CIS 563: Introduction To Data Science
Term Project Proposal

“ Predict whether client subscribes a long term deposit through telemarketing ”

by:
Prathma Rastogi (409745940)

Introduction

Telemarketing is an interactive technique of direct marketing via the phone which is widely used by banks to sell long term deposits. Marketing selling campaigns constitute a typical strategy to enhance business. Centralizing customer remote interactions in a contact center eases operational management of campaigns. The bank telemarketing data used here is related with direct marketing campaigns of a Portuguese bank institution.

Problem Description

The objective of this project is to predict if the client will subscribe to a long-term deposit through telemarketing. Within a campaign, the human agents execute phone calls to a list of clients to sell the deposit, or if the client calls the contact center for any reason, he is asked to buy a subscription of long term deposit. Hence, the result is binary successful or unsuccessful contact.

Further, the methodologies will be implemented to analyze the characteristics of clients who are predicted to invest in long-term deposits. The bank can utilize this information to increase revenue and provide other profits to such customers.

Since, the result is binary successful or unsuccessful contact, the problem can be termed as classification problem.

Dataset Description

The dataset is related with direct marketing campaigns of a Portugese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ‘yes’ or not (‘no’) subscribed.

There are two datasets:

- **bank-additional-full.csv** comprising 41188 examples and 20 inputs, ordered by date.
- **bank-additional.csv** comprising 10% of the examples randomly selected.

The dataset encompasses 4 main groups of features:

- Demographic information: *age, job, marital, education, default, housing, balance, loan*
- Time characteristics of the last call: *day, month, duration, contact type (‘cellular’, ‘telephone’)*
- Campaign characteristics: *contact, campaign, pdays, previous, poutcome*
- Social and economic context: *employment variation rate, consumer price index, consumer confidence index, number of employees*

Dataset Source

https://archive.ics.uci.edu/ml/datasets/bank+marketing#

Problem Approach

Leveraging CRISP-DM (Cross-industry standard process for data mining) methodology of data mining that includes,
Business understanding ❖ *Data understanding* ❖ *Data preparation* ❖ *Modeling* ❖ *Evaluation* ❖ *Deployment*

Data preparation/preprocessing:

- Data standardization and normalization
- Handling missing values and outliers
- Feature engineering
- SMOTE oversampling as this data is imbalanced

Data mining algorithms / modeling:

- Logistic Regression
- Principal Component Analysis
- K-Nearest neighbours
- Support Vector Machine (SVM)
- Ensemble Learning methods: Random Forest Classifier & Light Gradient Boosting Machine

Evaluation

- Dataset will be divided into 80% training dataset and 20% test dataset.
- K-fold cross validation will be performed on training data for tuning of hyperparameters.
- I will be using Area Under Curve, Balanced Accuracy, and f1-score as my evaluation metrics to evaluate the model.

Planned Submissions

Final submission by the end of semester will be including:

- Final report containing the results with accuracy calculated and comparison of performance among models.
- All the dataset used in the above problem description.
- Cleaned dataset used for implementation of algorithms.
- Graphs for data visualization.
- Separate codes for each algorithm implementation.