Descriptive and Predictive analytics

Group 1: JOHN Nixia Sancy RAMANNA Nintesh ZHAKSYBEK DILDA

December 2021

Agenda

- Problem statement
- Data Preparation
- Data Timeline
- Models
- Lift, Cumulative Gain and Response
- Profiling
- ROI

Problem statement

- Communications agency based in Belgium.
- Fundraising for 22 humanitarian organizations.

- Re-activation campaign by sending letters to those donors that have been **inactive for a long time.**
- Randomly selecting donors yields response rate of **1.8%.**
- Increasing response rate by **only targeting donors predicted by a model.**

Data Preparation

Data Cleaning

- Campaigns table
- Donors table
- Gifts table

Subset for Train and Test

- Gifts Train: Aug 4, 2015 to Aug 4, 2018
- Gifts Test: May 18, 2016 to May 18, 2019
- Gifts New: Jan 1, 2017 to Jan 1, 2020.

Creation of Target Variables

- Train: campaign ID = 6169, date >= Sept 4, 2018, only donors that donated
- Test: campaign ID = 7244, date >= June 18, 2019, only donors that donated

Basetable

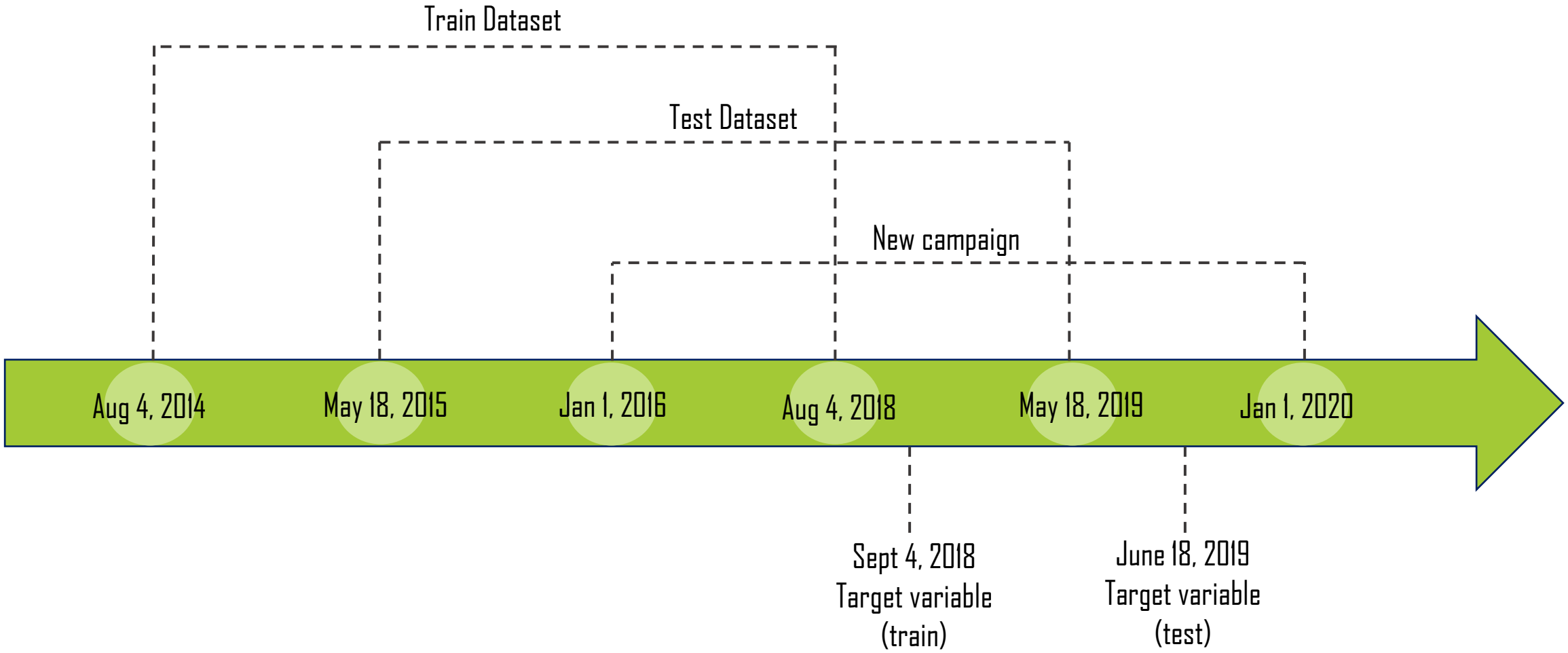
- Train dataset: gifts train + donors + target train
- Test dataset: gifts test + donors + target test
- New set: gifts new + donors + donor_scoring (campaign selection).

Feature Selection

- Pearson Correlation: 'gender', 'MISSING', 'Region__Missing', 'month', 'sum_3', 'No_campaigns', 'Total_Amount', 'Average_Amount', 'Maximum_Amount', 'Minimum_Amount', 'recency', 'Tor', 'age', 'age_group']

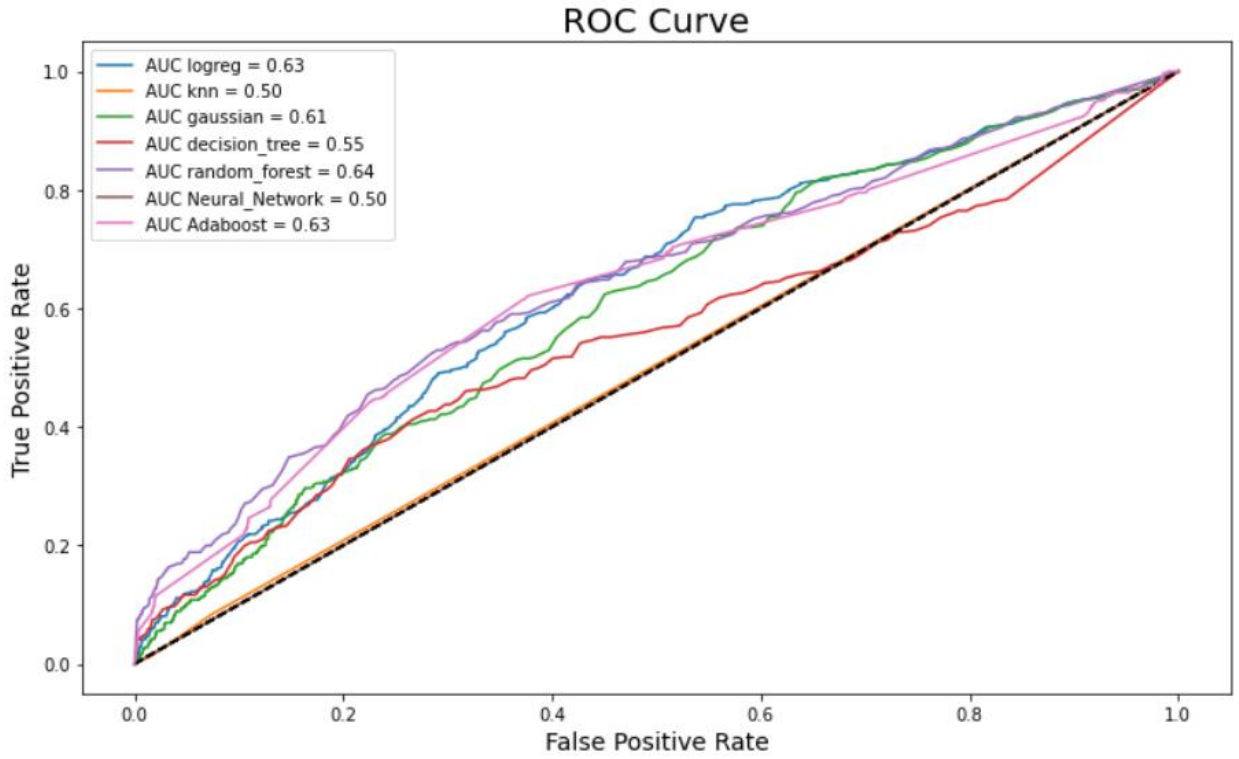
Apply models

Data timeline

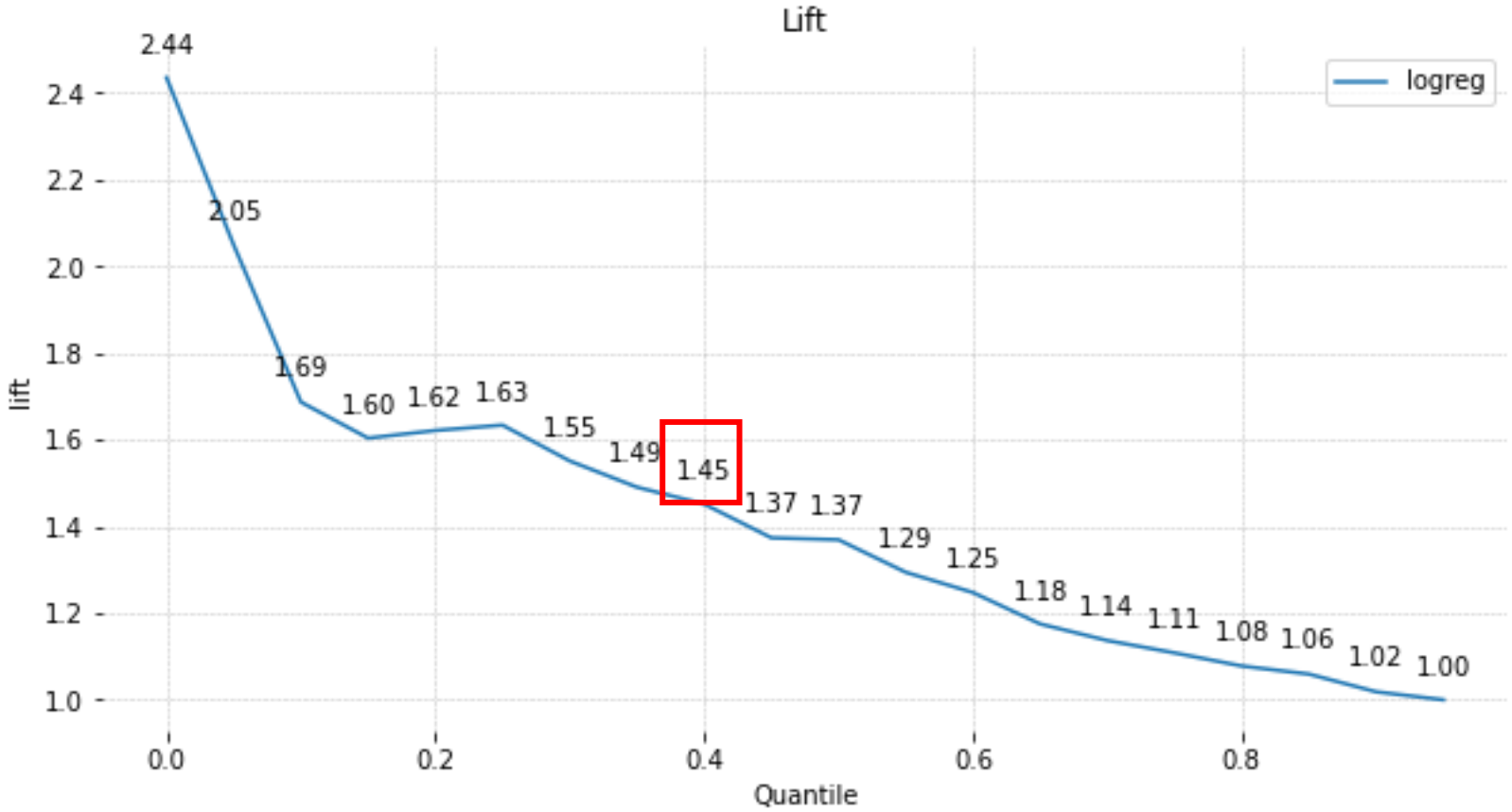


Models

Logistic Regression	KNN	Gaussian	Decision Tree	Random Forests	Neural Networks	Adaboost
0.63	0.50	0.61	0.55	0.64	0.50	0.63

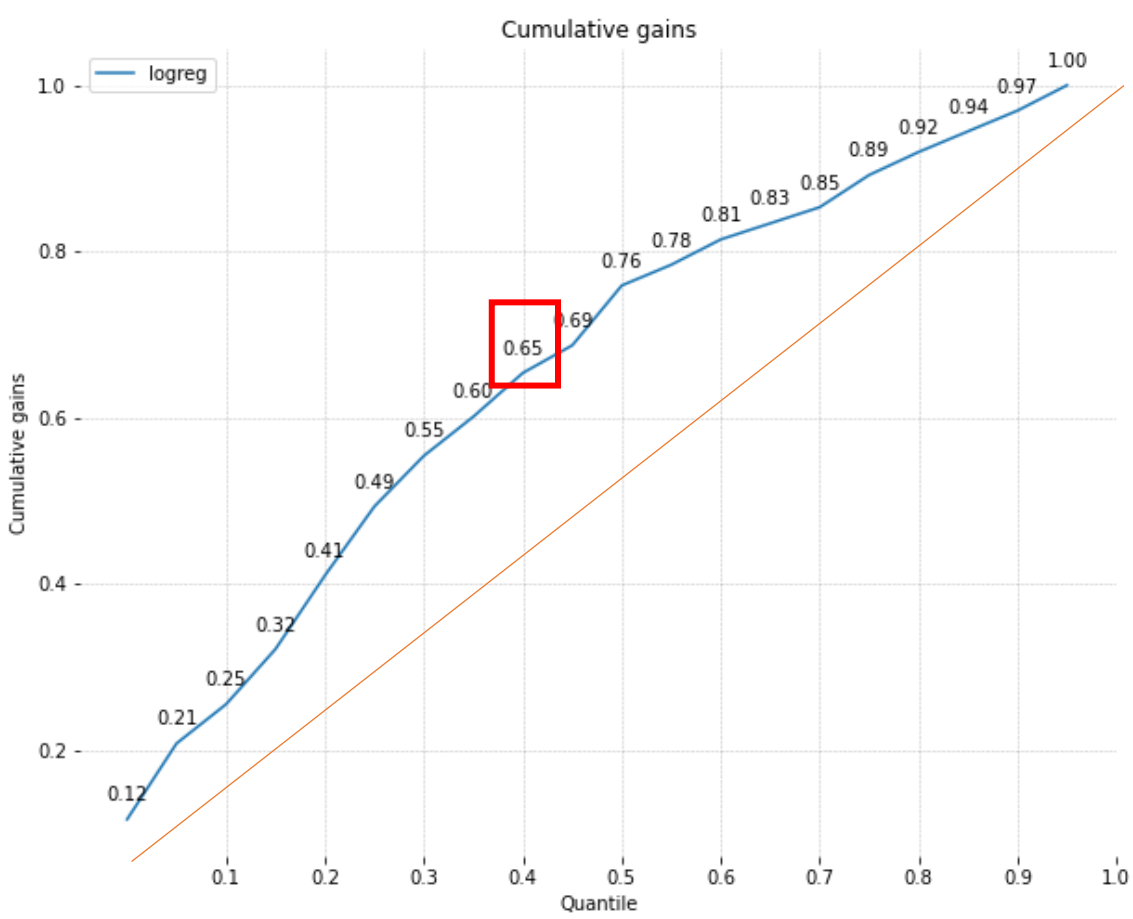


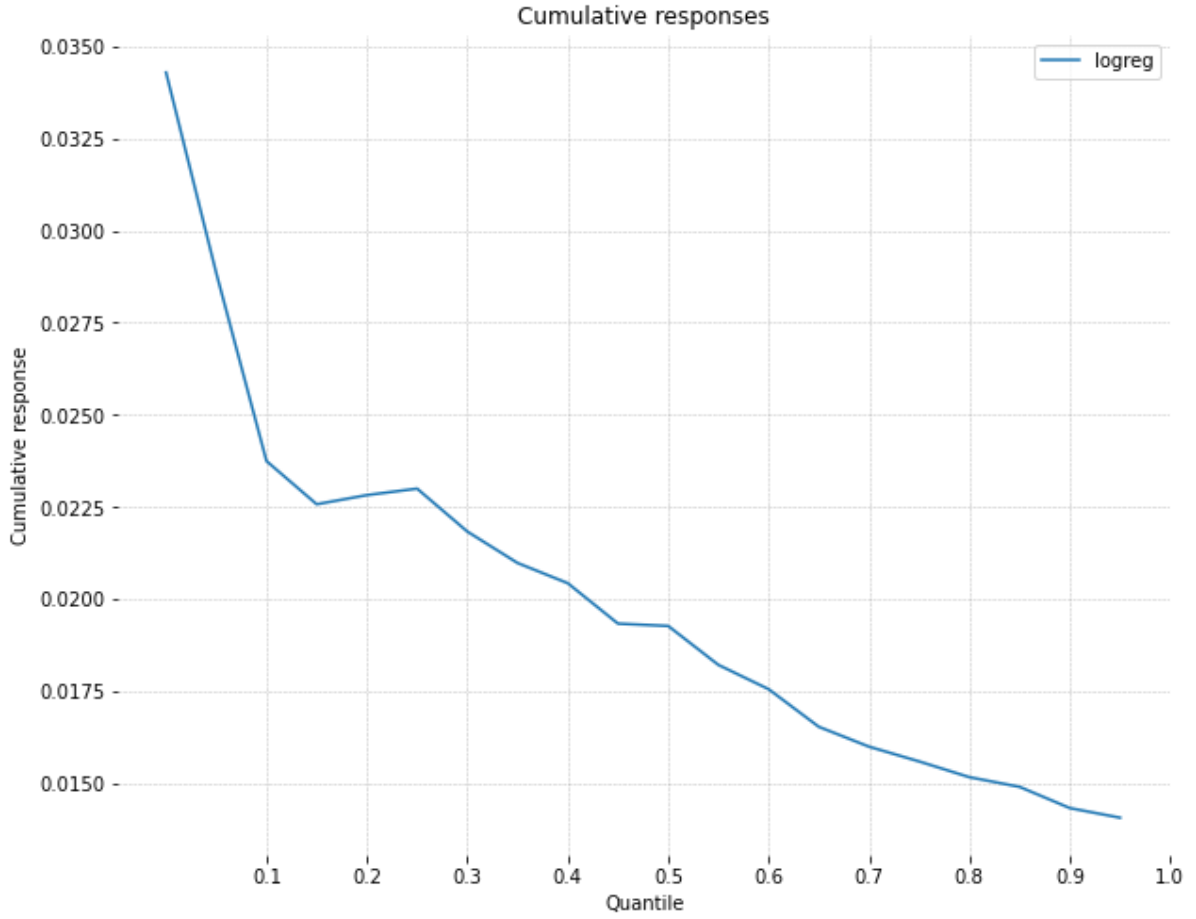
We're selecting logistic regression as our model as it has the highest AUC score and the best true positive rate. Also, type 2 error (false positive) is preferred to false negatives, which are higher in Random Forests model.



Lifts

- If 40% of the population contacted, logistic regressions model performs 1.45 times better than random selection. That is 1.45 times more potentials donors will be addressed.

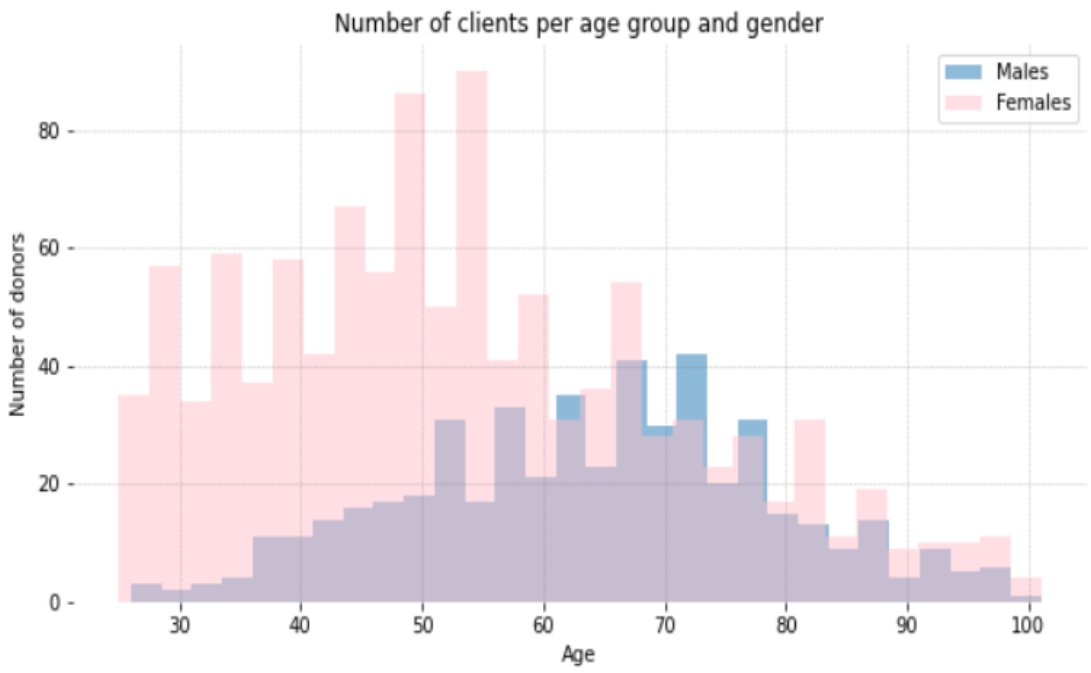


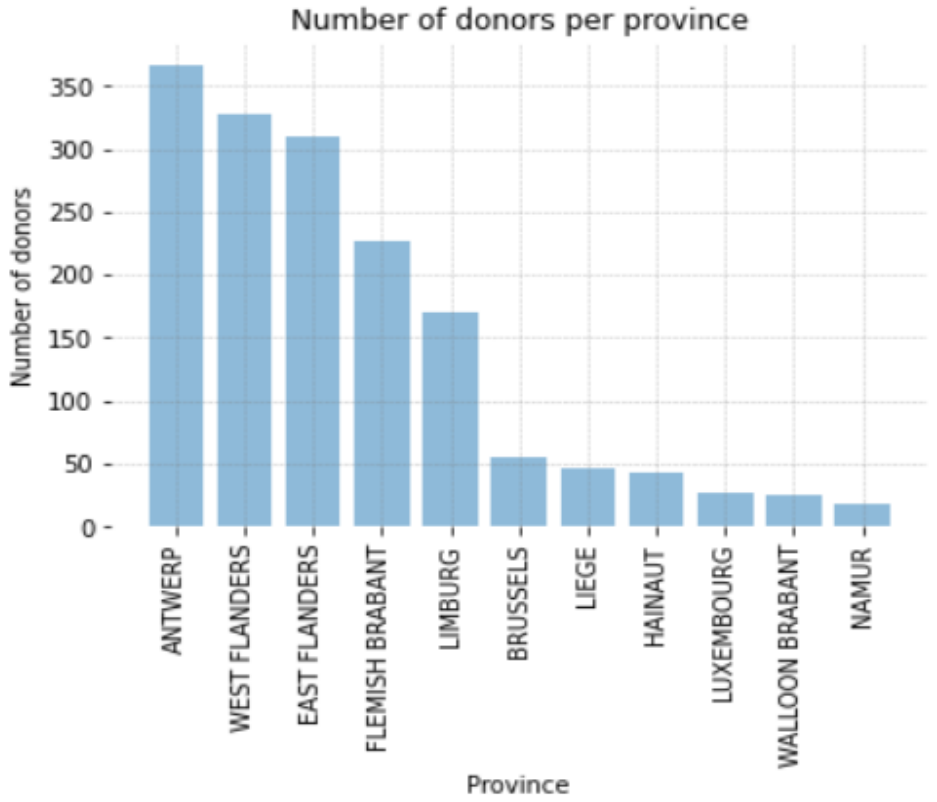


Cumulative Responses and Gains

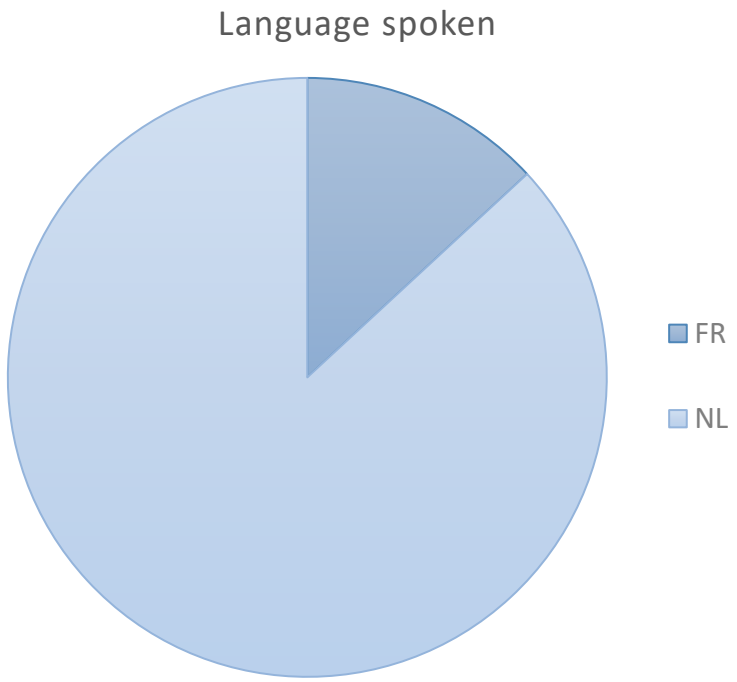
- For cumulative gains, applying the model, contacting 40% of population will lead to selecting 65% of people that are likely to donate, comparing with the base – random selection.
- For cumulative responses, applying the model, contacting 40% of population will cumulatively yield 2% donations

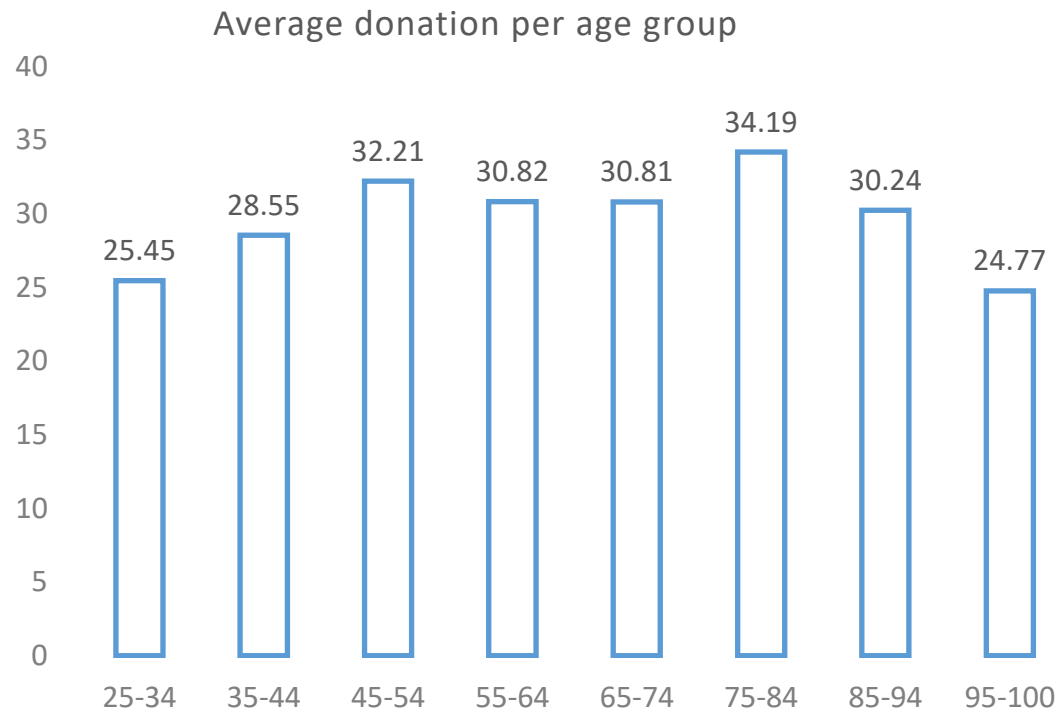
Profiling


Number of clients per age group and gender
Males
Females
Number of donors
Age
0
20
40
60
80
30
40
50
60
70
80
90
100


Number of donors per province
Number of donors
Province
0
50
100
150
200
250
300
350
ANTWERP
WEST FLANDERS
EAST FLANDERS
FLEMISH BRABANT
LIMBURG
BRUSSELS
LIEGE
HAINAUT
LUXEMBOURG
WALLOON BRABANT
NAMUR

Profiling


Language spoken
FR
NL


Average donation per age group
40
35
30
25
20
15
10
5
0
25.45
28.55
32.21
30.82
30.81
34.19
30.24
24.77
25-34
35-44
45-54
55-64
65-74
75-84
85-94
95-100

ROI

- Normal response rate calculated from the test dataset: 1.4%
- Average donation: EUR 51.57
- Selected fraction of people to be contacted: top 40% of people with the highest probability of donation

				MODEL			RANDOM			
Fraction of people contacted	Cumulative response	Cumulative gain	Amount of people contacted	Cost of offer	Gained money	Profit	Cost of offer	Gained money	Profit	Profit (model over random selection)
0.2	0.02	0.5	5,129	4,103	6,037	1,934	4,103	3,703	-400	2,334.1
0.4	0.02	0.7	10,258	8,206	10,810	2,604	8,206	7,406	-801	3,404.5
0.6	0.02	0.8	15,387	12,310	13,936	1,626	12,310	11,109	-1,201	2,827.2
0.8	0.02	0.9	20,516	16,413	16,056	-357	16,413	14,812	-1,601	1,244.2
1.0	0.01	1.0	25,645	20,516	18,616	-1,900	20,516	18,514	-2,002	101.8

Thank you!