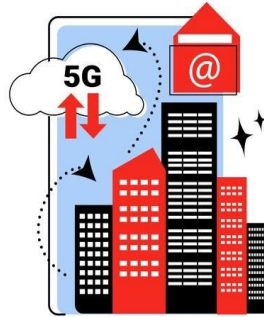


# Churn Prediction

Kowsalya Nitya Vootla



# Business Problem: How can xxx take a more proactive approach to retaining their customers?

## BACKGROUND

- Previous approach to churn management was **reactive**, offering discounts only when customers threatened to leave.
- The analytics team, proposed a **data-driven approach** to predict churn before customers even consider leaving.

## APPROACH

- **Develop a predictive churn model** using machine learning techniques, including logistic regression.
- **Identify key churn drivers**
- **Design targeted retention strategies** based on customer behavior, service usage, and demographics.
- **Quantify the impact of interventions** by simulating retention offers and estimating their effect on churn probability.
- **Evaluate the economic feasibility** using **Customer Lifetime Value (CLV) analysis** over a 5-year period.



# Data

## Demographics

Cust ID

Age

Female

Income

Education

Children

## Department-specific customer purchase history

Freq\_endurance, Freq\_eater,  
Freq\_team, Freq\_strength,  
Freq\_backcountry,  
Freq\_racquet

## Outcome Variable

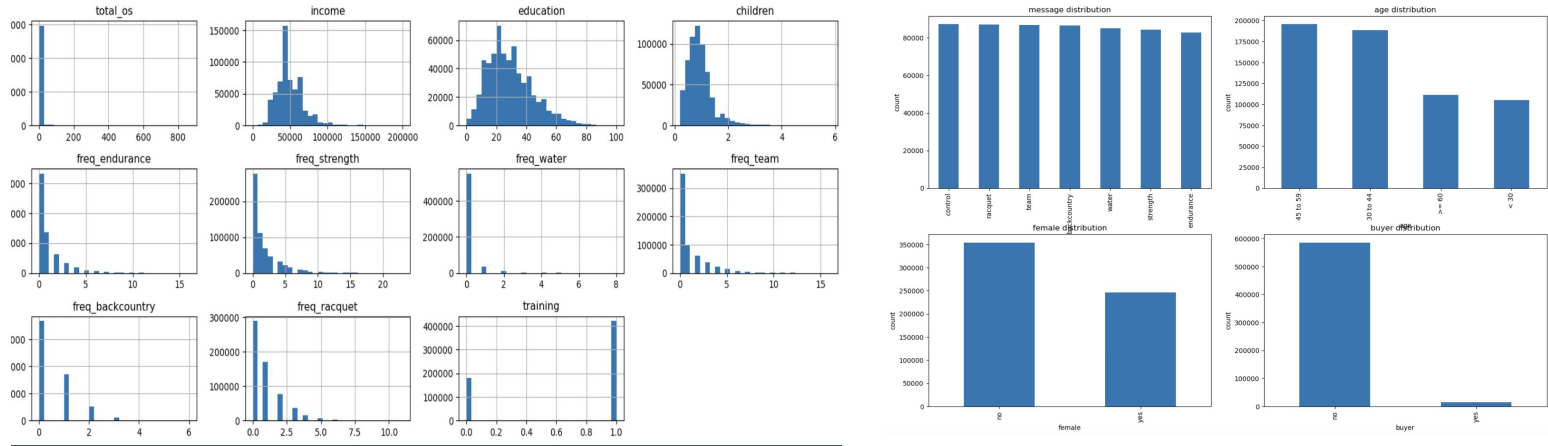
Purchase in 2 days of email  
receipt (y/n)

Total order Size  
(Euros)

Train:Test, Cross Validation Representative Sample

# Exploratory Data Analysis and Transformation

**Data heavily imbalance and skewed towards non-buyers - ONLY 2.4% buyers**



Income, children, education, total\_os left skewed

Multiple Categorical variables (female, age, message)

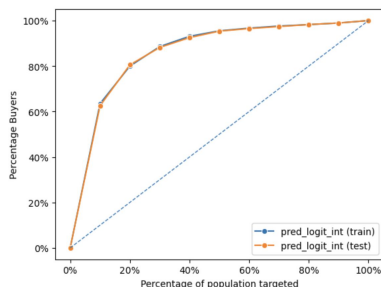
# Predicting Probability of Buying

- **Entire Train Dataset**
- **Rvar:** buyer = yes
- **Evar:** 'income', 'education', 'children', 'freq\_endurance', 'freq\_strength', 'freq\_water', 'freq\_team', 'freq\_backcountry', 'freq\_racquet', 'message\_control', 'message\_endurance', 'message\_racquet', 'message\_strength', 'message\_team', 'message\_water', 'female', 'age\_30\_44', 'age\_45\_59', 'age\_60a'
- **Exclude total\_os from evar**

## Logistic Regression

Log transform

Int terms added:  
message: (all other variables)

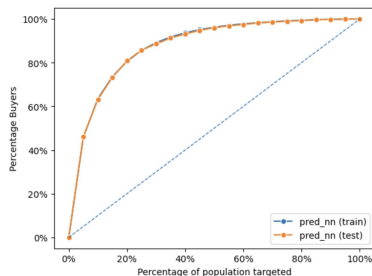


AUC: 0.884

## Neural Network

Scaled df

Hls = (3,3), Alpha = 0.1

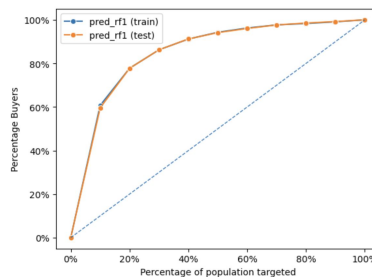


AUC: 0.891

## Random Forest

One-hot encode + convert to  
binary

Max\_features = 6,  
max\_samples = 0.5

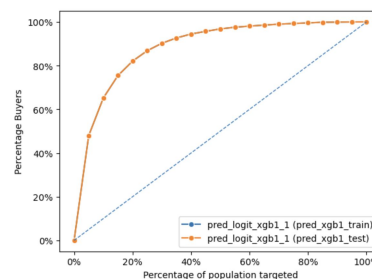


AUC: 0.871

## XGBoost

One-hot encode + convert to  
binary

Max\_depth = 5,  
n\_estimators = 51



AUC: 0.899

## % of customers for whom the message maximizes probability of purchase

Utilizing the predicted probability of purchase to determine the best message per customer:

Message Type	Logistic Regression	Neural Network	Random Forest	XGBoost
Endurance	73.20%	45.34%	87.86%	52.62%
Strength	19.39%	30.49%	2.34%	23.65%
Racquet	4.57%	0.52%	0.15%	6.26%
Water	1.07%	0.17%	5.96%	1.92%
Team	1.05%	23.48%	2.75%	1.78%
Backcountry	0.72%	0.00%	0.002%	7.64%
No message	~0.00%	0.00%	0.912%	6.12%

# Predicting Total Order Size

- **Train Dataset, filter on buyer == 'yes'**
- **Rvar: total\_os**
- **Evar: 'income', 'education', 'children', 'freq\_endurance', 'freq\_strength', 'freq\_water', 'freq\_team', 'freq\_backcountry', 'freq\_racquet', 'message\_control', 'message\_endurance', 'message\_racquet', 'message\_strength', 'message\_team', 'message\_water', 'female', 'age\_30\_44', 'age\_45\_59', 'age\_60a'**

## Linear Regression

Int terms added: message: (all other variables)

### Baseline

$R^2$ : 0.000

$R^2$ : 0.118

MAE: 0.653

MAE: 0.611

MSE: 0.649

MSE: 0.573

## Neural Network

Hls= (3), Alpha = 0.1

### Baseline

$R^2$ : -0.000

$R^2$ : 0.065

MAE: 3.03

MAE: 0.2881

MSE: 22.66

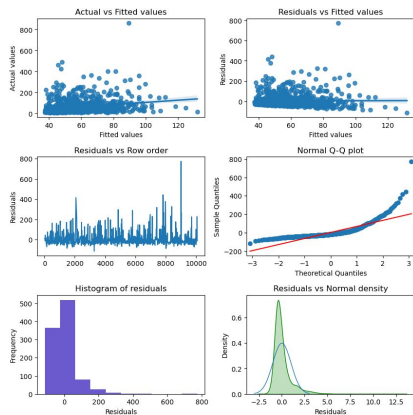
MSE: 21.197

$R^2$  improves, MAE and MSE decreases compared to baseline. However,  $R^2$  are still very low.

# Predicting Total Order Size

## Random Forest

Max\_features = 6 , max\_samples = 0.7



### Baseline

$R^2$ : -0.0001

$R^2$ : 0.703

MAE: 39.531

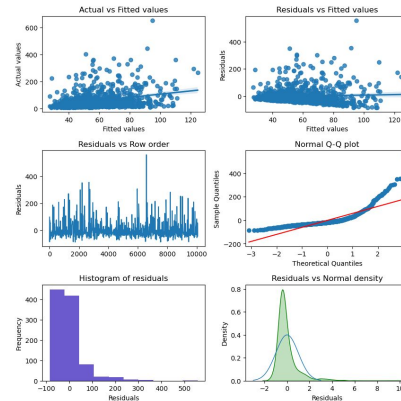
MAE: 37.668

MSE: 3859.43

MSE: 3587.72

## XGBoost

Learning\_rate = 0.2 , max\_depth = 1



### Baseline

$R^2$ : -0.0001

$R^2$ : 0.0744

MAE: 39.53

MAE: 37.48

MSE: 3859.43

MSE: 3572.61

$R^2$  improves, MAE and MSE decreases compared to baseline. However,  $R^2$  are still very low and residual plots are suboptimal with visible heteroscedasticity.

Decision (approach to calculating order size)

Utilize Average order size for each message type from historical buyer data



## % of customers for whom the message maximizes expected profit

Expected Profit = probability of purchase \* average order size \* (1-COGS)  
COGS = 60%

Message Type	Logistic Regression	Neural Network	Random Forest	XGBoost
Endurance	56.47%	33.49%	0%	37.79%
Strength	12.5%	4.56%	0%	17.28%
Racquet	17.48%	2.18%	0%	13.1%
Water	6.43%	0.17%	100%	17.61%
Team	3.44%	34.40%	0%	3.3%
Backcountry	0.72%	~0.00%	0%	8.57%
No message	~0.00%	0.00%	0%	2.35%

# Expected profit on average (Euros) is highest for customized messages

	Logistic Regression	Neural Network	Random Forest	XGBoost
Customized message	0.68	0.65	0.61	0.70
Randomised message	0.55	0.55	0.55	0.55
Same message				
Endurance	0.598	0.604	0.518	0.603
Strength	0.576	0.583	0.525	0.581
Racquet	0.568	0.584	0.589	0.583
Water	0.618	0.639	0.610	0.627
Team	0.574	0.598	0.580	0.585
Backcountry	0.545	0.570	0.570	0.569
No message	0.426	0.447	0.482	0.447

## Extrapolated profits to 5M customers & Improvement metrics

Extrapolated profit when...	Logistic Regression	Neural Network	Random Forest	<u>XGBoost</u>
<b>Personalized message is sent</b>	€3,394,907.78	€3,267,902.93	€3,051,997.78	€3,486,245.57
Improvement in profit (%) when <b>personalized message</b> is sent against				
<b>Same message</b>	9.81%	2.33%	0%	11.14%
<b>Random message</b>	23.03%	18.42%	10.60%	26.34%
<b>No message</b>	59.43%	46.31%	26.70%	56.00%

- Personalized messaging seems to be the best approach
- XGBoost delivers the highest % improvement in profit compared to same/random/no message



# Evaluating and Improving the New E-mail Policy Proposal

## STRENGTHS

### ✓ **Data-Driven Decision Making**

– Allocates promotional emails based on expected profitability

### ✓ **Fair Departmental**

**Distribution** – Split emails between the top two high-profit messages.

### ✓ **Continuous Learning** –

Monthly re-evaluation refines predictions using recent data.

## WEAKNESS

✗ **Bias Toward High-Performing Messages & Customer fatigue** – Other messages may never be tested.

✗ **Short-Term Profit Focus** – May lead to aggressive marketing strategies that hurt long-term engagement.

✗ **No Built-in Experimentation** – Lacks A/B testing to improve predictions and customer response modeling.

## IMPROVEMENTS

♦ **Introduce A/B Testing** – Randomly assign some customers a different message type to explore effectiveness.

♦ **Control for Confounding Factors** – Consider seasonality, self-selection bias, and prior purchases.

♦ **Implement Uplift Modeling** – Measure impact by comparing  $P(\text{Purchase} \mid \text{Message})$  -  $P(\text{Purchase} \mid \text{No Message})$  in randomized trials.

