# Introduction

Banks regularly conduct direct marketing campaigns to get the potential customers to subscribe to certain products. It is therefore important to analyse the dataset to predict the effectiveness of direct marketing campaign.

# Objectives

The classification goal is to predict if the direct marketing campaign is effective to get the potential customers will subscribe to a bank product.

For this particular project the direct marketing campaign is done by phone calls and the bank product is term deposit.

# Dataset

Exploratory data analysis is done using Python, files are saved as python notebook (.ipyn).

All the analysis is done in R, files saved as R Markdown (.Rmd).

There are two datasets:

* ‘Bank-full.csv’ with all examples, ordered by date (from May 2008 to November 2010).
* ‘Bank.csv’ with 10% of the examples (4521), randomly selected from ‘Bank-full.csv’.

Due to limited computing power, smaller dataset (‘Bank.csv’) is used to explore different machine learning techniques. Full dataset (‘Bank-full.csv’) is used for more detailed prediction.

# Methodologies

There are several machine learning techniques employed for comparison using smaller dataset (‘Bank.csv’):

* K-nearest neighbours (KNN)
* Logistic Regression
* Decision tree using tree()
* Decision tree using rpart()
* Random Forest
* Gradient Boosting
* Adaptive Boosting (AdaBoost)
* Extreme Gradient Boosting (XGboost)
* Support Vector Machine (SVM) with linear kernel
* Support Vector Machine (SVM) with polynomial kernel
* Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel

Based on result comparisons, two machine learning techniques are selected to perform more detailed analysis using full dataset (‘Bank-full.csv’).

# Results Comparisons

|  |  |  |
| --- | --- | --- |
| ML Technique | Misclassification Error | Type-II Error |
| KNN | 0.12 | 0.89 |
| Logistic regression | **0.10** | **0.62** |
| Decision tree | 0.13 | 0.58 |
| Decision tree with rpart | 0.11 | 0.63 |
| Random forest | **0.10** | **0.60** |
| Gradient boosting | 0.11 | 0.65 |
| Adaboost | 0.11 | 0.64 |
| XGBoost | 0.11 | 0.64 |
| SVM with linear kernel | 0.11 | 0.81 |
| SVM with polynomial kernel | 0.12 | 0.97 |

The detail analysis is focused on techniques which yielded best results in terms of high accuracy and low Type-II error:

* Logistic Regression.
* Random Forest.

## Logistic Regression

## Random Forest

# Findings

Generally the accuracy based on TP and TN are very good regardless of the prediction models. However, it is most likely caused by the data imbalance: there are way more ‘No’ in label ‘y’.

The marketing campaign classification prediction ‘y’ is based on probability of potential customers will subscribe to bank term deposit. Current assessment ‘y’ = ‘Yes’ is based on p(variables) > 0.5.

If the bank wishes to be conservative in making the prediction, lower threshold might be chosen, for example p(variables) > 0.2

As the threshold is reduced, the error rate of ‘Yes’ prediction (potential customers that will subscribe to term deposit) also decreases. However, at the same time the ‘No’ prediction increases. In order to

The reason we are choosing AUC over accuracy is because, as we will see in Exploratory data analysis, the dataset we are working with is an imbalanced dataset with the class “no” being the majority class. If we use accuracy as our metric, any random model can give us a very good accuracy. But at the end, it will be a random model. AUC gets over this problem by looking into both the True positive rate (TPR) and False positive rate (FPR). Only if both the TPR and FPR are well above the random line in the ROC curve, we will get a good AUC. Accuracy does not guarantee that.

Target: high TPR, low FPR (Type-I error)

This is given by area under ROC curve -> AUC

A classifier that performs no better than chance to have an AUC of 0.5

# Conclusion/Discussion

# References

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

<https://towardsdatascience.com/machine-learning-case-study-a-data-driven-approach-to-predict-the-success-of-bank-telemarketing-20e37d46c31c>

<https://www.kaggle.com/janiobachmann/bank-marketing-campaign-opening-a-term-deposit>

<https://www.kaggle.com/psqrtpsqrt/bank-marketing-eda-classification-pr-f-score#model-selection>

<https://github.com/z-o-e/bank_data_analysis/blob/master/Linear_Models_Discriminants_Additive_Models_trees.R>

# Attachment 1: Dataset

The dataset is taken from UCI Machine Learning Repository:

[https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing)

The data is related with direct marketing campaigns (by phone calls) of a Portuguese banking institution from May 2008 to November 2010.

Input variables:

#bank client data:

1 - age (numeric)

2 - job: type of job (categorical: "admin", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services")

3 - marital: marital status (categorical: "married", "divorced", "single"; note: "divorced" means divorced or widowed)

4 - education (categorical: "unknown", "secondary", "primary", "tertiary")

5 - default: has credit in default? (binary: "yes", "no")

6 - balance: average yearly balance, in euros (numeric)

7 - housing: has housing loan? (binary: "yes", "no")

8 - loan: has personal loan? (binary: "yes", "no")

# related with the last contact of the current campaign:

9 - contact: contact communication type (categorical: "unknown", "telephone", "cellular")

10 - day: last contact day of the month (numeric)

11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

12 - duration: last contact duration, in seconds (numeric)

# other attributes:

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)

15 - previous: number of contacts performed before this campaign and for this client (numeric)

16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown", "other", "failure", "success")

Output variable (desired target):

17 - y - has the client subscribed a term deposit? (binary: "yes", "no")

# Attachment 2: Exploratory Data Analysis

Output variable comparison

Total output variable ‘y’ based on **job**:

Percentage of output variable ‘y’ for each **job** group:

Comment:

Out of all job categories, ‘student’ (29%) and ‘retired’ (23%) have the highest percentage of subscribing to term deposit.

Total output variable ‘y’ based on **marital status**:

Percentage of output variable ‘y’ for each **marital status** group:

Comment:

Marital status does not seem to have made any difference on the output variable: most (85% to 88%) will say ‘No’ to term deposit regardless of marital status.

Total output variable ‘y’ based on **month** the campaign is conducted:

Comment:

Highest ‘Yes’ on month March (52%), December (47%), and September (46%).

Total output variable ‘y’ based on **average balance**:

Comment:

The potential customers will more likely say ‘Yes’ to term deposit if they have more balance.

Total output variable ‘y’ based on existing **housing loan**:

Comment:

Potential customers who **does not** have existing housing loan will more likely subscribe to term deposit.

Total output variable ‘y’ based on existing **personal loan**:

Comment:

Potential customers who **does not** have existing personal loan will more likely subscribe to term deposit.

Total output variable ‘y’ based on call duration:

Comment:

As expected, the call duration is strong predictor for the potential customers to actually subscribe to term deposit.