Advanced Analytics in Fundraising

IESEG School of Management - MBDA 2020
10 December, 2020

ROHIT BHALERAO

ALESSIO ROSATO

POOJA SHAH

ROXANE SMADJA

OUR TEAM

TABLE OF CONTENTS

PROJECT DEFINITION

01

Client: Direct Social Communication - A company that raises funds for charities

**Action:** Send multiple mails to people and gather funds in a return mail

Campaign: Physical mails sent to prospective donors

Recurrence: Up to 3 times a month

Project Goals:

- Reduce costs of ineffective mailings by optimizing the target and increase the donations

- Target the donors for campaigns that are more likely to donate based on their profiles

- Create a model that overperforms the 1,8% current success rate of campaigns

TABLES

CAMPAIGNS

FILTER
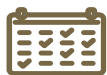
- Gifts
Data of donations made from year 2000 for different campaigns

- Campaigns
Data from year 2004 for different donation campaigns

- Campaign 6169 (training set) took place on 04/09/2018

- Campaign 7244 (test set) took place on 18/06/2019

- Each dataset dates filtered to get useful information from the past for each donors

- For training data, only the data before campaign 6169 is used while for 7662 campaign, entire data is used in our models

DATA PREPARATION

02

DONORS TABLE

Missing Zipcodes and Region: Only 6 donors in total had the language 'EN' including both these missing entries. The missing values were replaced by information from remaining 4 profiles - Region: Flanders, Province: Flemish Brabant and Zip: 3080.

Age: Calculated the donor's age from the DOB.

Dummy encoded variables:

- Gender – IsMale, IsFemale, IsMissingGender

- Region – IsFromFlanders, IsFromBrussels, IsFromWallonia, IsMissingRegion

- Language – Speaks_NL, Speaks_FR

- Province – 11 new columns, one for each province and a column IsMissingProvince

DONORS TABLE

	donorID	NewZip	IsMissingZipCode	age	IsMale	IsFemale	IsMissingGender	province	IsMissingProvince	IsFromFlanders	IsFromBrussels	IsFromWallonia	IsMissingRegion	Speaks_NL	Speaks_FR
0	100001	1861	0	52	0	1.0	0	Flemish Brabant	0	0	0.0	0.0	0	1.0	0.0
1	100002	2260	0	67	1	0.0	0	Antwerp	0	0	0.0	0.0	0	1.0	0.0
2	100003	1780	0	33	1	0.0	0	Flemish Brabant	0	0	0.0	0.0	0	1.0	0.0
3	100004	1020	0	67	0	1.0	0	Brussels	0	1	1.0	0.0	0	0.0	1.0
4	100005	1020	0	33	0	1.0	0	Brussels	0	1	1.0	0.0	0	0.0	1.0

CAMPAIGNS TABLE

campaignCost: Calculated total cost of campaign as the product of unit cost and the number of letters sent

	campaignID	date	lettersSent	CostUnit	campaignCost
0	153	13/12/2004	6873	0.17	1168.41
1	154	01/01/2005	7656	0.30	2296.80
2	180	18/01/2005	9933	0.20	1986.60
3	433	29/01/2005	7448	0.59	4394.32
4	476	12/02/2005	6605	0.51	3368.55

GIFTS TABLE

- **Missing Campaign IDs:** 2278 donations were between 2000 – 2004 for which campaigns data is unavailable and 13359 donations are not associated with any campaign. These missing fields assigned value 0 under new variable IsMissingCampaignID, keeping original intact

- **Outliers:** Removed donation amounts with outliers using z-score / IQR method

- **Date:** Converted to the right format

- **Aggregation** to find the total number of donations and total amount donated by each donor in the past and in the last 5 years, 3 years and 1 year

GIFTS TABLE

|	donorID	Number_of_donations	Total_amount	date	IsMissingCampaignID	campaignIDNew	in_last1_years	Total_amount1	in_last3_years	Total_amount3	in_last5_years	Total_amount5
0	100001	11	178.49	2004-03-23	1	0	3	11.0	1	20.0	1	20.0
1	100002	1	20.00	2007-10-11	0	1577	1	5.0	3	11.0	3	11.0
2	100003	5	18.00	2007-12-28	0	1634	1	40.0	1	5.0	1	5.0
3	100004	25	716.22	2008-06-14	0	1778	2	15.0	1	5.0	1	5.0
4	100005	4	14.96	2008-07-09	0	1808	1	20.0	2	70.0	4	130.0
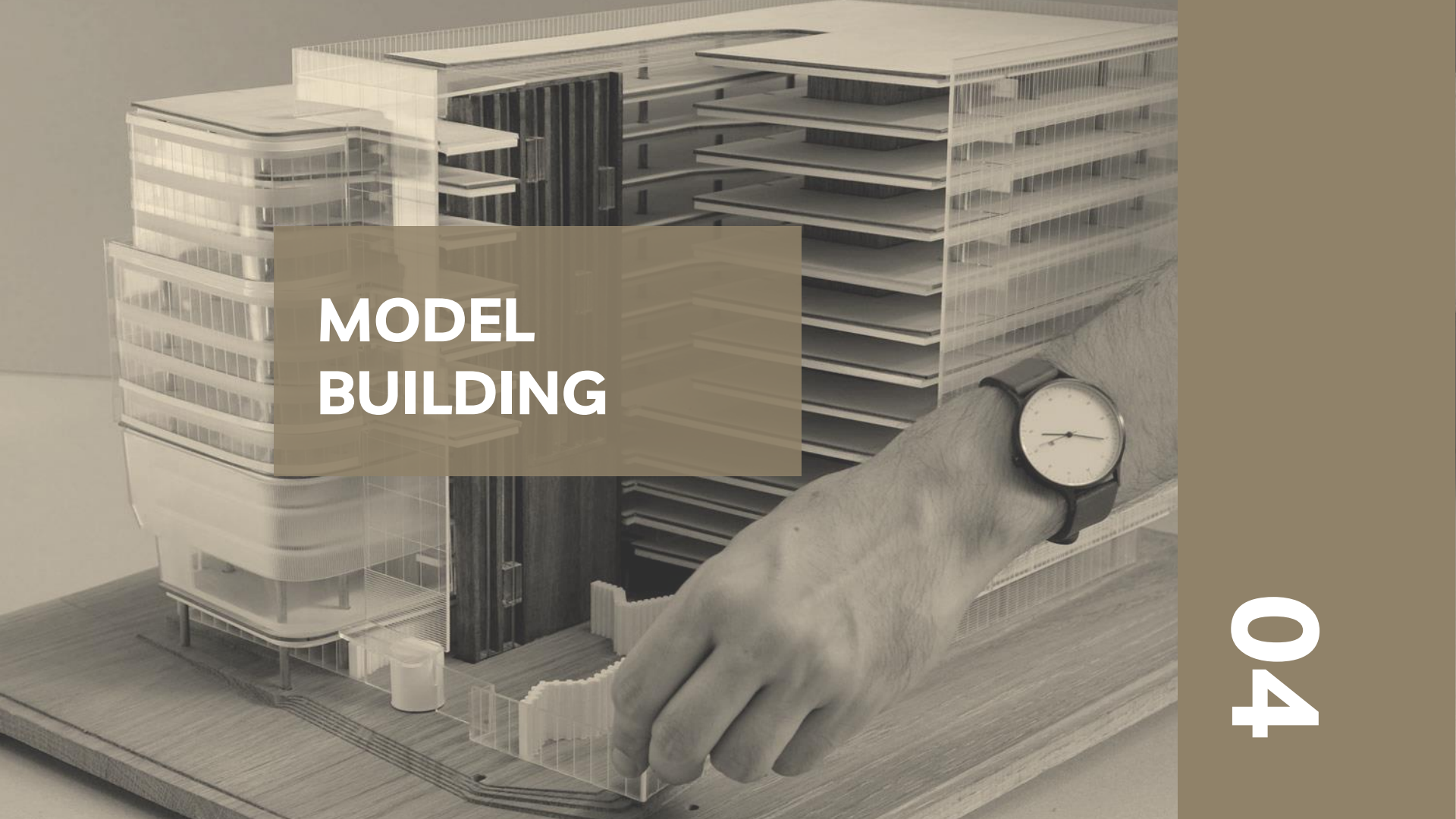
FEATURE SELECTION

03

- For selecting the most impactful features for our model, we assessed the AUC score of each of the features on the training and test set using different models and selected one with good AUC scores and also flexible – **Logistic Regression**

- We used **single-model univariate feature selection** process to get AUC score for each independent variable under logistic regression and kept only those features that scored above **50% probability** and used them for training our models: **from 29 to 20** features

- To verify the consistency of the results from the selected features, we also did **multi-model univariate feature selection** process. As the outcome was positive, we continued with **20 features** shortlisted by the previous model including age regions, last donation period, language and gender

FEATURE SELECTION
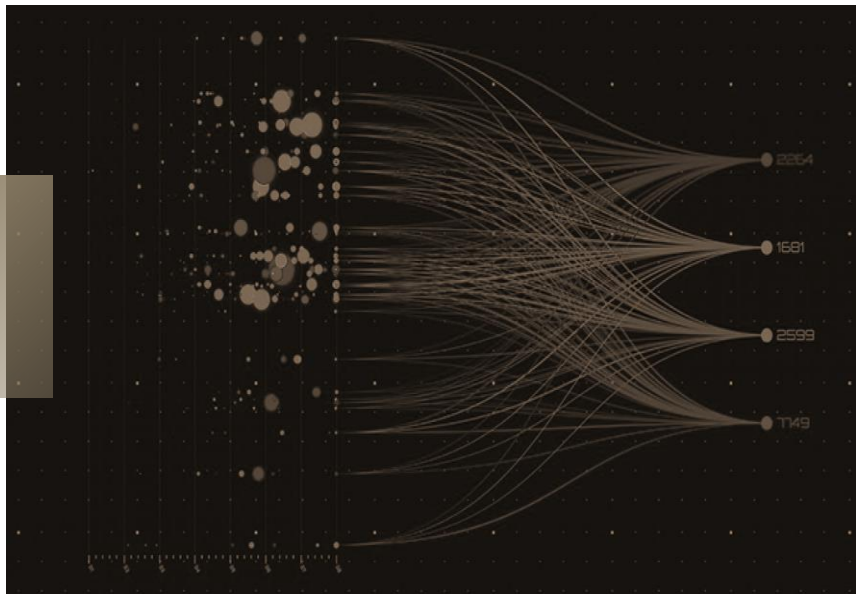
MODEL
BUILDING

MODELS USED

- 4 different training and testing datasets to compare the performance of different models

- ML models used:

- ✓ Decision Tree Classifier
- ✓ Logistic Regression Classifier
- ✓ Random Forest Classifier

- ✓ Gradient Boosting Classifier
- ✓ SVM
- ✓ Neural Networks
- ✓ K-Nearest Neighbors (KNN)

SET 1

- **Equal** number of donors (1191) and non-donors (1191) from the campaign 6169 training data randomly selected
- This ensures that the model is not biased towards predicting 0's as the original training data has 34697 non-donors (0's) and only 1191 donors (1's).
- **75% data used for training** and 25% for testing

SET 2

- **Equal** number of donors (1191) and non-donors (1191) from the campaign 6169 training data randomly selected
- We use **the entire data to train** the models and test the model on the campaign 7244 data.

SET 3

- **Entire training data** of campaign 6169 with ~ 35K rows and **entire testing** data of campaign 7244 with ~ 25K rows

SET 4

- **Entire training** data of campaign 6169 campaign
- **75% to train** and 25% to test the data
- Then we use the best model on the test data (7244 campaign)

MODELS RUN ON 4 SETS

MODEL VALIDATION

05

Set 1 Result:

Success Rate : 2.88 %
Population targeted: 21.49 %
Cost saved: 4 429.50 €

- Best model : Boosted Trees
- Accuracy: 77-80 %
- On the test set, model predicts that 21.49% of the people will donate of while 2.88% donate.

Set 2 Result:

Success Rate : 3.22 %
Population targeted: 19.35 %
Cost saved: 4 550. 20 €

- Best model : Boosted Trees
- Accuracy: 76-79 %
- On the test set, model predicts that 19.35% of the people will donate of which 3.22% donate.

DONORS = # NON-DONORS

Set 3 Result:

Success Rate : 55.55 %
Population targeted: 0.00035 %
Cost saved: NA as the targeted
rate is impossible to reach

- Best model : Logistic Regression
- Accuracy: 98.6%
- On the test set, the model predicts that 0.00035% (very specific) of the people of will donate, of which 55.55% donate

Set 4 Result:

Success Rate : 2.88 %
Population targeted: 21.48 %
Cost saved: 4430.90 €

- Best model : Boosted Trees
- Accuracy: 80%
- On the test set, the models predict that 21.48% of the people will donate of which, 2.88% donate

3.22%

CAMPAIGN SUCCESS RATE

-80.65%

COST OF CAMPAIGN

55.55%
Success Rate

IF WE HAVE A LARGE DONOR
DATA ~10 million ID's

PERFORMANCE OF
BEST MODEL

MODEL USAGE

90

Campaign 7662

Dec 15, 2020

Number of donors in the mailing list :
26522 donor IDs

Number of donors @ 1.8% success rate : **477**

Average cost of unit in the last 3 years : 0.22 €

Cost of campaign : **5834.84 €**

Number of donors targeted :
7713 donor IDs

Predicted number of donors : **250**

Success Rate : 3.22 %

Population Targeted : 29.08%

New campaign cost : **1696.86 €**

Cost saved : **4137.98 €**

THE PROFILE OF THE IDEAL DONOR

Female

72 years old

Speaks Flemish

From Region Flanders

Thank you!