Predictive Modeling for Direct Marketing Campaigns

Problem statement:

The goal is to create a model that will help a Portuguese banking institution to determine clients who will subscribe a long term deposit before contacting them in a direct marketing campaign, using data from previous campaigns.

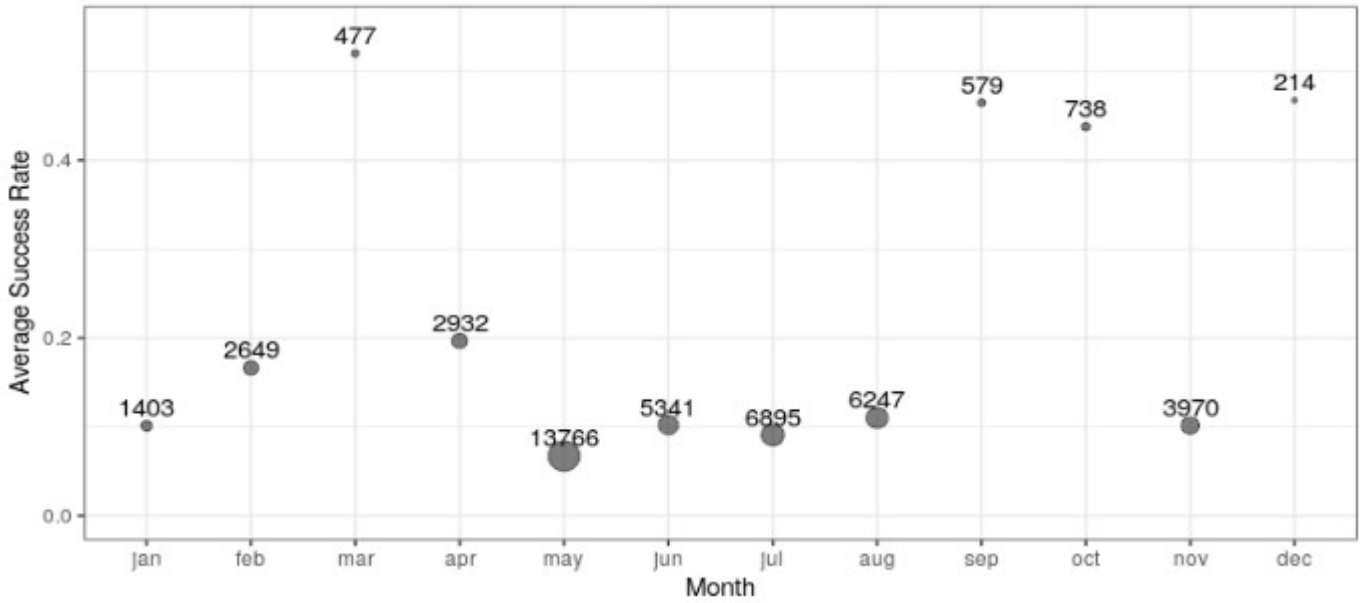
Dataset

- Source: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

- Why choose bank over bank-additional
  - The social economic indicators in bank-additional dataset are not available before predictions.
  - The bank dataset has feature balance, which is likely to be a strong predictor

Removing features

- **Duration:** not available before contact.
- **Month**: outcome strongly depends on month in the dataset but no reason to believe so; may learn wrong relationship from the data.



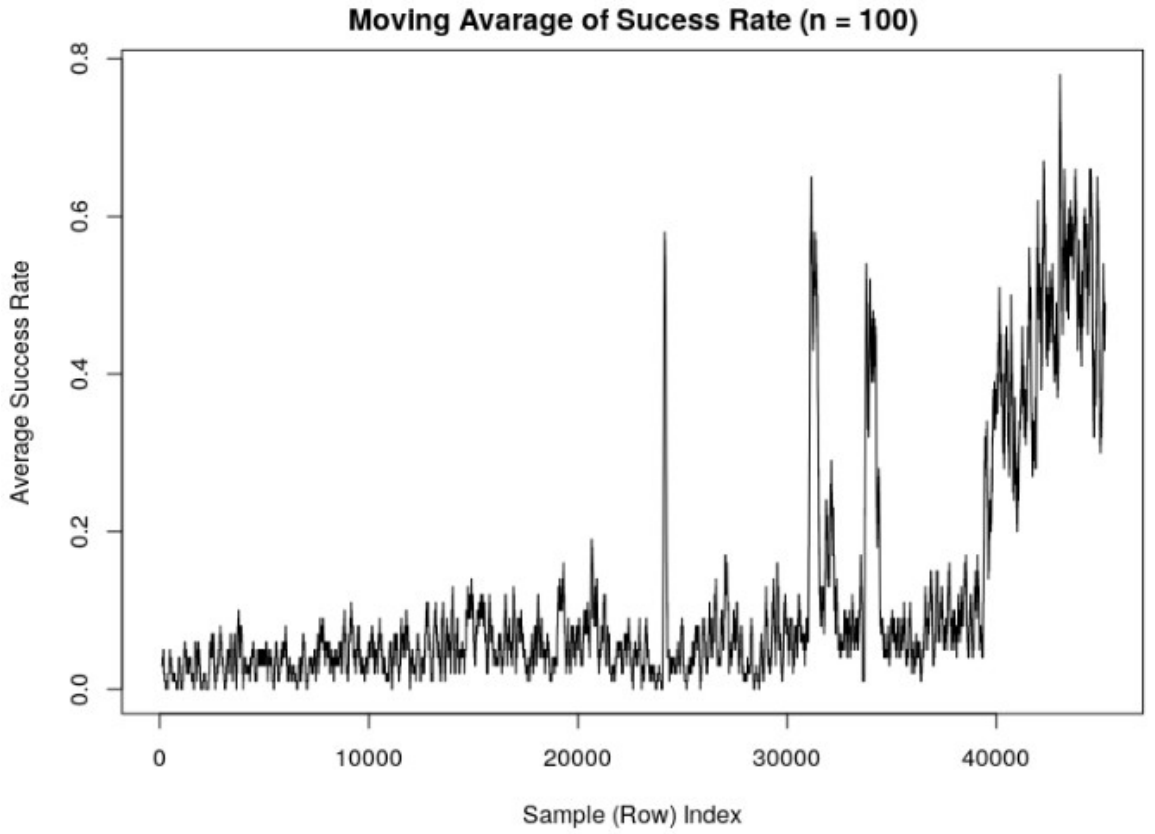
- **Day**: too many categorical values, weak dependence.

Train – test split

Ideally, use newer records as test. But the outcome is not uniform over time so use **random split instead**.

old → new

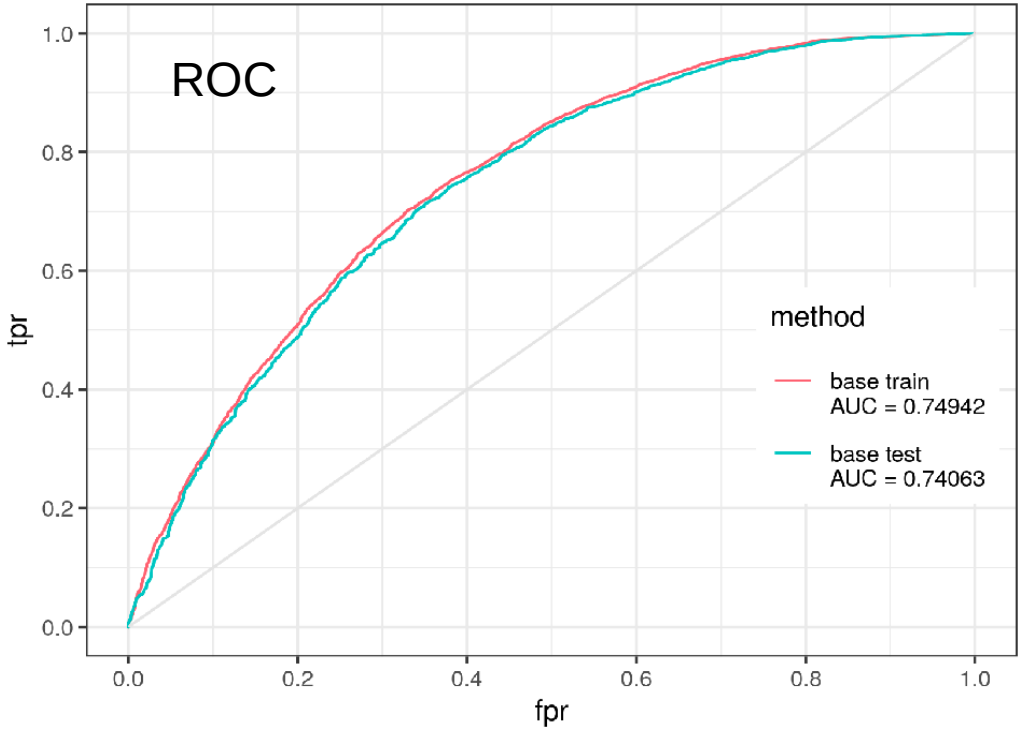
11.7%
success (“yes”)

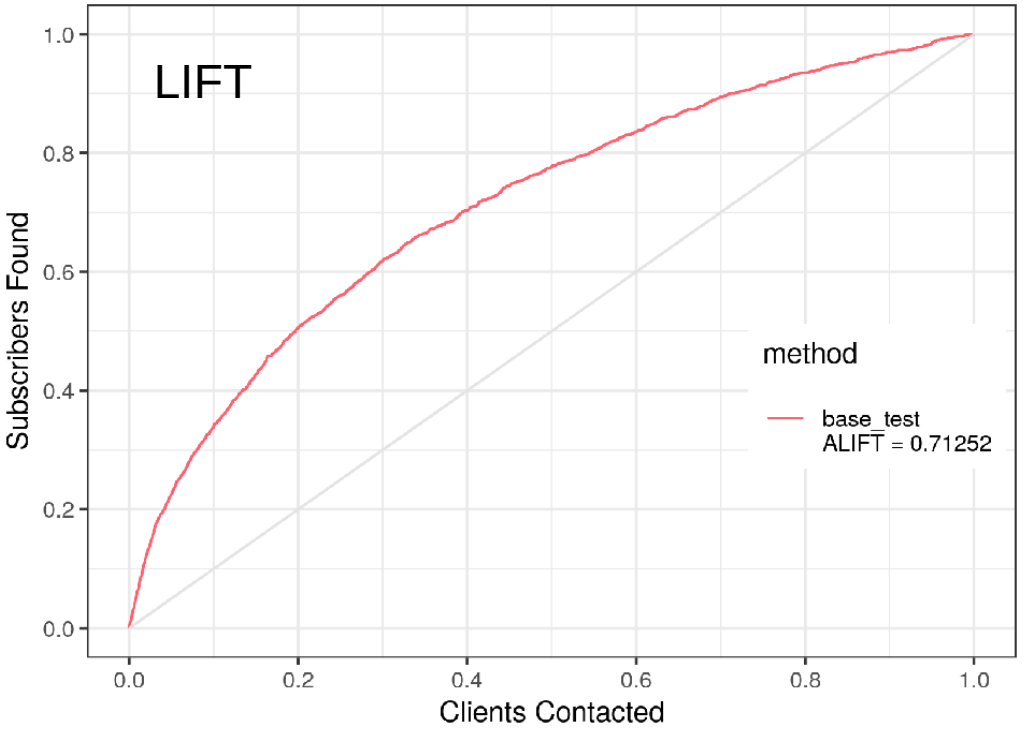


old → new

Baseline model

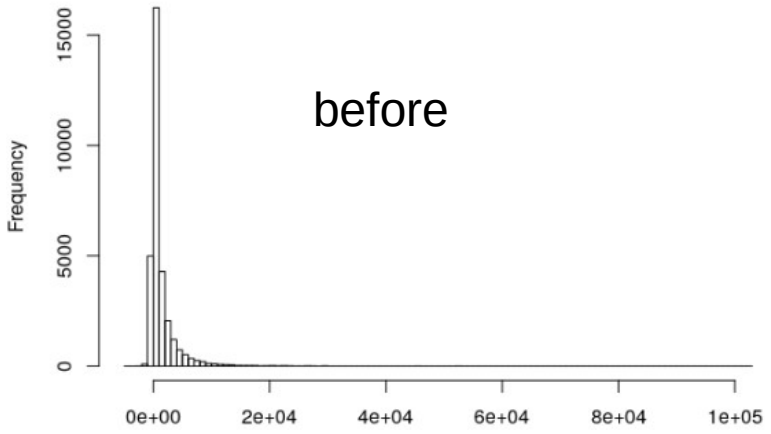
- Task: binary classification problem
- Baseline model: logistic regression
- Metrics: AUC and ALIFT
- No data preprocessing

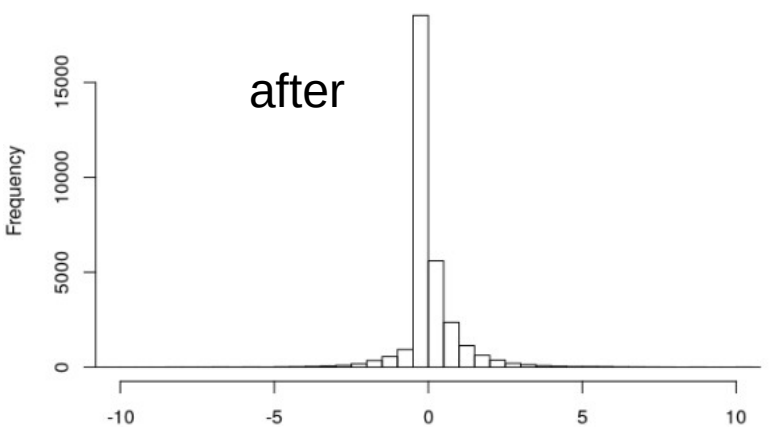




Data pre-processing

- Missing (unknown) as its own category
- Transform skewed data, for example balance



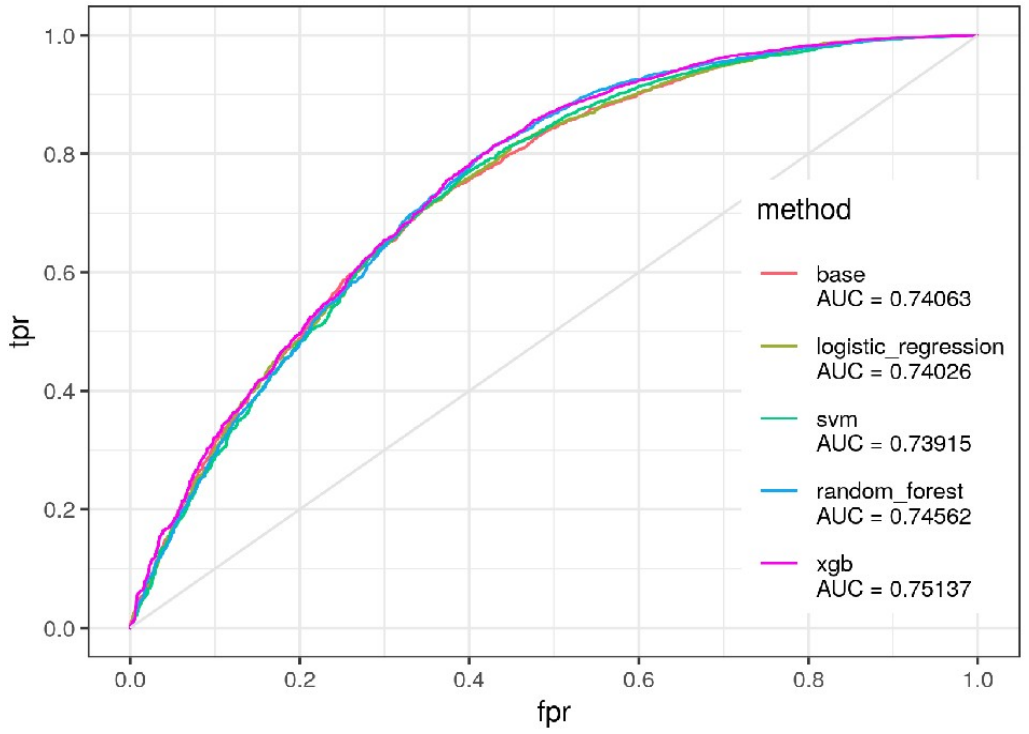


- pdays and previous
  - Create a new feature previous contact: “yes” or “no”
- Under-sampling: sufficient records, fast in training.

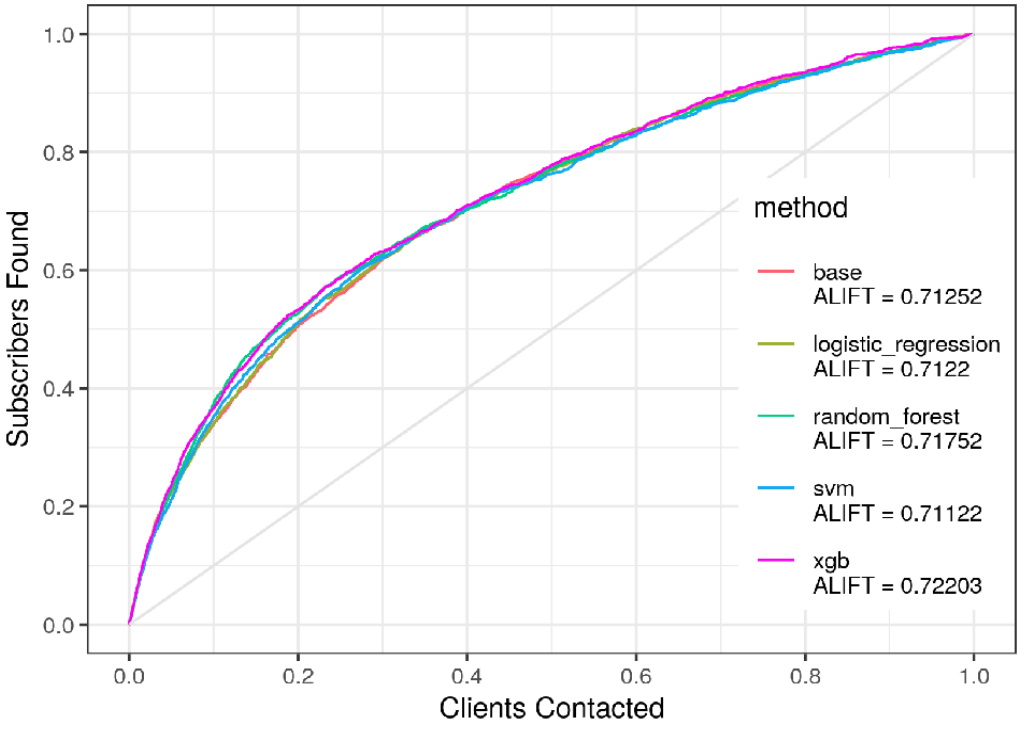
Try different algorithms

- XGBoost performed better than logistic regression, svm, and random forest, using both AUC and ALIFT

ROC



LIFT



Final model: XGBoost

Contacted --- Found

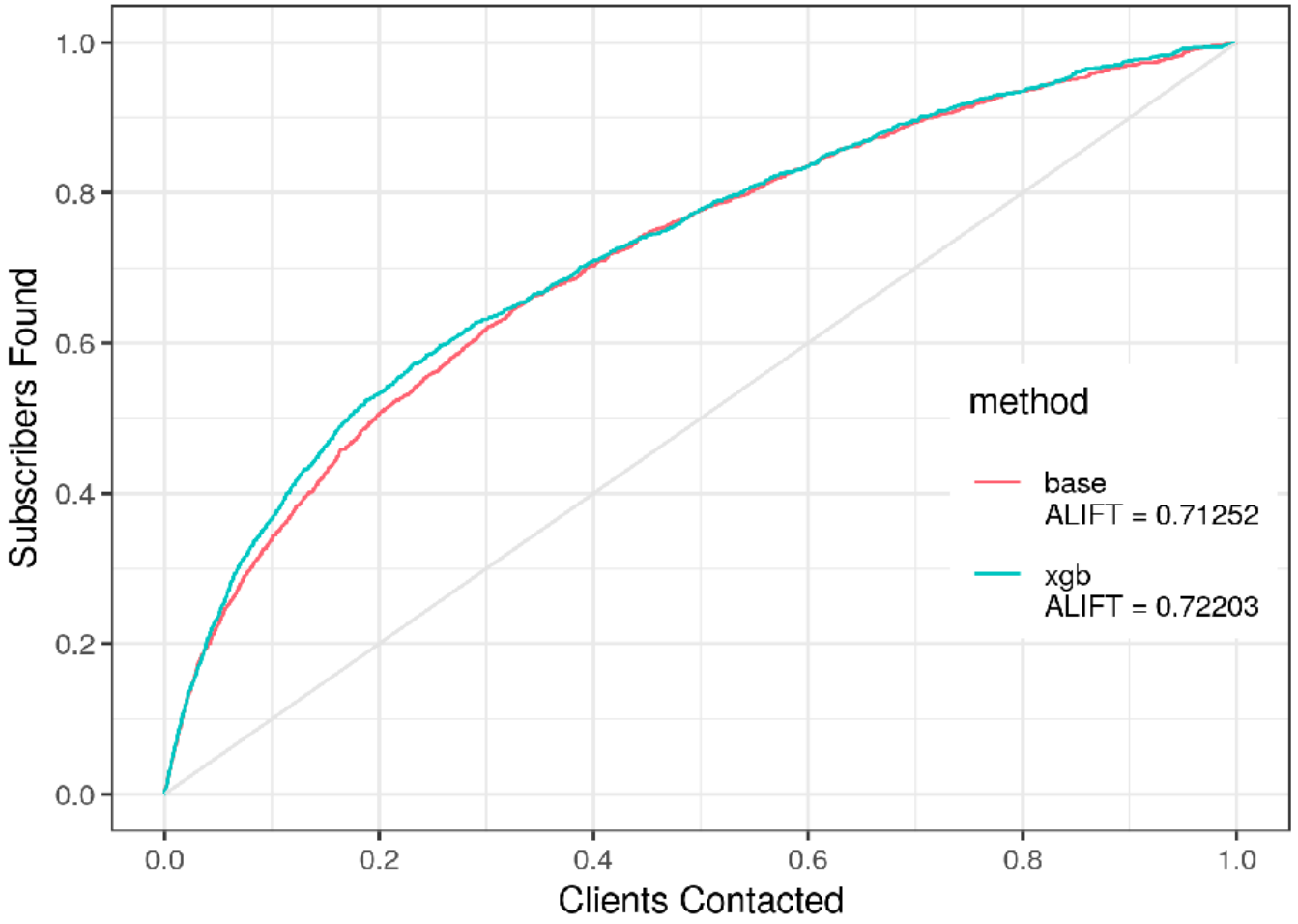
1% --- 6.4%

5% --- 23.6%

10% --- 36.6%

20% --- 53.3%

50% --- 77.8%



What else to consider:

- Profit from a subscriber
- Cost of contact
- Total number of subscribers

Further improvement

- Understand the original data, especially how campaigns differed from each other
- More time on feature engineering and hyper parameter tuning
- Understand profit from subscribers and cost of contact.