Predictive Modeling for Bank Telemarketing

By: Mohammad Zarei

CONTENTS

1. Data Overview
2. Exploratory Data Analysis (EDA)
3. Data Preprocessing
4. Model Development
5. Model Deployment
6. Model Monitoring
7. Potential Improvement Suggestions

Data Overview: From Portugal bank marketing campaigns (41176, 21)

Bank Client Info (7)

Age, job, marital, education, loan, housing, default

Current Campaign (4)

Contact, month, day of week, **duration**

Socioeconomic (4)

EVR, CPI, CCI, Euribor, # of Employees

Other (4)

Campaign, pdays, previous number of contacts, previous outcome

Target (1):

client subscription (yes:11.3%, no:88.7%)

EDA results (1)

Success rate for the calls is more for clients upto 20 and above 60 years of age.

—Age

The probability of success reduces far greatly as the number of calls increase

—# of calls

32% of students and 25% of retirees say 'yes'!

—Job

In relative terms singles was responded better

—Marital

Educated clients are more likely to subscribe.

—Education

Best communication channel is cellular

—Contact

EDA results (2)

Home ownership and having personal loan does not greatly affect performance

—Housing, Loan

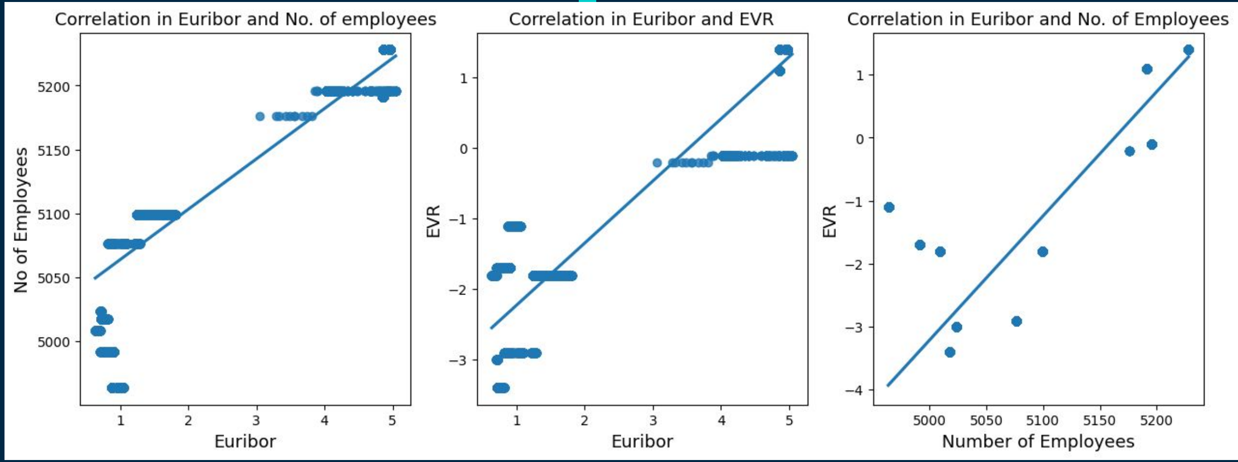
65% of the people who agreed for previous campaign agreed for this campaign as well.

—Previous Outcome

It seems it's more likely to get 'yes' during Mar, Dec, Sep, Oct.

—Month, Dayofweek

EDA results (3)



	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
age	1.000	-0.001	0.005	-0.034	0.024	-0.000	0.001	0.129	0.011	-0.018	0.030
duration	-0.001	1.000	-0.072	-0.048	0.021	-0.028	0.005	-0.008	-0.033	-0.045	0.405
campaign	0.005	-0.072	1.000	0.053	-0.079	0.151	0.128	-0.014	0.135	0.144	-0.066
pdays	-0.034	-0.048	0.053	1.000	-0.588	0.271	0.079	-0.091	0.297	0.373	-0.325
previous	0.024	0.021	-0.079	-0.588	1.000	-0.421	-0.203	-0.051	-0.455	-0.501	0.230
emp.var.rate	-0.000	-0.028	0.151	0.271	-0.421	1.000	0.775	0.196	0.972	0.907	-0.298
cons.price.idx	0.001	0.005	0.128	0.079	-0.203	0.775	1.000	0.059	0.688	0.522	-0.136
cons.conf.idx	0.129	-0.008	-0.014	-0.091	-0.051	0.196	0.059	1.000	0.278	0.101	0.055
euribor3m	0.011	-0.033	0.135	0.297	-0.455	0.972	0.688	0.278	1.000	0.945	-0.308
nr.employed	-0.018	-0.045	0.144	0.373	-0.501	0.907	0.522	0.101	0.945	1.000	-0.355
y	0.030	0.405	-0.066	-0.325	0.230	-0.298	-0.136	0.055	-0.308	-0.355	1.000

Data Preprocessing

Categorical Features

- Binary variables encoded to (1,0). Unknowns are counted as 0.
- *pdays* = 999 is replaced by zero.
- *previous* # of calls mapped to 0, 1 (>0)
- *job*, *marital*, *education* is encoded using **target encoding**
- *month*, *dayofweek* are encoded using **sine/cosine** to keep cyclical information.

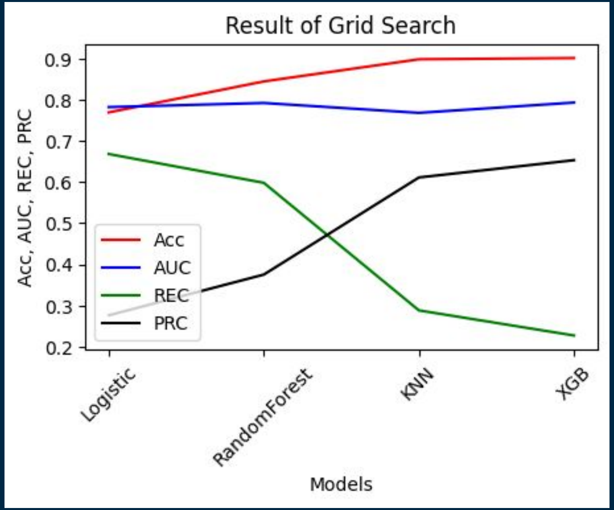
Numerical Features

- *nr.employed*, *age*, *CPI* are log-transformed.
- *CCI* signed is changed to be positive.
- *duration* dropped to avoid data leakage.

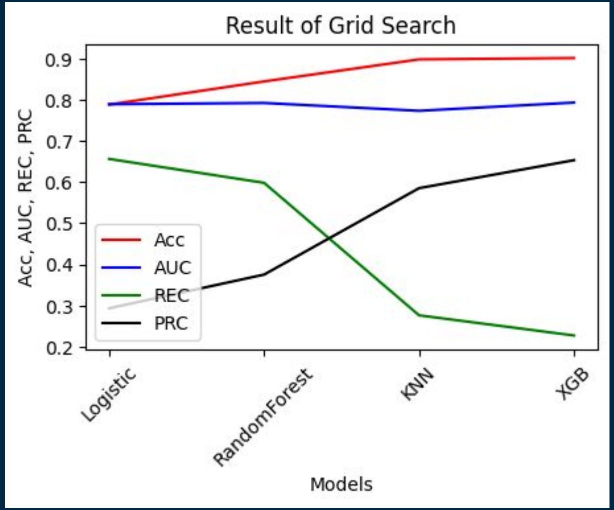
Scaling and Balancing

- X_train: without any scaling or balance
- X_train_scaled: scaled using StandardScaler
- X_train_balanced: balanced with SMOTE

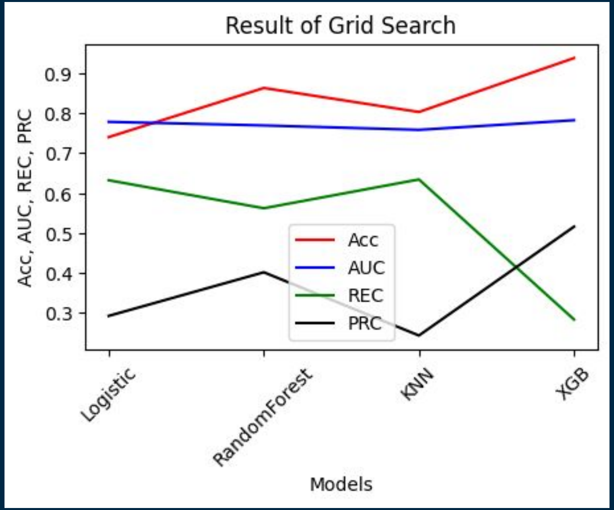
Model Development (1)



X_train



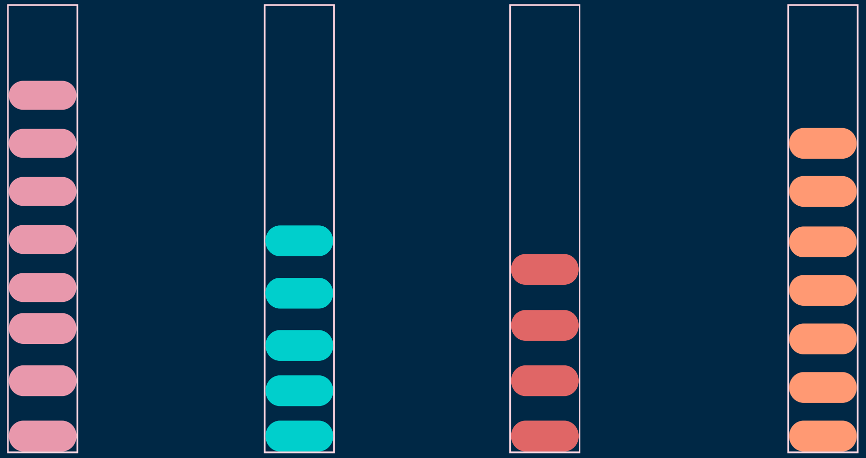
X_train_scaled



X_train_balanced

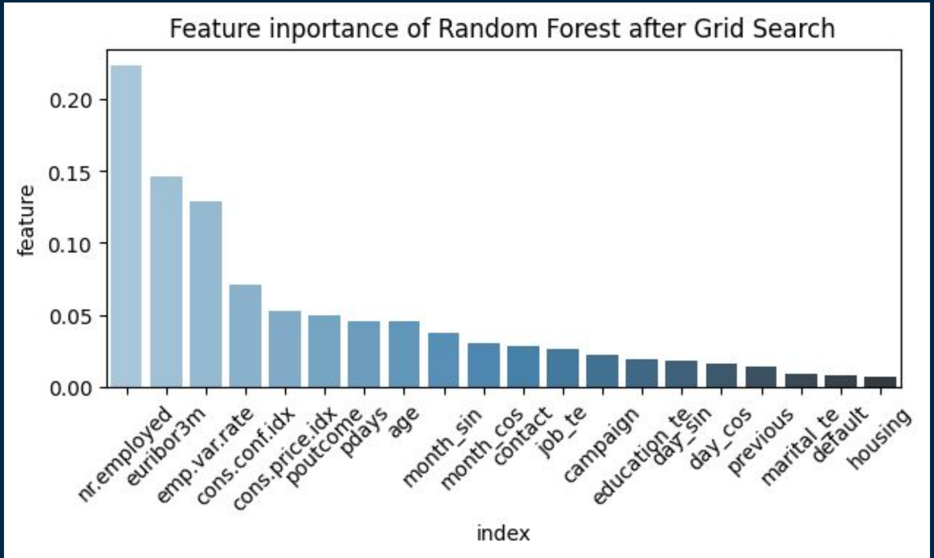
Model Development (2)

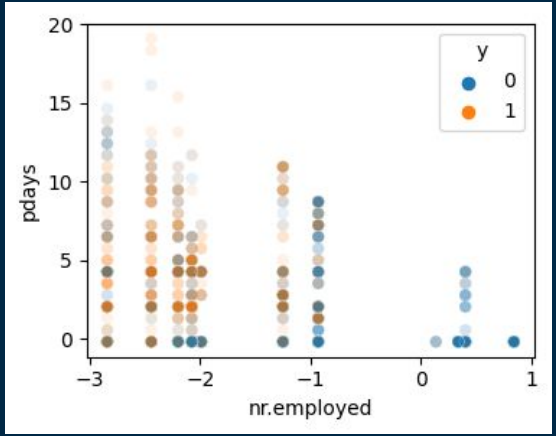
Selected RF Model

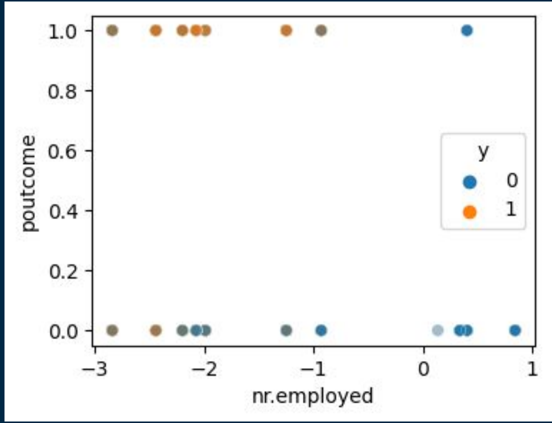


86%	56%	40%	77%
Accuracy	Recall	Precision	AUC

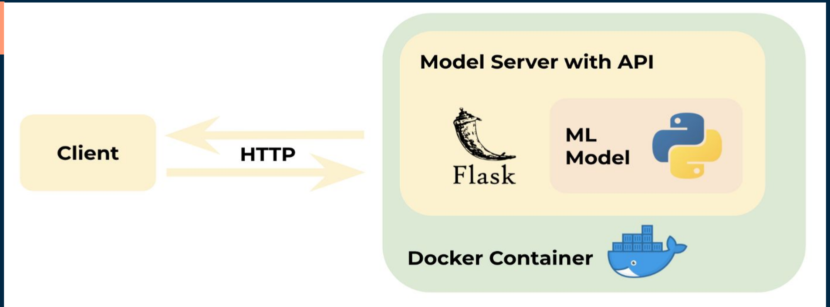
Model Development (3)







Model Deployment

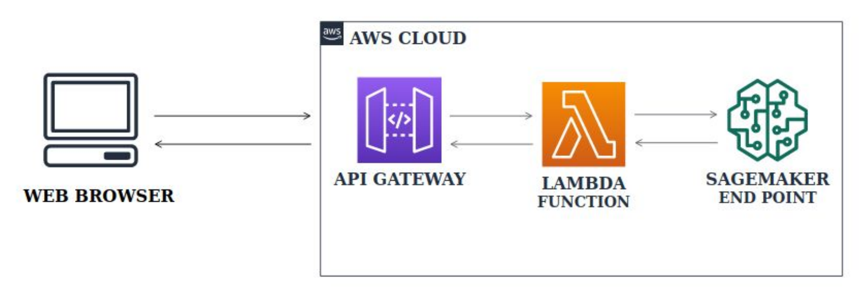


Locally

Using Docker/Flask and a local machine as server

Cloud (AWS)

Register model image, deploy in an endpoint



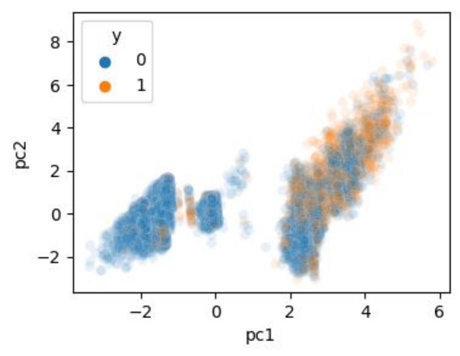
Model Monitoring

- **Data/Feature Drift:** perform distribution tests by measuring distribution changes using distance metrics - mean/std/min/max/correlation comparison, KL/KS for continuous, Chi2/entropy for category variables, PCA - Retrain new model using new data or give higher weight to new data and compare new models using A/B testing

- **Model/Concept Drift:** learned relationship/patterns have changed over time - Instantaneous (data issue, new domain), Gradual (preference change), Recurring (seasonal), Temporary (adversarial attack, unintended use) - monitor prediction metrics, label distribution

Potential Improvement Suggestions

- Binning numerical features like age, EVR, etc which different ranges seems to have different impact
- Perform feature selection (drop collinear features such EVR, nr_employees)
- Explore larger hyper parameter space (random grid search)
- Data augmentation techniques (GANs, VAE)
- Use complex models such as NN, kernel-SVM



- Other Data sources
  - Location/geospatial data (states, cities, postal code)
  - Account related data (type, number of holders, credit score, balance, product_num)
  - Other data (gender, estimated salary)