SCORING MODEL

Matteo Racioppi (5410938)
Sara Dellacasa (5409694)
Giulia Franceschina (5407529)
Aberto Edoardo Faoro (5406685)

PRELIMINARY ANALYSIS

01 PRELIMINARY ANALYSIS

02 DATA AUDIT

03 MODEL ESTIMATING

04 LIFT CHART

05 CAMPAIGN MANAGEMENT

EXECUTIVE SUMMARY

In this project, our goal is to support the bank's upselling marketing strategy by developing a **predictive model** that identifies customers who are **most likely to respond positively** to the offer. We will begin with an **exploratory data analysis** to understand the structure of the dataset and **key indicators** such as the overall acceptance rate and the distribution of relevant KPIs.

Following this, a **thorough data audit** will be conducted to **ensure data quality** by removing constant variables, handling missing data, and eliminating highly correlated features that may impact model performance.

With the cleaned dataset, we will train and evaluate a **logistic regression model** to predict customer acceptance. The model's effectiveness will be assessed using a **lift chart** to quantify its added value over random targeting. Additionally, we will estimate the **optimal number of customers** to contact in order to maximize the campaign's profitability.

This data-driven approach aligns analytical insights with business objectives to enhance the efficiency and return on investment of the bank's upselling initiatives.

1. PRELIMINARY ANALYSIS

In this part of the analysis, we were asked to **evaluate the dataset's structure** by reporting the number of rows and columns, calculating the **average acceptance rate**, and exploring the univariate **distribution of key performance indicators (KPIs)**.
For the univariate analysis, we selected a **subset** of numerical variables based on their business relevance and potential predictive power. Specifically, we focused on variables that reflect **customers' financial capacity** (e.g., total assets, annual inflows), their activity and engagement with the bank (e.g., number of logins, total transactions), and specific **behavioral indicators** (e.g., withdrawals, inflows in the subsequent period). These variables are logically associated with a customer's likelihood to respond to an upselling offer and offer valuable insights into both their financial profile and level of involvement with the bank.

NUMBER OF ROWS AND COLUMNS

Rows = 19120

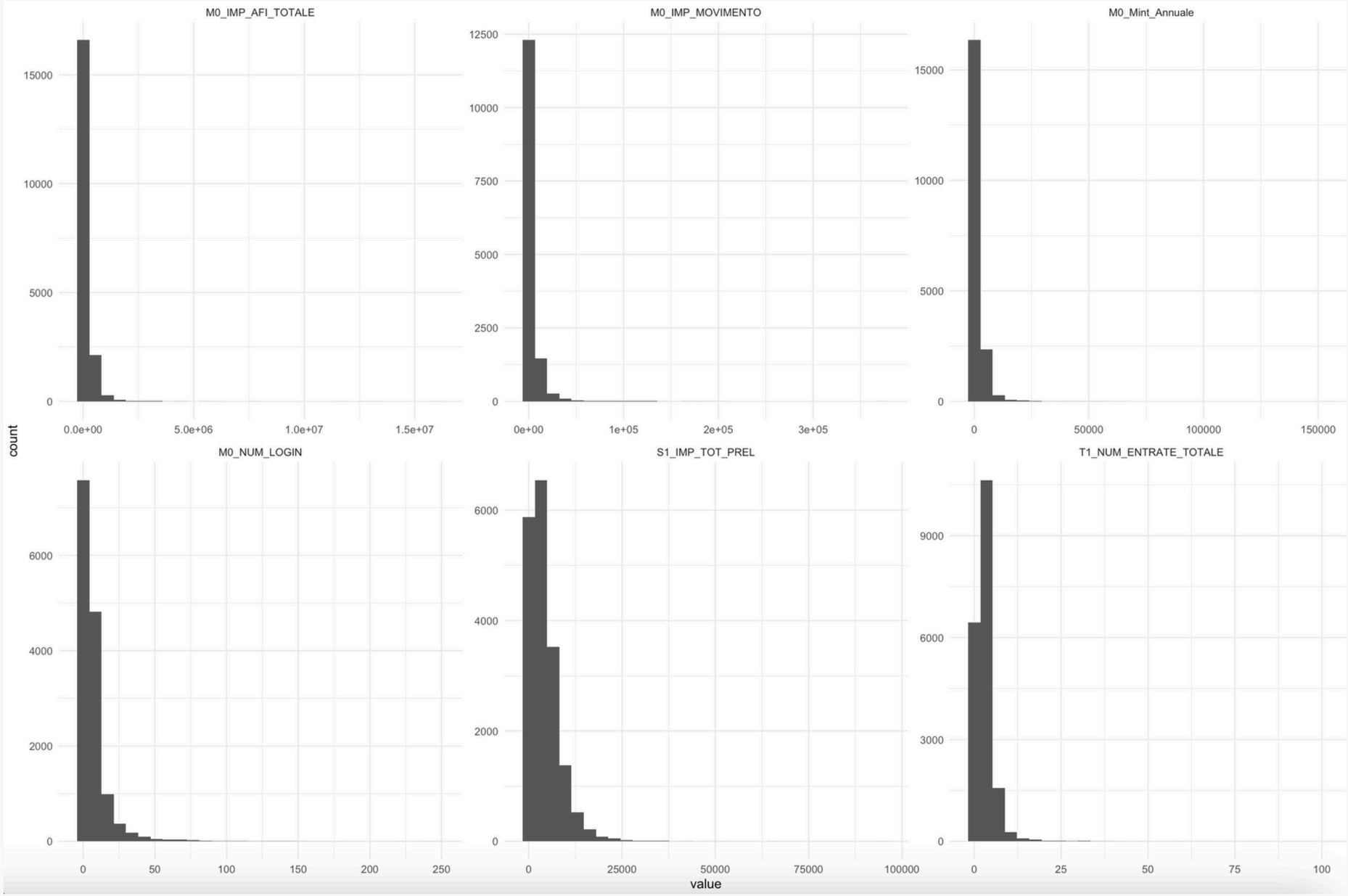
Column = 35

AVERAGE ACCEPTANCE RATE

0.15

The dataset contains a total of **35 columns**, including one target variable, **FLG_TARGET**, which indicates whether a customer accepted the upselling offer. The remaining 34 columns consist of **key performance indicators (KPIs)** that are considered essential for predicting the likelihood of a positive response.
Additionally, the average acceptance rate (representing the proportion of customers who responded positively to the campaign) was calculated to be approximately **15%**.

PRELIMINARY ANALYSIS



Univariate analysis of the KPIs

The univariate analysis of the selected numerical KPIs reveals that **all variables exhibit strong right-skewness**. Most customers are concentrated at the **lower end of the distribution** for financial metrics (such as total assets, annual inflows, total movements, and withdrawals), engagement indicators (e.g., number of logins), and activity measures (e.g., number of inflows in T1).

In contrast, a **small subset of customers** displays **significantly higher values**, suggesting the presence of **outliers or extreme cases**. This result is reasonable, since In most customer datasets, the majority of customers have low or moderate values, and only a few customers have very high values.

This distribution highlights a **heterogeneous customer base**, where the majority show low to moderate levels across these dimensions, while only a few demonstrate substantially higher activity or financial capacity. **These patterns underscore the variability in customer behaviour and support the use of a predictive model to effectively identify those more likely to respond to the upselling campaign.**

2. DATA AUDIT

We are checking for:

- Presence of constant columns (i.e. standard deviation = 0)
- Presence of columns with many Nasty (i.e. > 70%)
- Presence of sets of highly correlated features (i.e. ρ > 0.6/0.7)

} **DROP THEM**

CONSTANT STANDARD DEVIATION

```
> numeric_cols <- sapply(data, is.numeric)
> std_dev <- sapply(data[, numeric_cols], sd, na.rm = TRUE)
> constant_columns <- names(std_dev)[std_dev == 0]
> constant_columns
character(0)
```

HIGH PRESENCE OF NA_S

```
> missing_pct <- sapply(data, function(x) mean(is.na(x)))
> high_missing <- names(missing_pct[missing_pct > 0.7])
> print(high_missing)
character(0)
```

HIGH CORRELATION

```
> highly_corr_names <- findCorrelation(corr_matrix, cutoff = 0.7, names = TRUE)
> cat("Highly correlated columns to remove: ", highly_corr_names, "\n")
Highly correlated columns to remove:  T1_IMP_AFI_TOTALE M0_IMP_AFI_TOTALE M0_IMP_ATTIVITA_FINANZIARIE_12M M0_IMP_AFI_TO
TALE_12M M0_IMP_FONDI_COMUNI_12M T1_IMP_FONDI_COMUNI T1_IMP_ENTRATE_TOTALE S1_NUM_TOT_ACQ_STANDARD
```

NO columns with constant SD or with **more than 70% of NAs** have been founded.
In contrast, the **set of variables** here indicated, are the one that we **needed to remove** from the dataset due to **high correlation**

3. MODEL ESTIMATION

1. We started by splitting the dataset into two parts:

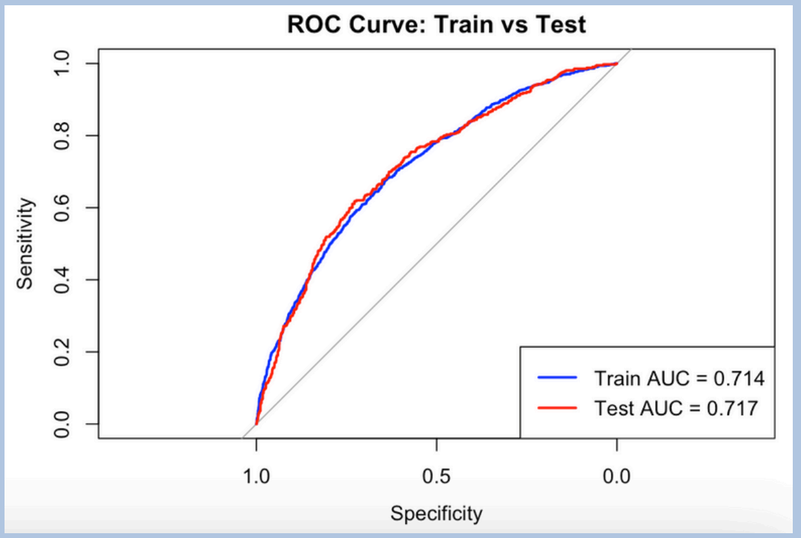
TRAIN DATASET 70% OF OBS

TEST DATASET 30% OF OBS

We fitted a full logistic regression model on the training dataset:

2. **Full model**

```
# Fit logistic regression
model <- glm(FLG_TARGET ~ ., data = train_data, family = "binomial")
summary(model)
```



The ROC curve shows **very similar performance** between the training and test datasets, with both curves following a **similar trajectory** and achieving nearly identical AUC values. This indicates that the model generalizes well and is **not overfitting** the training data. The AUC being stable across both datasets confirms that the model captures the underlying patterns in the data without fitting noise. Overall, the model demonstrates good discriminatory power and robustness when applied to unseen data.

3. **Variable selection**

Since from the resulting summary of the full model, a **distinct number of variables were not significant**, we decided to perform **variable selection**, based on the Akaike Information Criterion (AIC) because it balances model fit and complexity, by removing unnecessary predictors. The model with the lowest AIC found is the following:

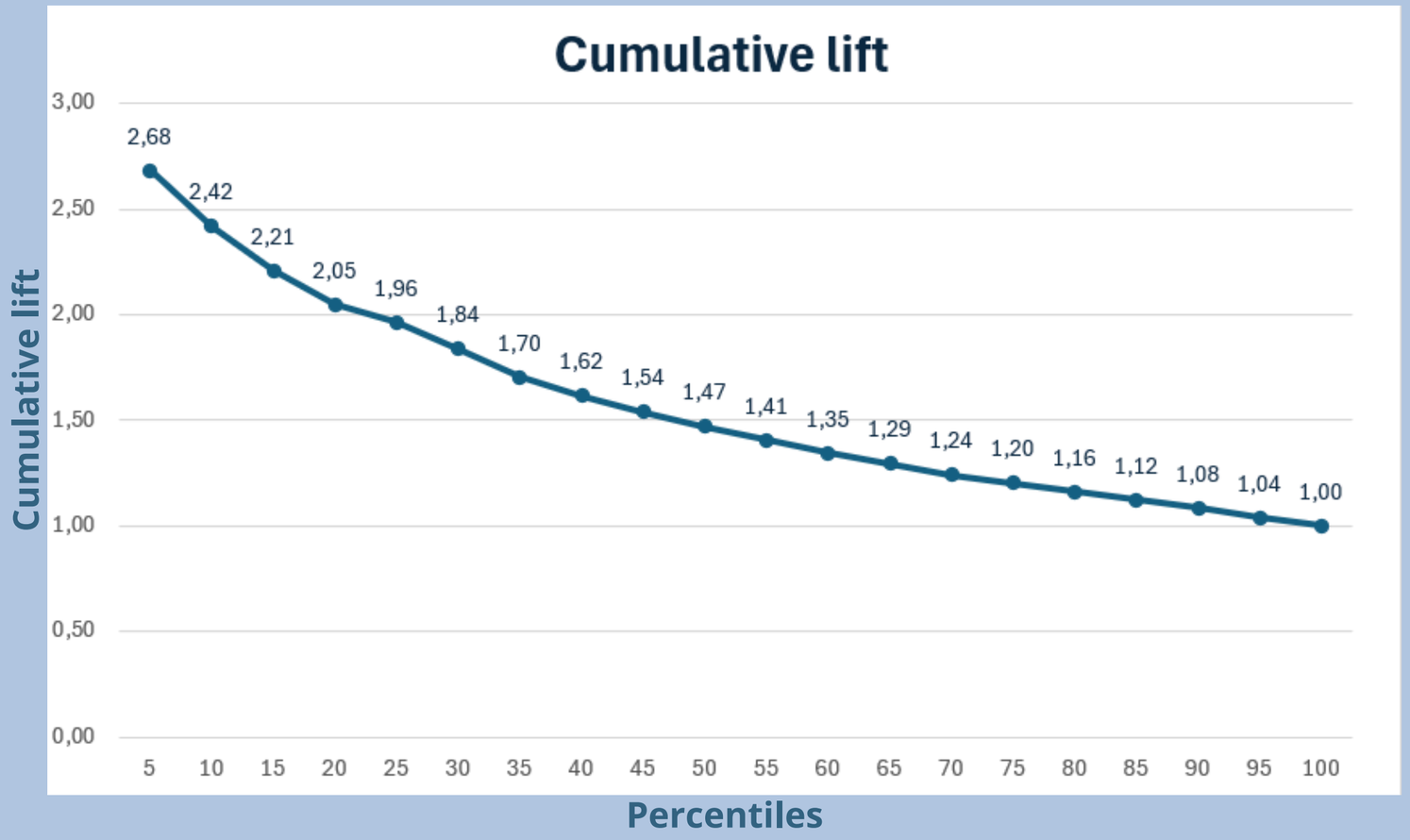
```
stepwise_model <- stepAIC(model, direction = "both")
summary(stepwise_model)
```

```
Step:  AIC=4659.02
FLG_TARGET ~ M0_IMP_FONDI_COMUNI + M0_P_INV_GESTITI + M0_IMP_MOVIMENTO +
    M0_NUM_LOGIN + SM1_NUM_CONTRATTI + SM2_NUM_CONTRATTI + M0M1_NUM_CONTATTI_OUTB +
    S1_NUM_INCONTRI_OUTB + M0_NUM_STIPENDI_TOTALE_12M + T1_NUM_ENTRATE_TOTALE +
    T1_NUM_FILIALE_USCITE + SM1_NUM_MOVIMENTI_TOTALE + ST3_IMP_SPEND_CARTE_CREDITO_TOT +
    S1_IMP_TOT_PREL + S1_IMP_TOT_ACQ_STD_ABBIGL + S1_IMP_TOT_ACQ_STD_TLIBERO +
    S1_NUM_TOT_ACQ_STD_CASA_FAM + TIPO_MULTIB
```

4. LIFT CHART

Percentile	Customers	Cum_Customers	Upsells	Cum_Upsells	Redemption	Cum_Redemption	Captured_Response	Cum_Captured_Response	Lift	Cum_Lift
5	379	379	183	183	0,482850	0,134164	0,134164	0,134164	2,682930	2,682930
10	379	758	147	330	0,387863	0,241935	0,107771	0,241935	2,155141	2,419036
15	379	1.137	122	452	0,321900	0,331378	0,089443	0,331378	1,788620	2,208897
20	379	1.516	107	559	0,282322	0,409824	0,078446	0,409824	1,568708	2,048850
25	379	1.895	111	670	0,292876	0,491202	0,081378	0,491202	1,627351	1,964550
30	379	2.274	82	752	0,216359	0,551320	0,060117	0,551320	1,202187	1,837490
35	379	2.653	61	813	0,160950	0,596041	0,044721	0,596041	0,894310	1,702750
40	379	3.032	69	882	0,182058	0,646628	0,050587	0,646628	1,011597	1,616356
45	379	3.411	63	945	0,166227	0,692815	0,046188	0,692815	0,923632	1,539386
50	379	3.790	57	1.002	0,150396	0,734604	0,041789	0,734604	0,835667	1,469014
55	379	4.169	53	1.055	0,139842	0,773460	0,038856	0,773460	0,777024	1,406106
60	379	4.548	46	1.101	0,121372	0,807185	0,033724	0,807185	0,674398	1,345130
65	379	4.927	46	1.147	0,121372	0,840909	0,033724	0,840909	0,674398	1,293536
70	379	5.306	37	1.184	0,097625	0,868035	0,027126	0,868035	0,542450	1,239887
75	379	5.685	45	1.229	0,118734	0,901026	0,032991	0,901026	0,659737	1,201210
80	379	6.064	35	1.264	0,092348	0,926686	0,025660	0,926686	0,513129	1,158205
85	379	6.443	36	1.300	0,094987	0,953079	0,026393	0,953079	0,527790	1,121122
90	379	6.822	31	1.331	0,081794	0,975806	0,022727	0,975806	0,454485	1,084086
95	379	7.201	15	1.346	0,039578	0,986804	0,010997	0,986804	0,219912	1,038604
100	378	7.579	18	1.364	0,047619	1,000000	0,013196	1,000000	0,264593	1,000000

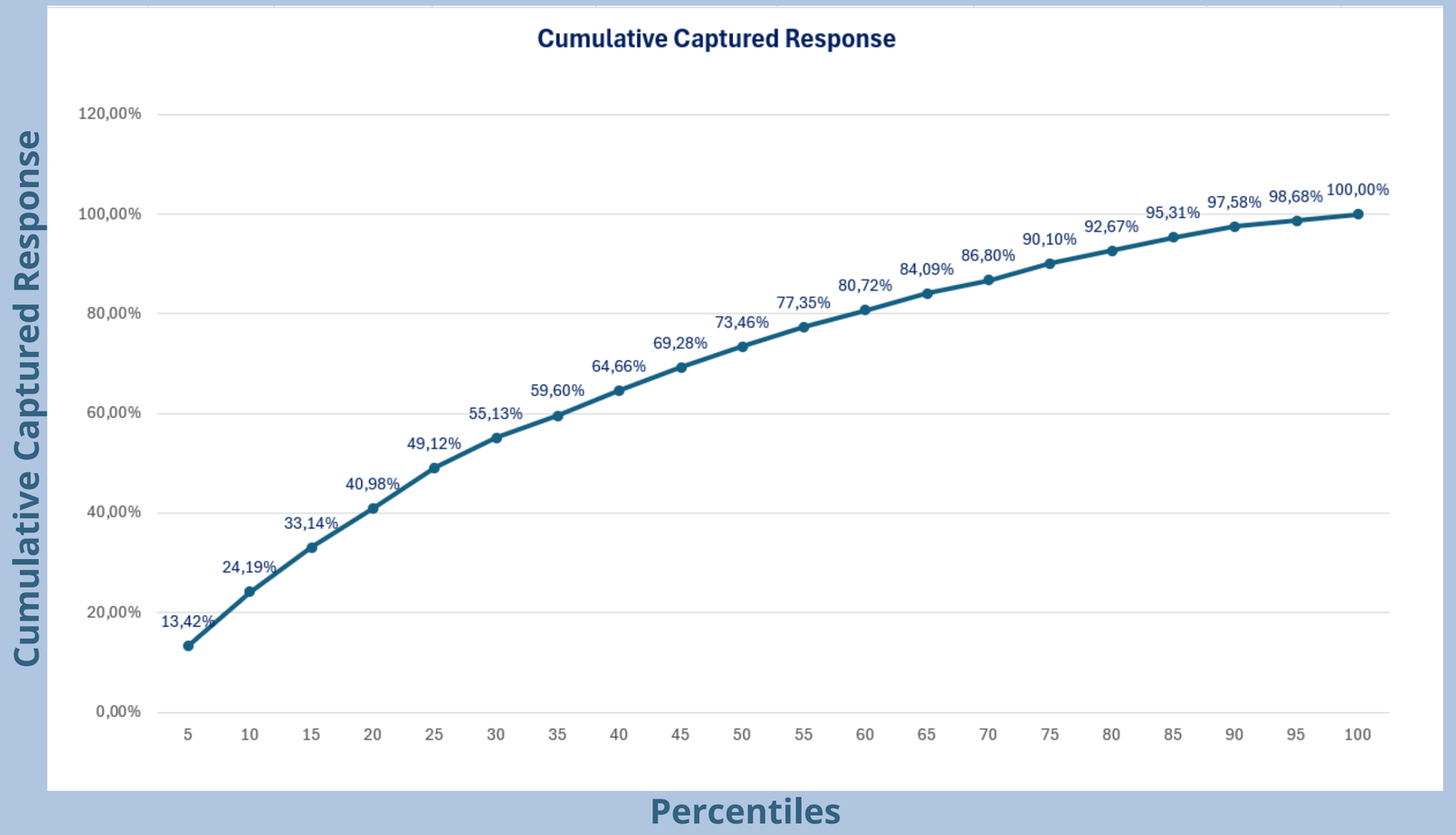
CUMULATIVE LIFT CHART



- For the Top 5% customers, the expected redemption is 2.68 times the overall redemption (Mean Value)

- For the Top 15% customers, the expected redemption is 2,21 times the overall redemption (Mean Value)

THE GAIN CHART



- Top 5% of customers concentrate 13.42% of Overall Positive Events
- Top 15% of customers concentrate 33.14% of Overall Positive Events

5. CAMPAIGN ASSUMPTION

Given that the trained model exhibits robustness without signs of overfitting, we can reasonably assume that the score distribution observed in our sample (representative of the broader customer base) also holds for the full target population of one million customers. By segmenting the customer base into ventiles, we can extrapolate comparable cumulative lift values across the entire population and estimate the expected number of positive targets within each ventile accordingly.

Number of customers:
1 Milion

Value of each customer:
120€

Percentage of respondents:
20%

Unit contact costs:
20€

Fixed contact costs:
100.000€

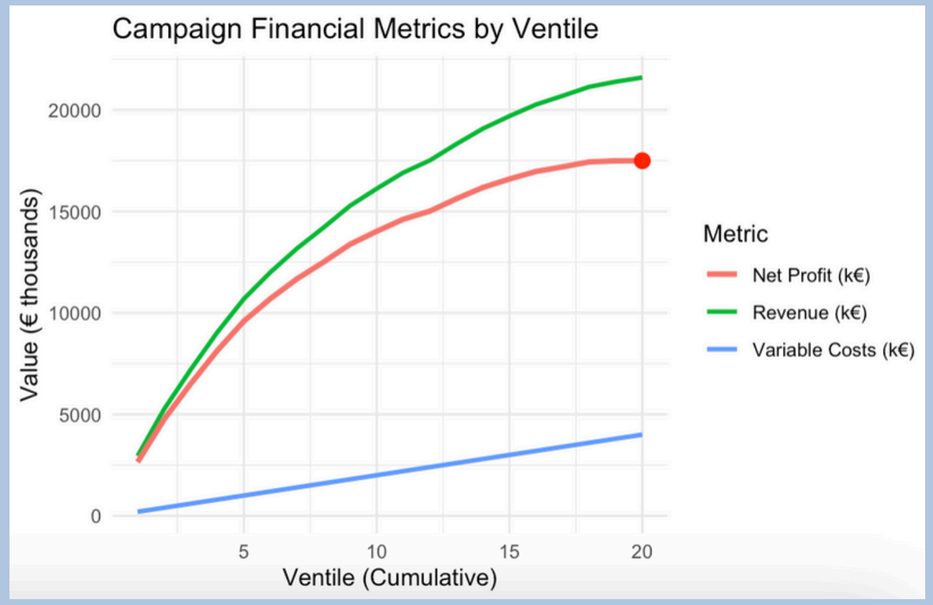
REVENUES AND COSTS

The plot illustrates three key financial metrics—Revenue, Net Profit, and Variable Costs—accumulated across ventiles, which divide the customer base into 20 equally sized segments ranked by model score.

- Revenue (green line) grows steadily and reaches its maximum at the 20th ventile, showing a strong cumulative return as more customers are included.
- Net Profit (red line) also increases, but begins to plateau around the final ventiles, suggesting diminishing marginal returns toward the lower-scoring segments. The red dot at the end marks the total net profit when the entire customer base is targeted.
- Variable Costs (blue line) increase linearly, reflecting a constant per-customer cost structure.

The gap between revenue and variable costs (i.e., net profit) narrows in the later ventiles, indicating that the most profitable segments lie in the top-scoring ventiles. This implies that a more selective targeting strategy—focusing on the highest-scoring ventiles—could improve campaign efficiency.

	ventile	customers_sample	upsells_sample	cum_customers_sample	cum_upsells_sample	cum_capture_rate	cum_lift	cum_customers_pop	expected_churns_cum	expected_revenue	expected_variable_cost	expected_fixed_cost	net_profit
1	20	378	13	7579	1364	1	1	1e+06	179971	21596517	4e+06	1e+05	17496517



THANK YOU!