank long-term deposits. It is our intent to utilize analytic methods to clean, explore the dataset and develop a model that indicates the effect of the nominal and categorical variables on the client’s response of ‘yes’ to making a bank term deposit from phone marketing campaigns.

**Methodology – Keerthi & Morgan**

1. Analyze information about a financial institution’s direct-phone marketing campaign to explore future strategies and to determine the likelihood of client term deposits.
   1. A term deposit is a fixed-term investment that requires an initial deposit into an account at a financial institution.
2. Import Dataset from: <https://www.kaggle.com/janiobachmann/bank-marketing-dataset>
3. Data Cleaning & Exploratory Data Analysis:
   1. Randomly select a subset of 600 to 3,000 records from our dataset to analyze, containing both quantitative and qualitative features.
   2. Address null values by either removing the invalid data or replacing the null value with the mean or median value for quantitative independent variables.
      1. Null values for qualitative variables will be removed.
   3. Remove duplicate values if the redundancy is consistent across the features.
   4. Encode categorical values to numerical values through the creation of dummy variables.
4. Model Approach:
   1. Analyze the histogram distribution to determine the necessity for variable transformations
   2. Create scatterplots and correlation matrices to observe variable relationships
   3. Fit full regression model with all significant variables; analyze parameter estimates, significance, goodness-of-fit, and Adj. R2 values.
   4. Determine if multi-collinearity among the independent variables is of significant concern; compute VIF statistics.
   5. Apply various selection methods to find a subset of optimal predictors.
   6. Analyze residual plots and the normal probability plot to confirm independence, constant variance, linearity, and normality.
   7. Determine/justify if removal of outliers and influential points are necessary.
   8. Verify the strongest/most influential predictors for the response variable.
   9. Split data into training and testing sets for model generation.
5. Validation Method:
   1. Compute the PRESS values and cross-validate across n models, where n is equal to the number of independent variables. Eliminate predictors that are not significant in a majority of the generated models. Verify ASE plot and values to determine performance.
   2. Use the fitted regression model to predict the dependent variable. Using SAS to compute the predicted dependent variable, 95% confidence interval and prediction interval for our estimate
   3. Confusion matrix
   4. Research resource

**Data Schema and Size – Crystal**

Dataset name: bank-additional.csv

UCI Dataset URL: <https://www.kaggle.com/janiobachmann/bank-marketing-dataset>

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 (12 total, **11** w/o n/a)
  + admin
  + blue-collar
  + Entrepreneur
  + Housemaid
  + Management
  + retired
  + self-employed
  + Services
  + Student
  + Technician
  + Unemployed
  + Unknown
* marital status(4 total, **3** w/o n/a)
  + divorced || widowed
  + married
  + single
  + unknown
* education (8, 7)
  + basic.4y
  + basic.6y
  + basic.9y
  + high.school
  + illiterate
  + professional.course
  + university.degree
  + unknown
* default (has credit in default?) (3, 2)
  + no
  + yes
  + unknown
* housing (has housing loan?) (3, 2)
  + no
  + yes
  + unknown
* loan (has personal loan?) (3, 2)
  + no
  + yes
  + unknown
* contact (contact communication type) (2)
  + cellular
  + telephone
* month (last contact month of year) (12)
  + [jan, feb, mar, ..., nov, dec]
* day\_of\_week (last contact day of the week) (5)
  + [mon, tue, wed, thu, fri]
* Poutcome (previous campaign’s outcome) (3)
  + failure
  + nonexistent
  + success

**Project Timeline– Roshen**

| **Date** | **Activity** |
| --- | --- |
| Wednesday, May 5th | Team project proposal meeting |
| Saturday, May 8th | Initial project proposal draft completed |
| Sunday, May 9th | Project proposal rough draft team review |
| Monday, May 10th | Project proposal final draft submitted |
| Sunday, May 16th | Team meeting on data cleaning finalization |
| Monday, May 17th | Dataset cleansing and exploration completed |
| Sunday, May 23rd | Team meeting on dataset analysis done up to this point. Discussion on report. Discuss research references. |
| Monday, May 24th | Dataset analysis completed and team begins to work on rough draft of project report (including finding research references for graduate project requirements) |
| Sunday, May 30th | Team reviews draft for any edits. |
| Monday, May 31st | Rough draft of project report completed. |
| Sunday, June 6th | Finalization of all project deliverables including team evaluations |
| Monday, June 7th | Submission of all project deliverables including team evaluations |