Bank Telemarketing Decision Model

Predictive Modelling (2024 P3A)

HAN - Master Applied Data Science

29th of June 2025

Research takeover of Portuguese bank by investment company.

Point of focus is the telemarketing campaign to sell bank deposit subscriptions.

Reduce randomly chosen telemarketing campaign prospects,
by developing a predictive model,
that performs better than random selection,
to achieve profit maximization.

Exploratory Data Analysis

- classification task
- bank-additional-full.csv
- 41,188 rows
- 21 columns
- target: `y` (yes/no)
- features: 20 columns
- categorical: 10, numerical: 10
- missing values: 0

Feature engineering

- `year`: Extract `year` from data description and `month`, `day_of_week` features

“Proportions of instances per year”

Figure 1: “Proportions of instances per year”

“Proportions of target”

Figure 2: “Proportions of target”

Other Remarks

- Feature `duration` should be removed
  - Duration of last contact, not available at prediction time
- Mix of categorial and numerical values within `pdays` feature
  - `pdays` is the number of days since last contact within campaign
  - `pdays` = 999 means no previous contact
  - `pdays` = 0 means contact on the same day
  - instances with `pdays` = 999 did show previous contact, which is incorrect

Data Preparation

- Convert `y` to binary values
  - `yes` = 1, `no` = 0
- Remove `duration` feature
- Split dataset into `approached` and `not approached`
  - Filter dataset `pdays` = 999 for `not approached`
  - Rremove features `pdays`, `previous`, `poutcome` in `not approached`

Modelling

Train and Test Split

- Split based on `year` = 2010 as test set for both datasets
- ...

Feature Transformation

- Transform categorical features:
  - `job`, `marital`, `education`, `default`, `housing`, `loan`, `contact`, `month`, `day_of_week`, `year`
- Scale numerical features:
  - Approached: `age`, `campaign`,`emp.var.rate`, `cons.price.idx`, `cons.conf.idx`, `euribor3m`, `nr.employed`, `pdays`, `previous`, `poutcome`
  - Not approached: `age`, `campaign`, `emp.var.rate`, `cons.price.idx`, `cons.conf.idx`, `euribor3m`, `nr.employed`

Model Shortlist

- Random Forest
- AdaBoost
- XGBoost

Initially shortlisted due to the ability to handle imbalance in datasets.

Model Evaluation

Approached dataset

Metric	Random Forest	AdaBoost	XGBoost
Recall	0.891	0.717	0.796
ROC AUC	0.637	0.638	0.611

Not approached dataset

Metric	Random Forest	AdaBoost	XGBoost
Recall	0	0	0
ROC AUC	0.54	0.539	0.548

Model Selection

- Select **AdaBoost** as the best performing model for **appoached** dataset
- Select **XGBoost** as the best performing model for **not approached** dataset
- Due to highest ROC AUC
- ...

To be continued...