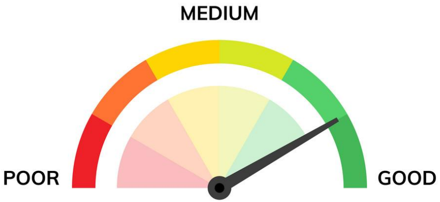
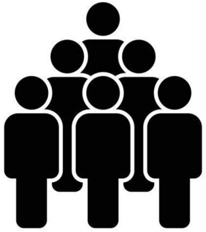
2nd Order Solutions - Applying EBM and GBM in Predicting Donations upon Receiving Mail Offers

Tego Chang, Ying Feng,
Weiliang Hu, Abhijith Tammanagari

Background

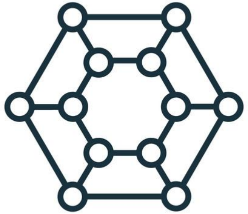


Low FICO Score



Individual

Financial Institution



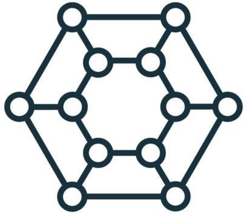
Algorithms

?

Delinquent Record

Insufficient Income

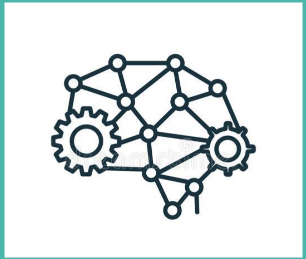
Main Project Objective - Model Comparison



GBM (Gradient Boosting Machine)

- Mainstream Algorithm
- Pros: Fast, high accuracy
- Cons: Black-box, hard to explain

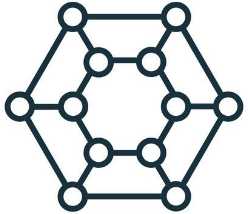
VS



EBM (Explainable Boosting Machine)

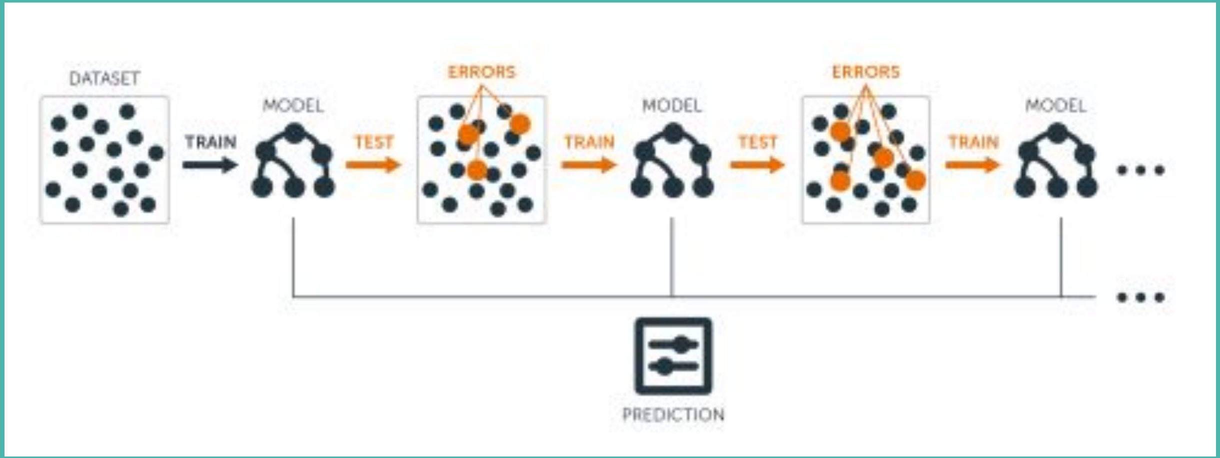
- Newer Model
- Potential solution on explainability

Model Introduction - GBM

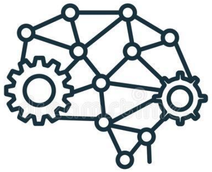


GBM (Gradient Boosting Machine)

- Ensemble Model
- Boosting Method - Convert weak learners to strong learners
- Gradients in loss function



Model Introduction - EBM



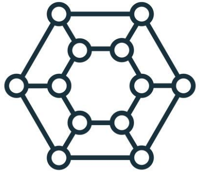
EBM (Explainable Boosting Machine)

- Generalized additive model (GAM)
- Learn each feature function f using modern techniques(bagging, gradient boosting, etc.)
- Auto-detect and focus more on interaction term
- Easy to reason about each feature contribution

$$g(E[y]) = \beta_0 + \sum f_i(x_i) + \sum f_{i,j}(x_i, x_j)$$

Executive Summary

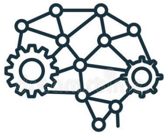
GBM



AUC: .61
Running Time: 229s (All)

AUC: .61
Running Time: 22s (Selected)

EBM



AUC: .62
Running Time: 71s (All)

AUC: .61
Running Time: 11s (Selected)

Objectives Recap

Comparison

AUC, Confusion Matrix, Training Time

EBM Explainability

Feature Importance

KDD Cup 1998 Dataset Introduction

Mail for Donations

Sending mail offers with different promotion cards to target users and ask for donations.

2

Response Variables

Target_B: whether the user donates

Target_D: how much the user donates

90K+

Observations

Size of training and validation dataset is fairly large.

Each row represents a potential donor

479

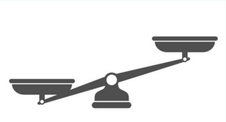
Features

219 **numeric** variables

254 **categorical** variables

Image from https://www.krone.at/1726171

Data Processing Challenges



Unbalanced Response Variable

95% are Non-Donors
5% are Donors

Ambiguous Variables and Variables with significant amount of missing values

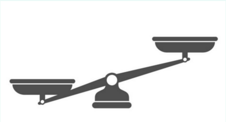
DW3: Percent Duplex Structure

Encoded Categorical Variables

2nd byte = Socio-Economic status of the neighborhood
1 = Highest SES
2 = Average SES
3 = Lowest SES

Image form:
https://www.flaticon.com/free-icon/exchange_1017368
https://time.ly
https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.pngegg.com%2Fen%2Fsearch%3Fq%3Drfm&psig=AOvVaw1p_j4B

Solutions



Unbalanced Response Variable

Undersampling

Ambiguous Variables and Variables with significant amount of missing values

Feature engineering/Feature selection

Encoded Categorical Variables

One-hot encode all the digits

Image form:
https://www.flaticon.com/free-icon/exchange_1017368
https://time.ly
https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.pngegg.com%2Fen%2Fsearch%3Fq%3Drfm&psig=AOvVaw1p_i4Bc

Encoded categorical variables

Let's look at RFA status for RFA_2, RFA_3……. to RFA_22:
It's a 3 digit code representing the recency/frequency/amount status of the donors.
present recency based on the date of the last gift, it has 6 categories,
represent frequency based on period of recency i has 4 categories, and
the amount in the last gift, it has 7 categories.

F=FIRST TIME DONOR Anyone who has made their first donation in the last 6 months and has made just one donation.

N=NEW DONOR Anyone who has made their first donation in the last 12 months and is not a

1=
2=
3=
4=

A=$0.01 - $1.99
B=$2.00 - $2.99
C=$3.00 - $4.99
D=$5.00 - $9.99
E=$10.00 - $14.99
F=$15.00 - $24.99
G=$25.00 and above

ecency

as not

S=STAR DONOR STAR Donors are individuals who have given to 3 consecutive card mailings.

Solution

Separate the code into multiple variables, suffix them according to it's digits order

→

One-hot encode all the digit-variables according to it's possible values

→

Merge all the data and make sure it is done correctly

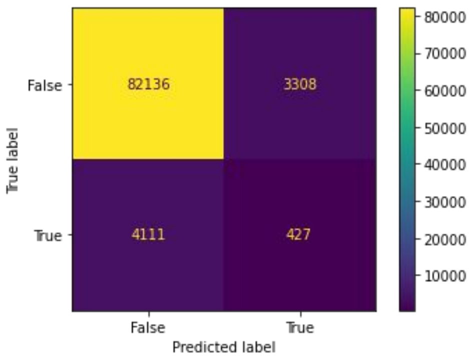
Two Selection Methods

- All Variables (Numerical and OHE Categorical)
- Subset Feature Selection
  - 23 key features that makes the most sense to common intuition in predicting donor/non-donors.
- More scientific approach to Feature Selection/Engineering (Future goal)
  - T-test, Permutation test for scouting important features
  - Lasso
  - Select top-features by training a random forest classifier based on their importance.

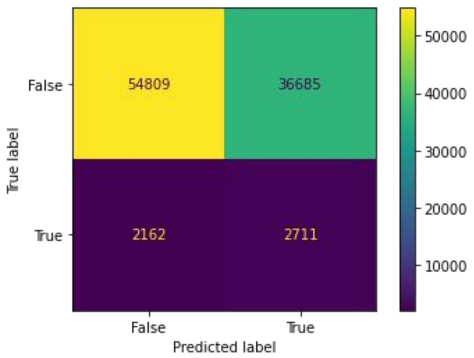
Comparison between EBM and GBM

GBM Models

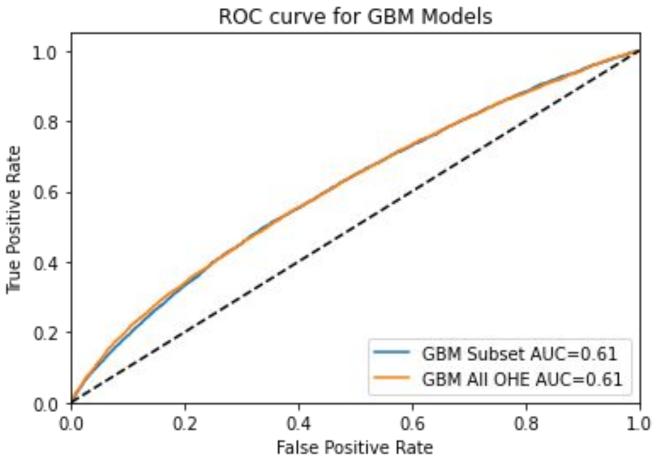
All Variable Stats



Feature Selection Stats

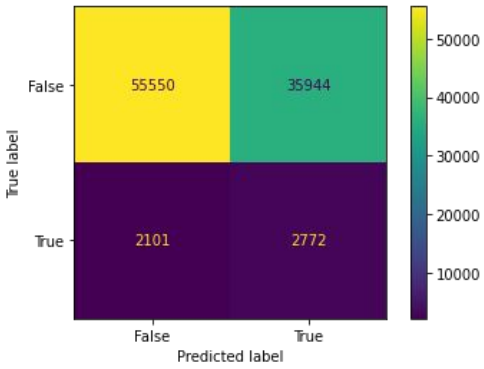


ROC Curve

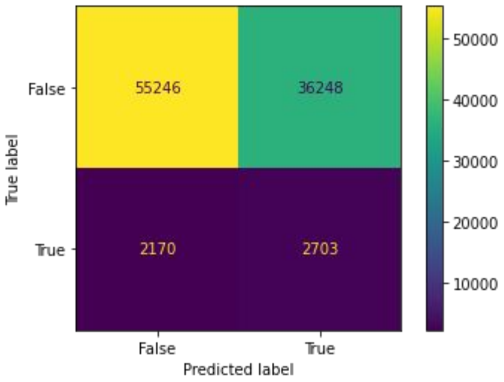


EBM Models

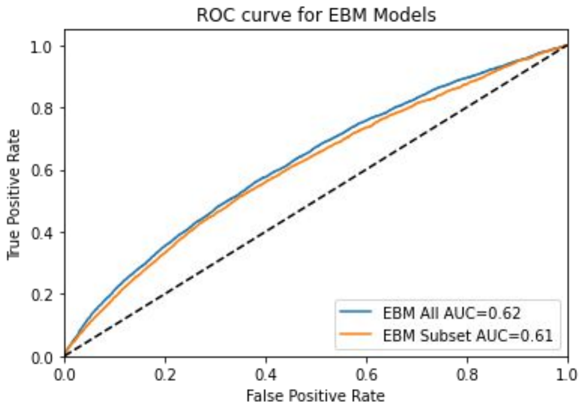
All Variable Stats



Feature Selection Stats



ROC Curve



EBM Important Features in All vs Subset

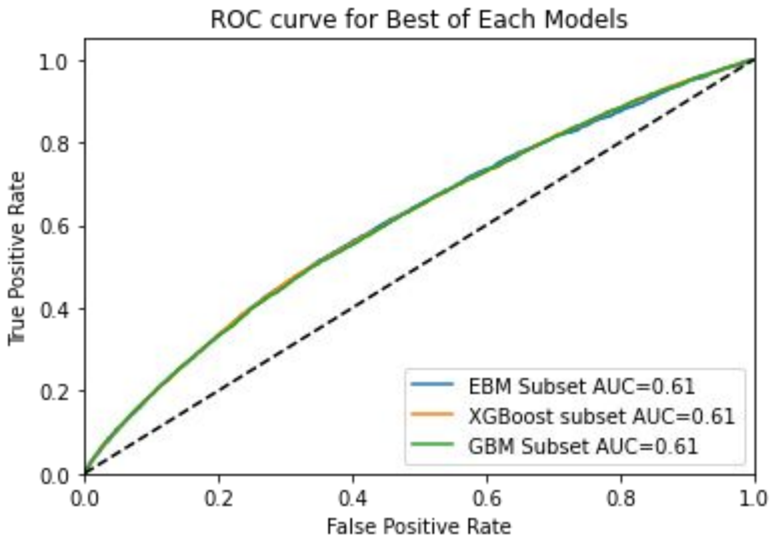
Most Important features using all Features

PEPSTRFL & CARDGIFT
AGE & RFA_17
PEPSTRFL & RAMNT_15
AGE & RFA_18
PEPSTRFL & MINRAMNT
RFA_2
PEPSTRFL
RFA_2F
PEPSTRFL & NGIFTALL
CARDGIFT
RAMNT_8
RFA_2A
LASTGIFT
PEPSTRFL & FISTDATE
ODATEDW & PEPSTRFL

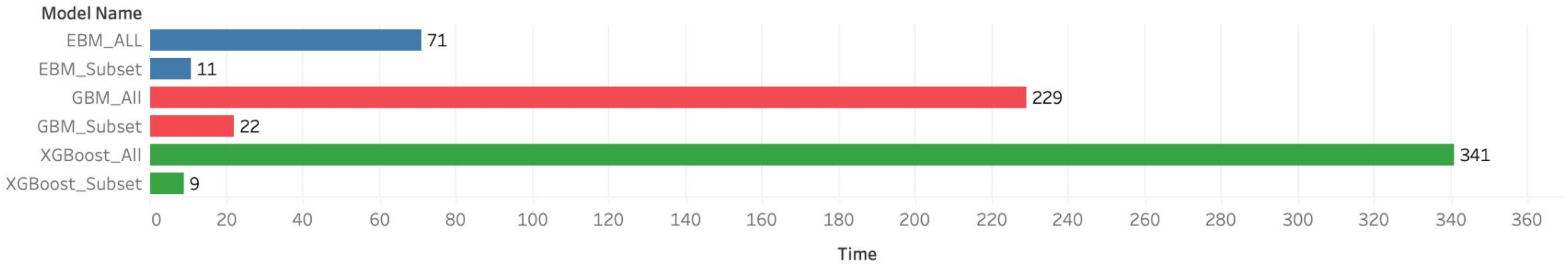
Most Important Features using subset

CARDGIFT
RFA_2_2
AGE
INCOME
DOMAIN_2
TCODE
NUMPRM12
RFA_5_3
MAXRAMNT
CARDPM12
WEALTH2
RFA_6_2
RFA_2_3
CLUSTER
RFA_4_2

EBM vs GBM (Performance)



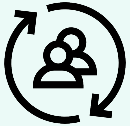
GBM vs EBM (Training Time)



Explainability of EBM

Feature Categories

We briefly classified the 479 features **into 3 categories**.



Social & Economic Status

Age
TCODE (Title)
DOMAIN_2 (Neighborhood)
CLUSTER (Donor group)
WEALTH2 (Family population)
INCOME (Household)

Promotion & Donor Statistics

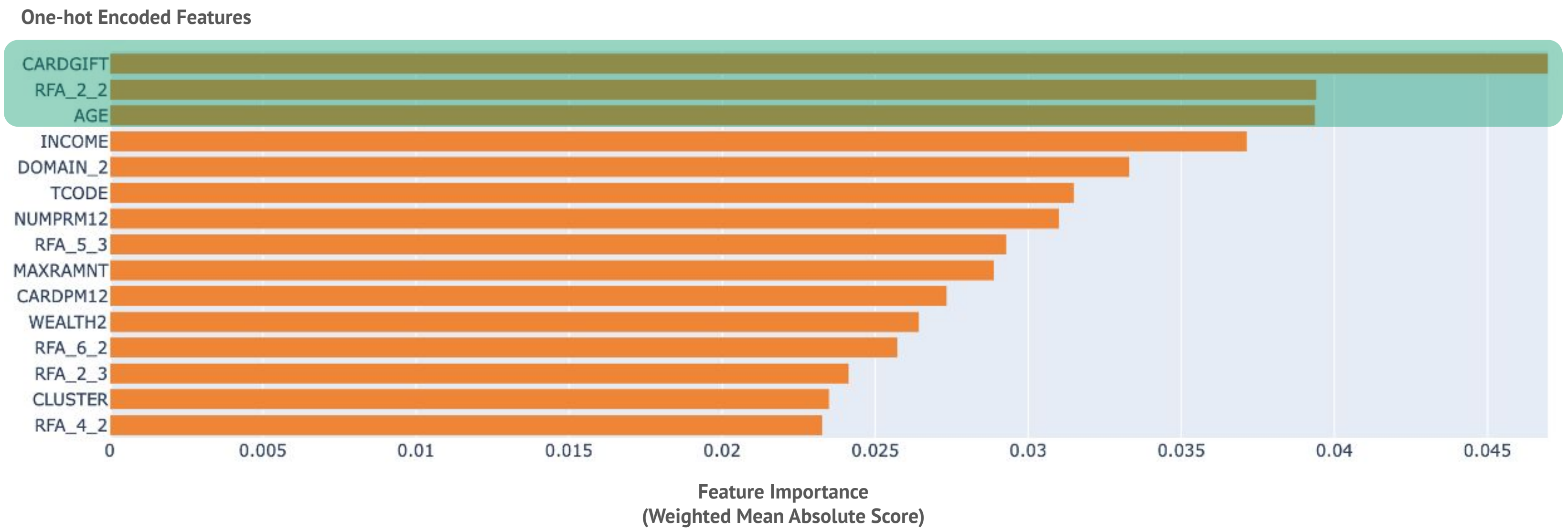
NUMPRM12 (Promotion Number)
CARDPM12 (Card promotion)
NGIFTALL (Donation lifetime)
CARDGIFT (Card to donation)
MAXRAMNT (Largest donation amount)

RFM Status

RFA
(Recency/Frequency/Monetary)

Image form:
https://www.flaticon.com/free-icon/exchange_1017368
https://time.ly
https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.pngegg.com%2Fen%2Fsearch%3Fq%3Drfm&psig=AOvVaw1p_i4B

What affects EBM in predicting donation?

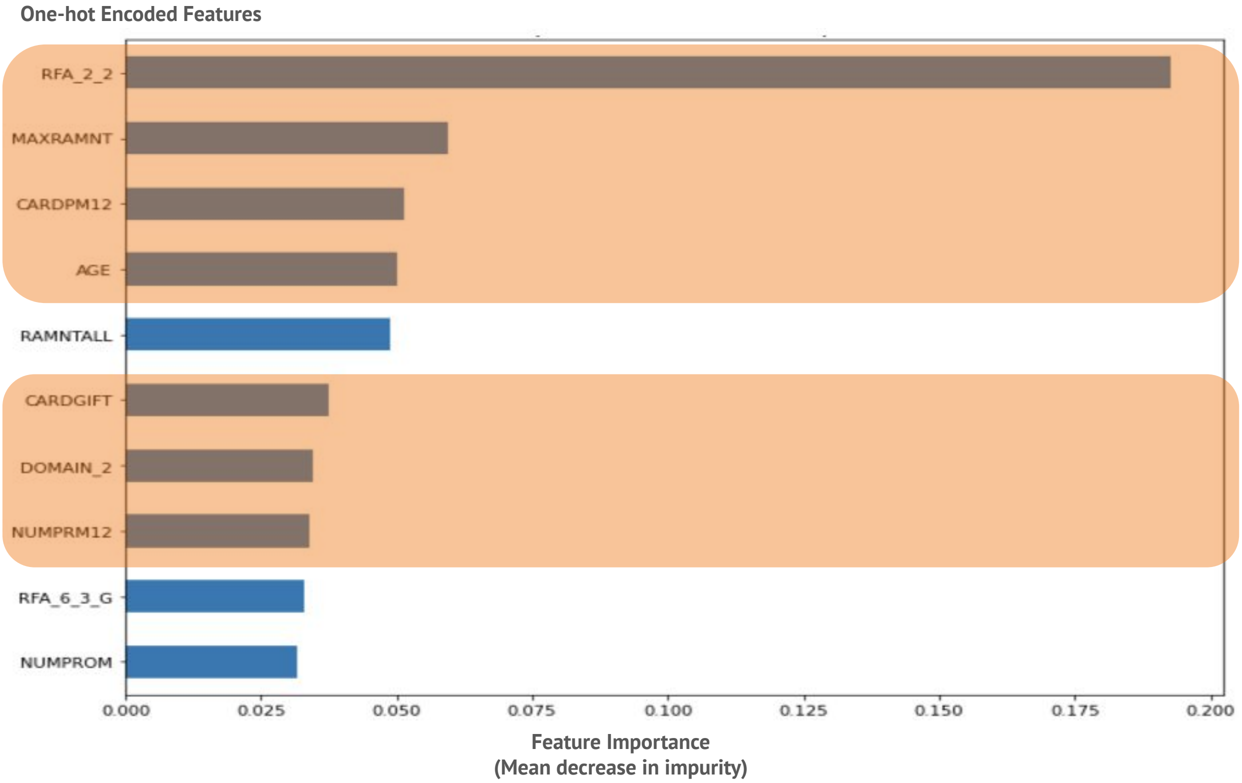


EBM considers:

1) total number of donations to card promotion (CARDGIFT)
2) donation frequency in the year of 1997 (RFA_2_2)
3) age

as the reference for its decision-making.

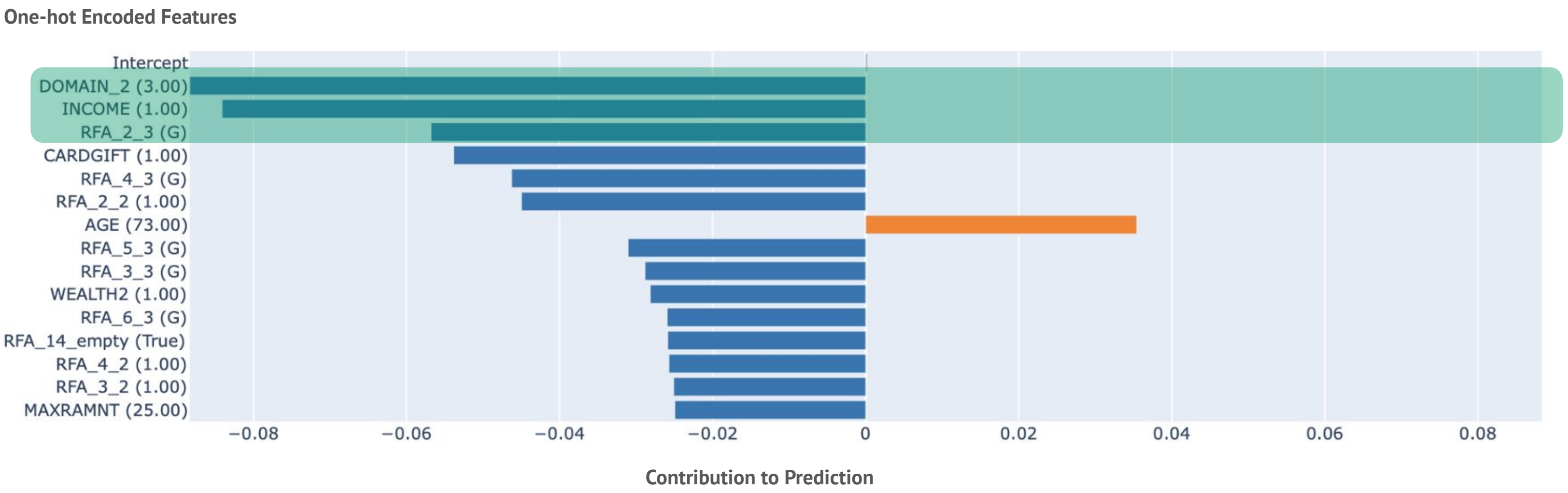
Comparing EBM's decision-making with GBM



Both EBM and GBM consider **similar features in predicting donations**.

What goes to predicting a non-donor?

On the case of **correctly identifying as a non-donor** with confidence level 0.67



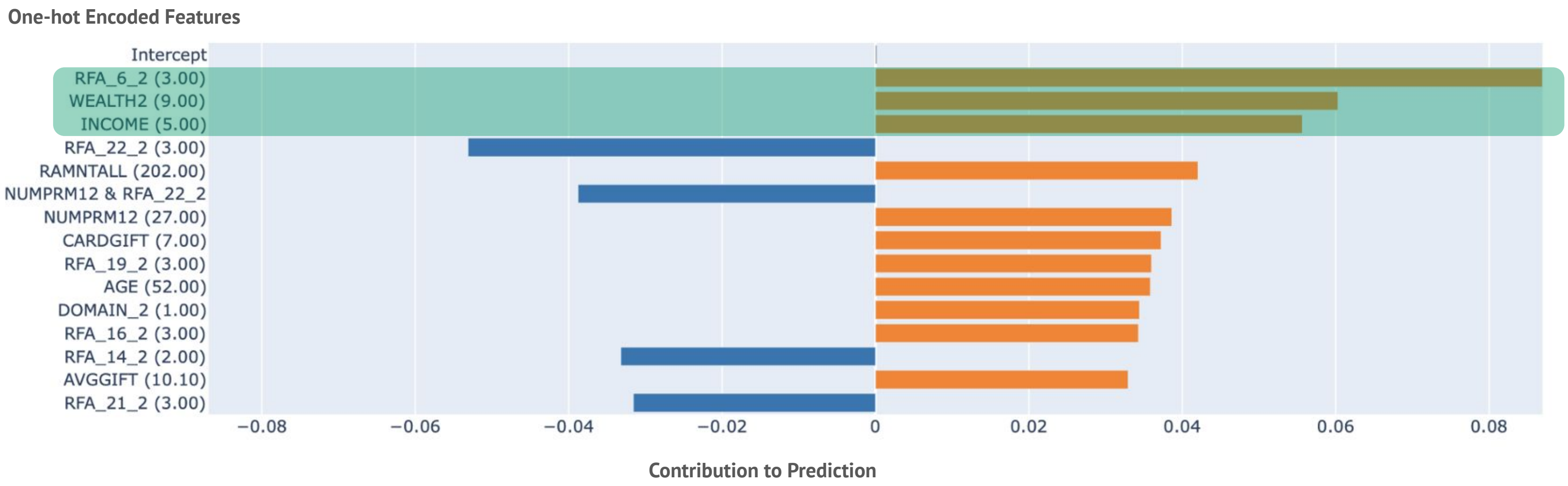
The person who:

1) lives in the neighborhood with the **lowest Socio-Economic status** (DOMAIN_2)
2) has a household **income at the level of 1** (INCOME, range from 0 to 7)
3) donates **more than 25.0 dollars** (amount of G) in the year of 1997 (RFA_2_3)

is predicted as a non-donor for the following promotion.

What goes to predicting a donor?

On the case of **correctly identifying as a donor** with confidence level 0.61

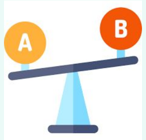


The person who:

1) donates **3 times** (second to the highest category) **in the year of 1996** (RFA_6_2)
2) has **the highest family wealth rating** (WEALTH2, range from 0 to 9)
3) has a **household income at the level of 5** (INCOME, range from 0 to 7)

is predicted as a donor for the following promotion.

Next Steps



Scientific Way of Feature Selection/Engineering

Selected variables with feature engineering

Algorithmic Level Comparison

- ‘What is causing the performance difference?”
- Algorithmic Differences

Explainability to Fintech

- Applicability

Image form:
https://www.flaticon.com/free-icon/exchange_1017368
https://time.ly
https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.pngegg.com%2Fen%2Fsearch%3Fq%3Drfm&psig=AOvVaw1p_i4B

Thank You!
Any Questions?