MUSEUM PASS

Piemonte Museum Pass

Daniele Iepre
870960

BIG DATA IN ECONOMICS
A.Y 24/25

MUSEUM PASS

About Project

- Business problem
- Dataset presentation
- Museum insights
- Clients insights
- Correlation analysis
- Clustering
- Network analysis
- Causality
- Churn prediction
- Retention Campaign

MUSEUM PASS

Business problem

The Museum Card gives access to a wide network of museums in Turin and surrounding areas. Users can visit museums unlimited times during the validity period (1 year).

Many users, however, do not renew, leading to high churn rates.

Our goal

- Identify the drivers of churn
- Build predictive models to anticipate churn
- Design a targeted marketing campaign to improve retention and profitability, balancing costs and incentives

Dataset presentation

data1.csv	General customer info + churn label (si2014)
an13.csv	Demographics, payment info, discount type
in13.csv	Museum visits: date, time, museum

MUSEUM PASS

Dataset preparation

Data1

- Removed an13 and ultimo_ing.x; replaced them with the corresponding month values
- Removed abb14 as it was not needed for modeling.
- Created the churn variable:
  - churn = 1 if the customer did not renew (Si2014 = 0)
  - churn = 0 otherwise

An13

- Removed cap and profession
- Filtered year of birth to retain only customers:
  - Aged 6 or older (no card needed for younger children)
  - Aged under 95 (to exclude unrealistic ages)
- Simplified discount and reduction variables by grouping similar types together.

In13

- Removed entries where orai was 00:00:00 (invalid visit times).
- Excluded province, and city fields to simplify the analysis

Feature engineering

- Aggregated visit data using groupby on Codcliente:
  - Created quota_risparmiata (total savings from discounts)
  - Calculated numero_visite (total number of visits)
  - Counted musei_unici (unique museums visited)
  - Extracted ultimo_ingr_mese (month of last visit)
- Generated a new dataset from these groupby features.

Final integration

- Performed inner join on CodCliente between:
  - Aggregated visit dataset
  - an13.csv
  - data1.csv
- Applied:
  - Binarization for discount (sconto) and reduction (riduzione) types
  - One-hot encoding for other categorical variables
  - Multicollinearity removal for cleaner model input

MUSEUM PASS

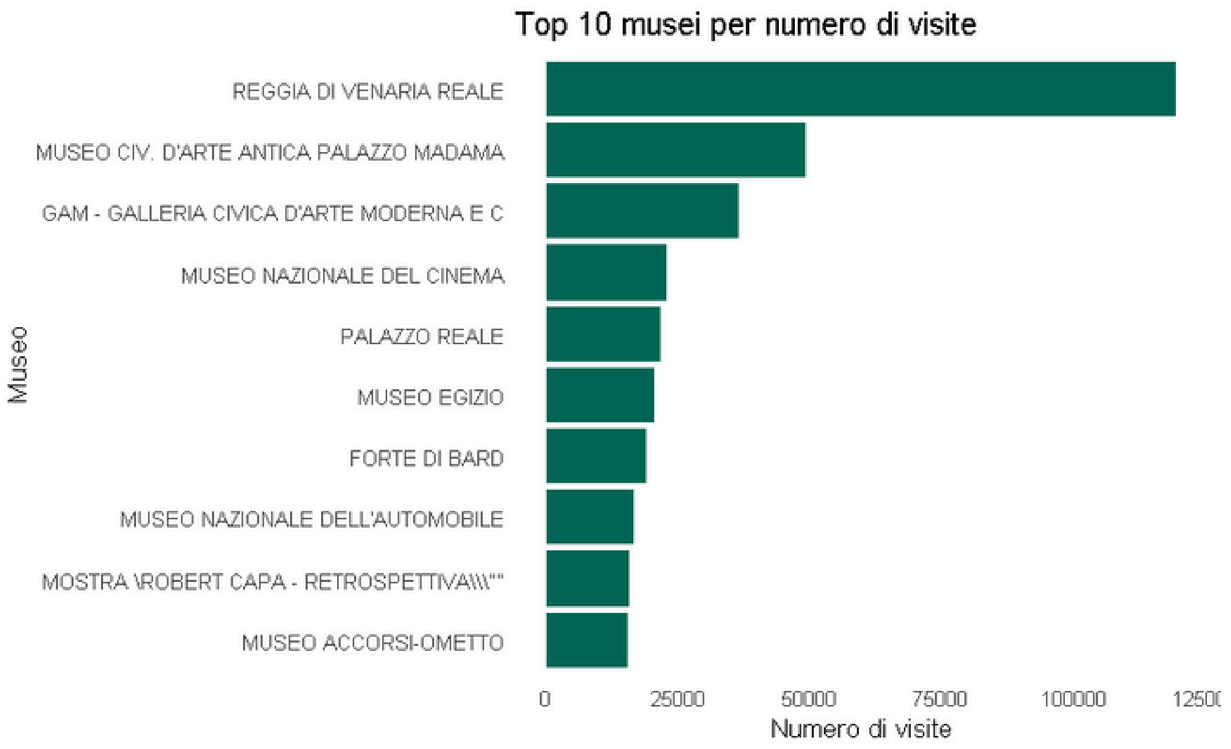
Museum insight

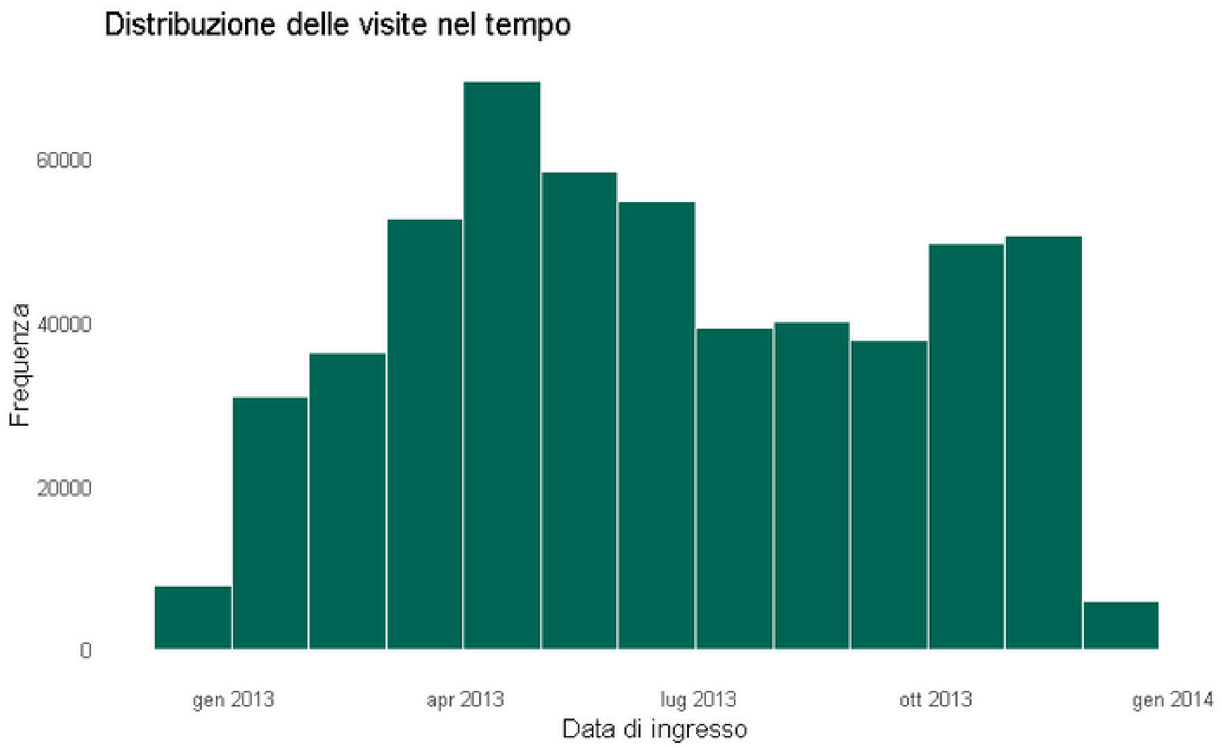
139 unique museums across **64** cities

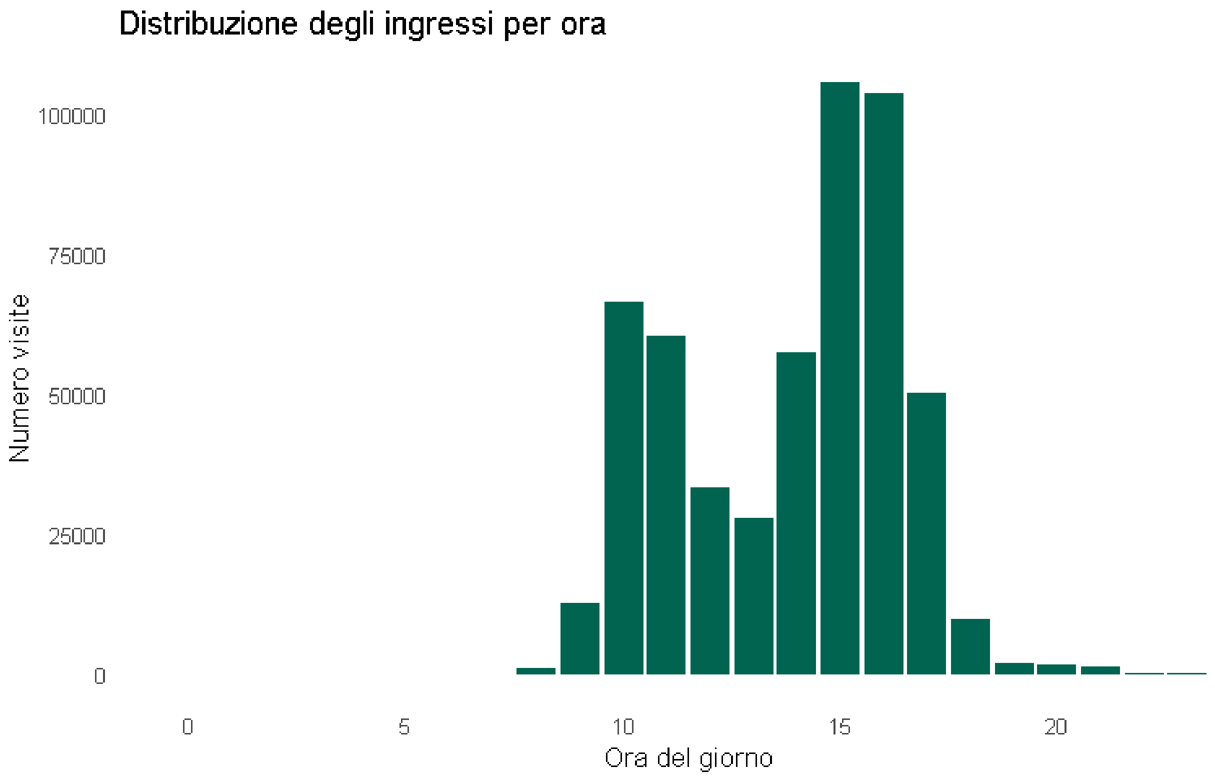
77,846 unique entries

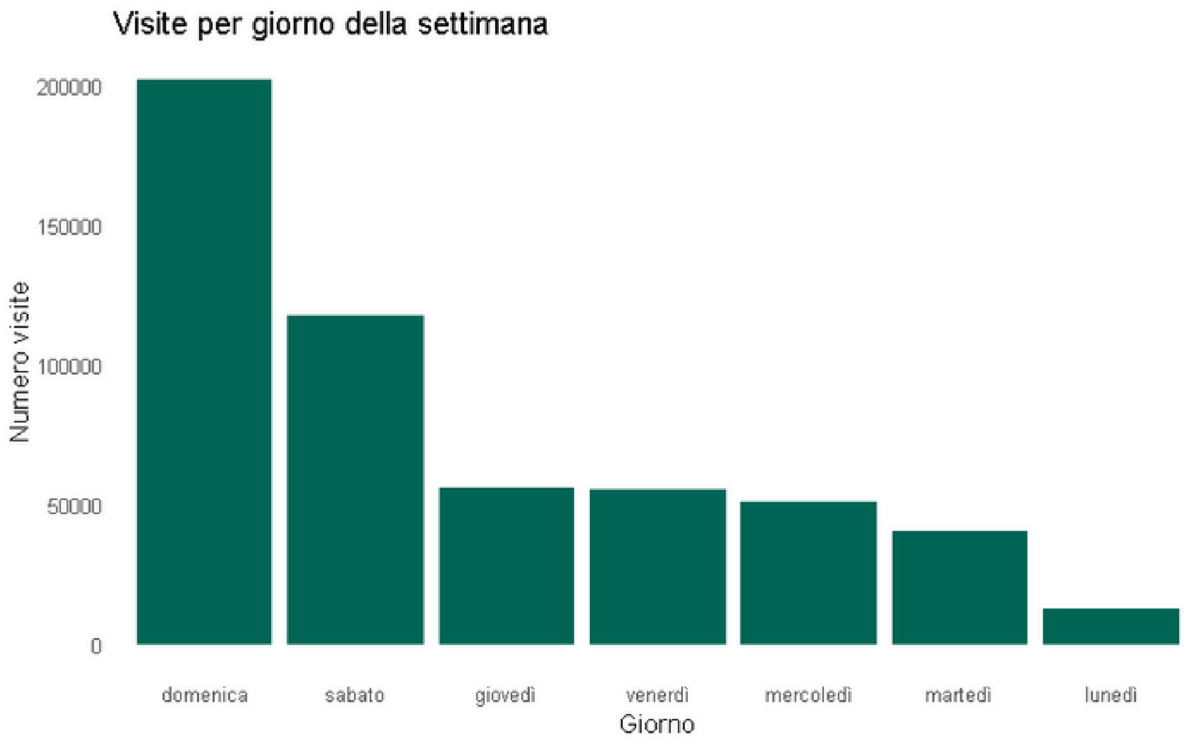
Visits generate a contribution of €0.20 each from the city of Turin

The museum card drives cultural participation with clear peaks during weekends and spring months.
Some museums act as key attractions, concentrating the majority of visits.


Top 10 musei per numero di visite


Distribuzione delle visite nel tempo


Distribuzione degli ingressi per ora


Visite per giorno della settimana

MUSEUM PASS

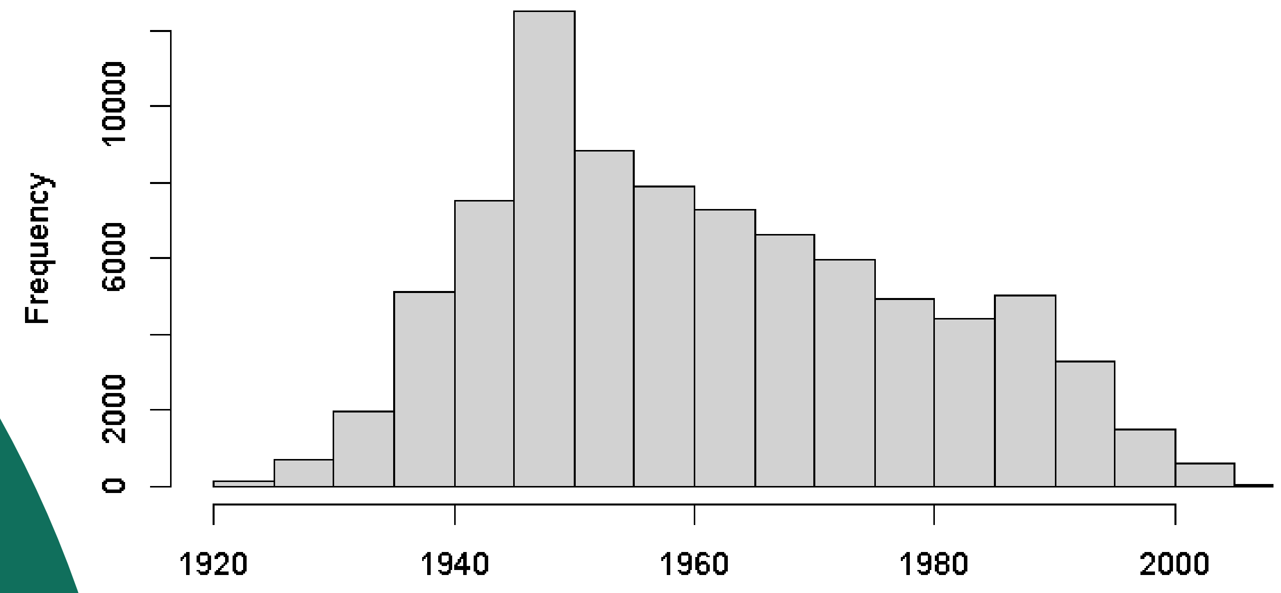
Costumers insight

87052 Costumers
(99% new subscribers)

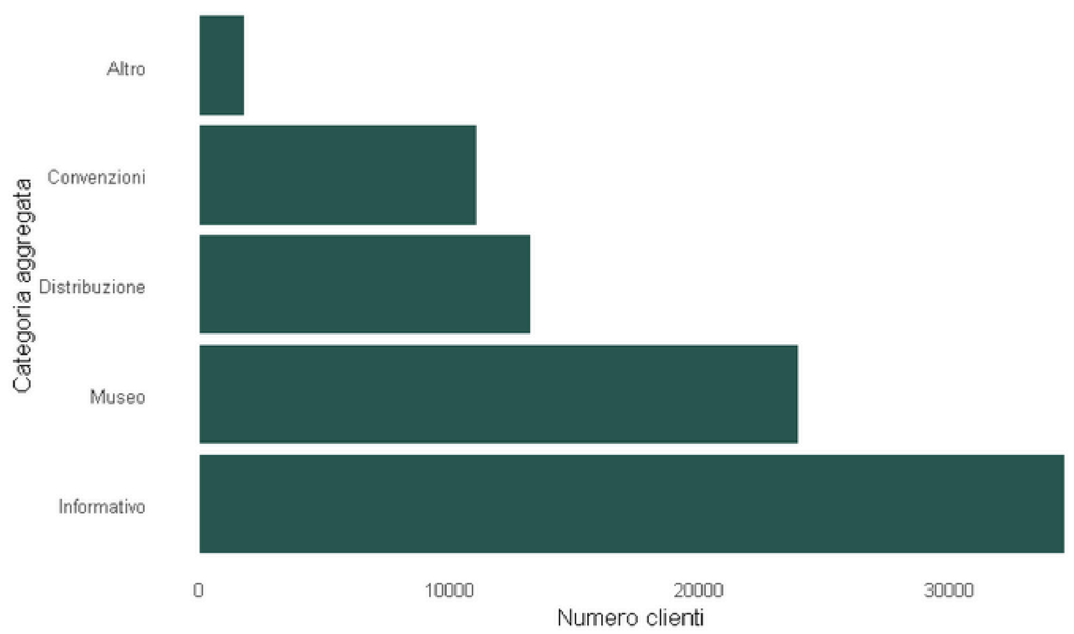
- 47701 Women
- 36732 Men

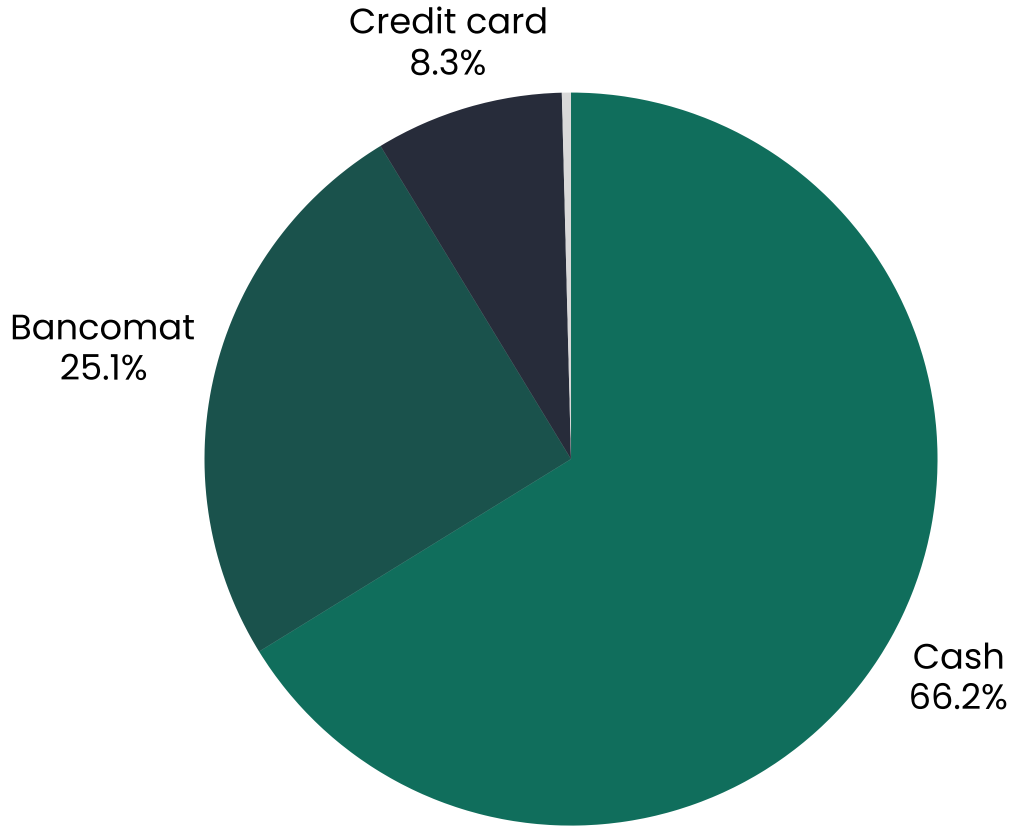
Card buyers' age: mostly between 6 and 93 years (born **1920–2007,** after cleaning))

Majority of subscriptions purchased via “**Informativo**” channels
Prefered payment methods: **cash**

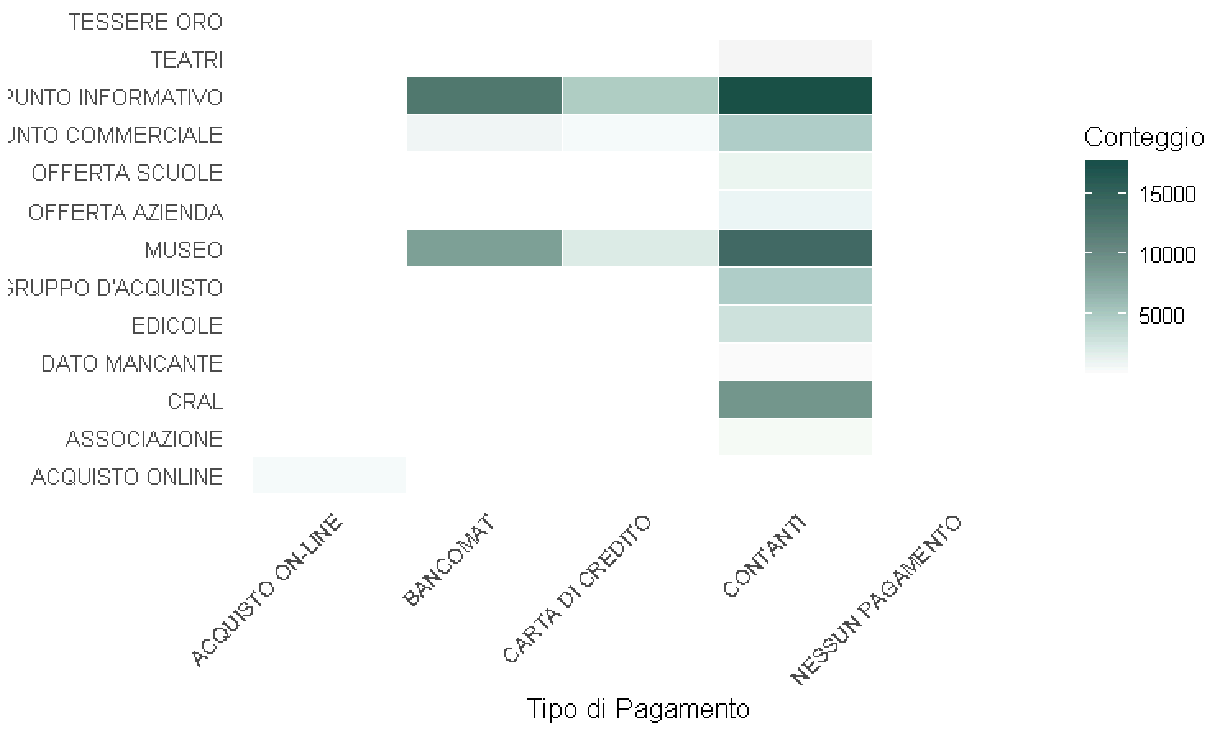


Distribuzione semplificata dei canali di acquisto





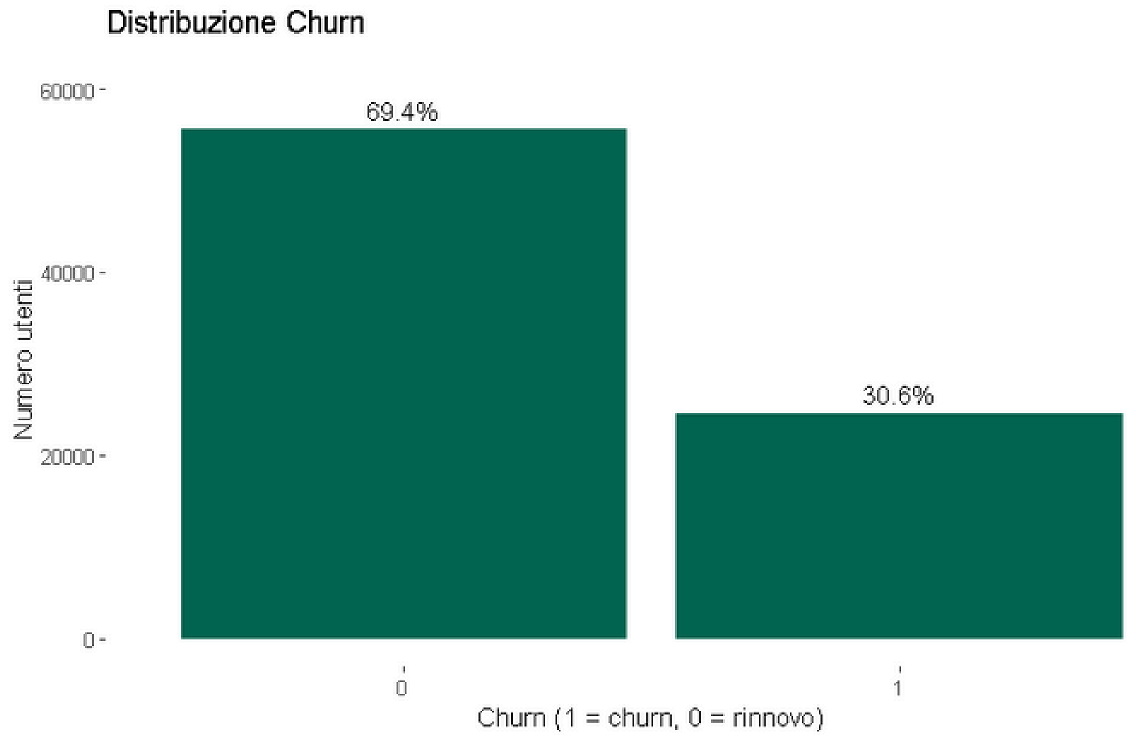
Co-occorrenza tra Tipo Agenzia e Tipo di Pagamento

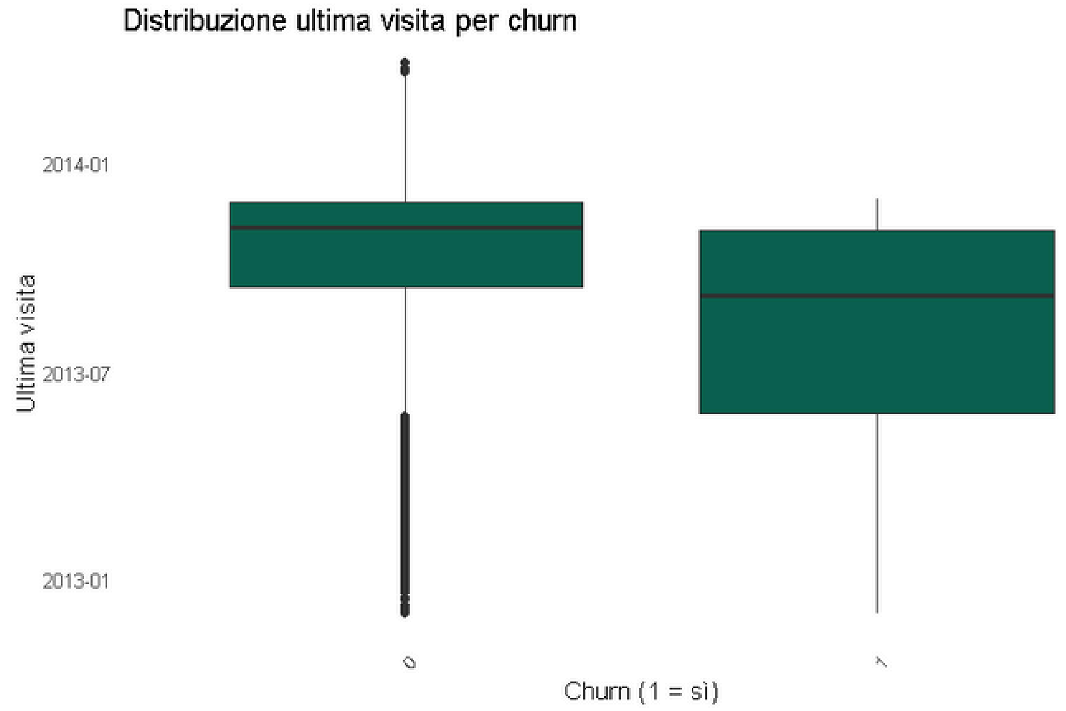


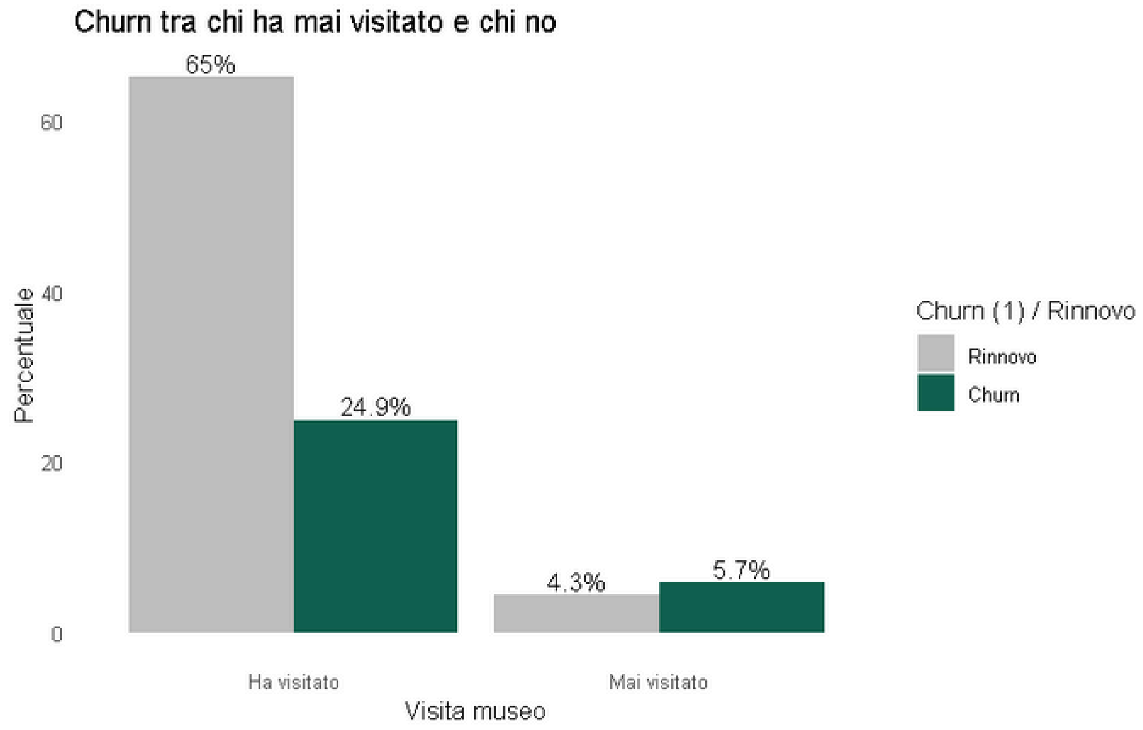
MUSEUM PASS

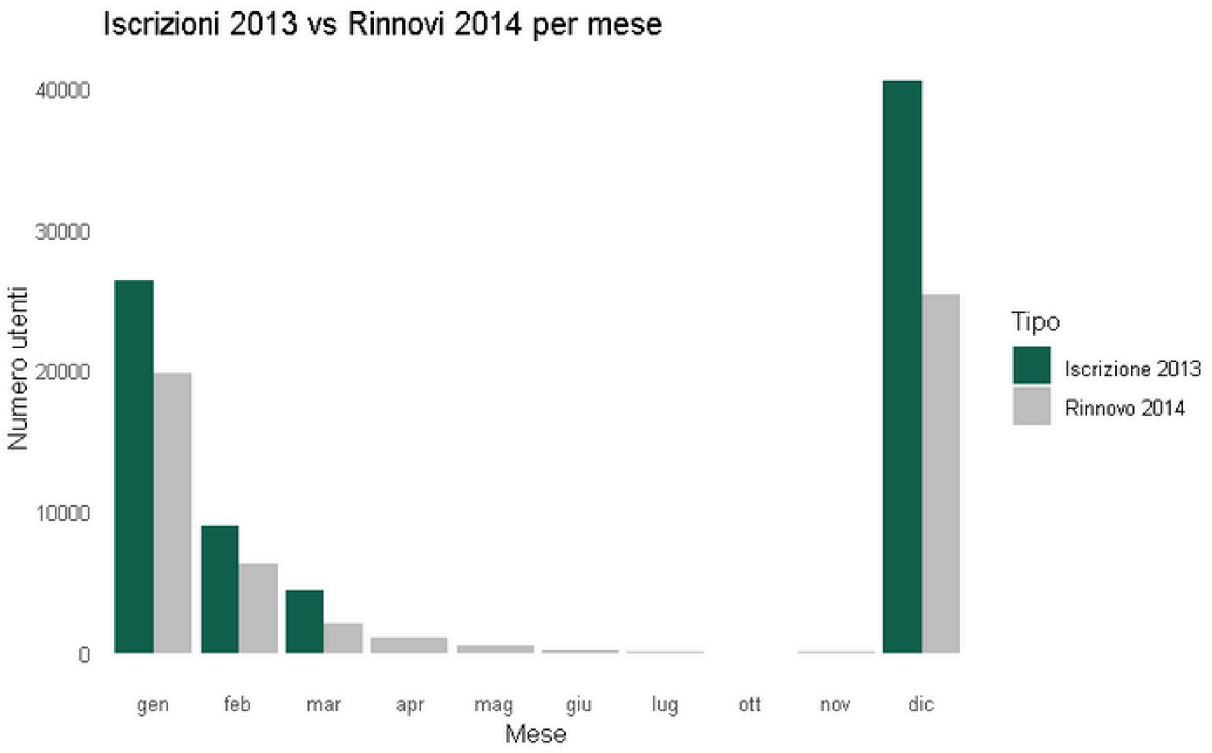
Churner insight

- The churn rate in 30.6%
- Churners generally had their last visit earlier than renewers, indicating lower recent engagement
- The average time between last visit and renewal is longer for churners (88.2 days mean gap).
- A small portion (5.7%) of churners never visited a museum.
- Renewal patterns show that late-year subscribers tend to churn less.


Distribuzione Churn


Distribuzione ultima visita per churn

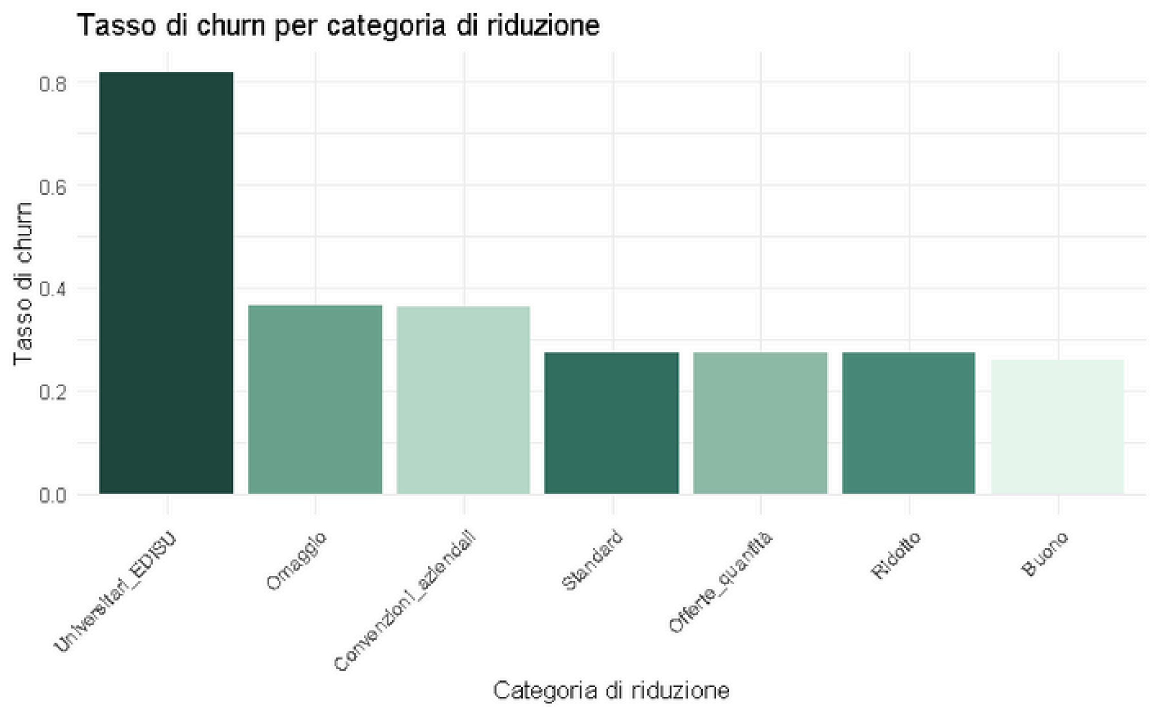

Churn tra chi ha mai visitato e chi no

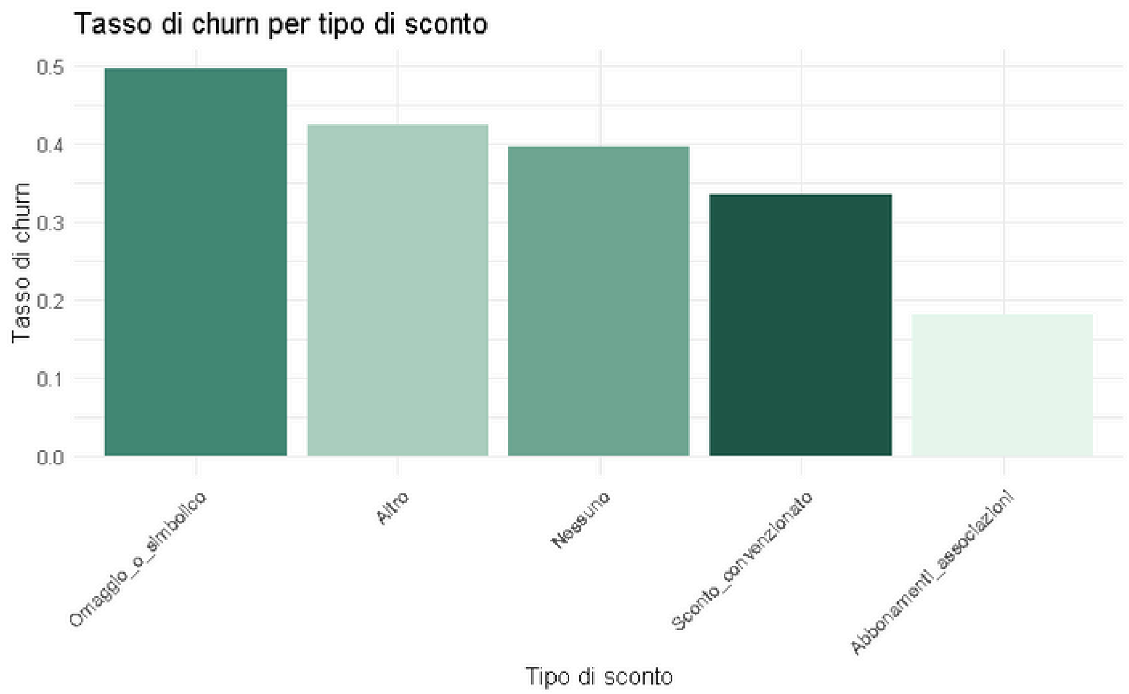

Iscrizioni 2013 vs Rinnovi 2014 per mese

MUSEUM PASS

Churner insight

- Customers with Universitari_EDISU reduction have the highest churn rate (82%).
- Omaggio_o_simbolico discounts are linked to the highest churn among discount types (~50%).

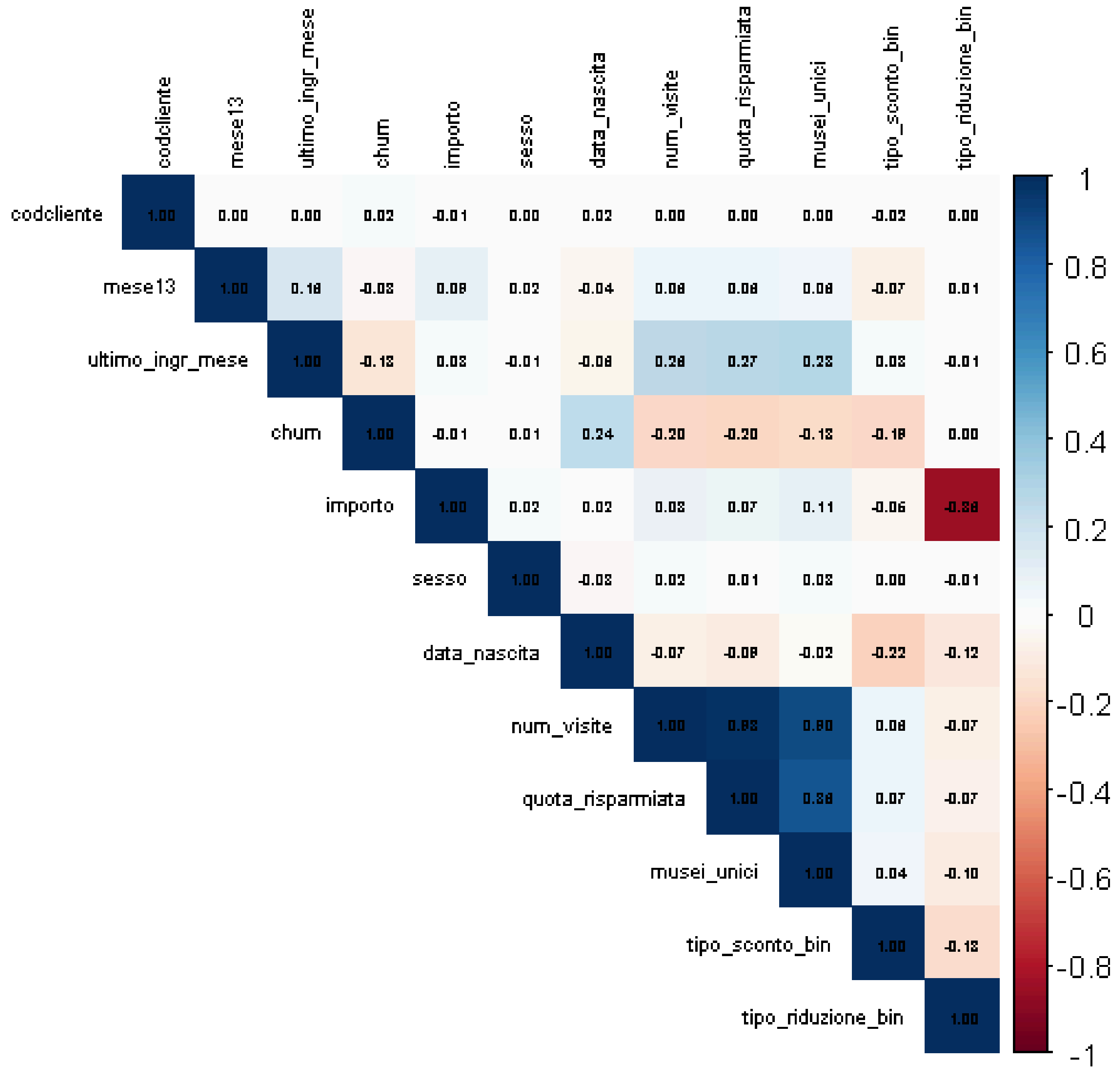

Tasso di churn per categoria di riduzione


Tasso di churn per tipo di sconto

MUSEUM PASS

Correlation analysis

- The churn variable shows low correlation with individual features.
- The highest correlations with churn are:
  - numero visite, quota_risparmiata (-0.20)
  - tipo_sconto_bin (-0.18)



	codcliente	mese13	ultimo_ingr_mese	churn	importo	sesso	data_nascita	num_visite	quota_risparmiata	musei_unici	tipo_sconto_bin	tipo_riduzione_bin
codcliente	1.00	0.00	0.00	0.02	-0.01	0.00	0.02	0.00	0.00	0.00	-0.02	0.00
mese13		1.00	0.18	-0.03	0.08	0.02	-0.04	0.08	0.08	0.08	-0.07	0.01
ultimo_ingr_mese			1.00	-0.13	0.03	-0.01	-0.08	0.28	0.27	0.23	0.03	-0.01
churn				1.00	-0.01	0.01	0.24	-0.20	-0.20	-0.13	-0.18	0.00
importo					1.00	0.02	0.02	0.08	0.07	0.11	-0.06	-0.88
sesso						1.00	-0.03	0.02	0.01	0.03	0.00	-0.01
data_nascita							1.00	-0.07	-0.08	-0.02	-0.22	-0.12
num_visite								1.00	0.93	0.80	0.08	-0.07
quota_risparmiata									1.00	0.38	0.07	-0.07
musei_unici										1.00	0.04	-0.10
tipo_sconto_bin											1.00	-0.13
tipo_riduzione_bin												1.00

Customers who visit more and save more are less likely to churn clear signal that active card usage protects against churn.

MUSEUM PASS

Clustering

SOM structure: **20x20**
Number of cluster: **2**

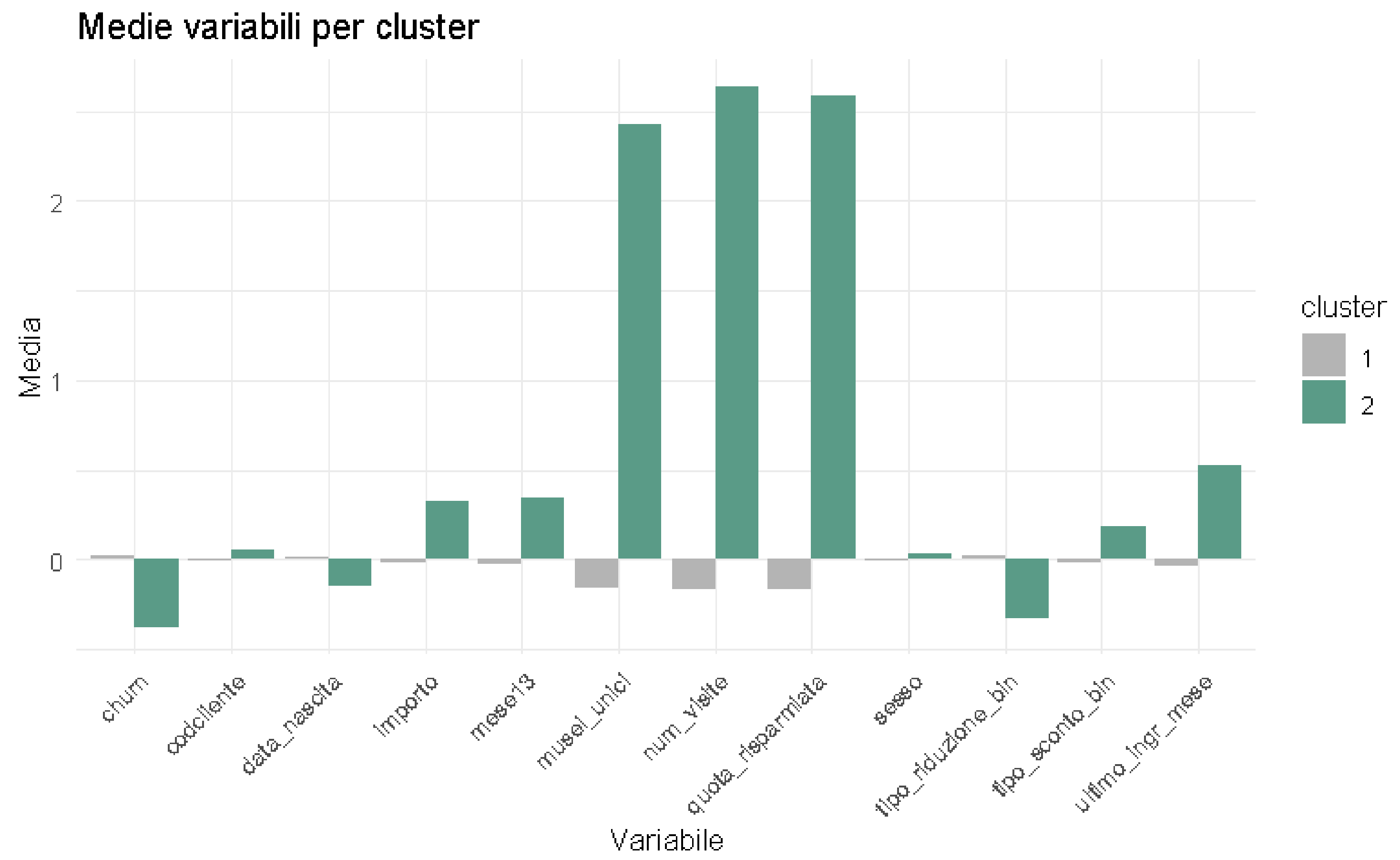
cluster	elements
1	65182
2	4190

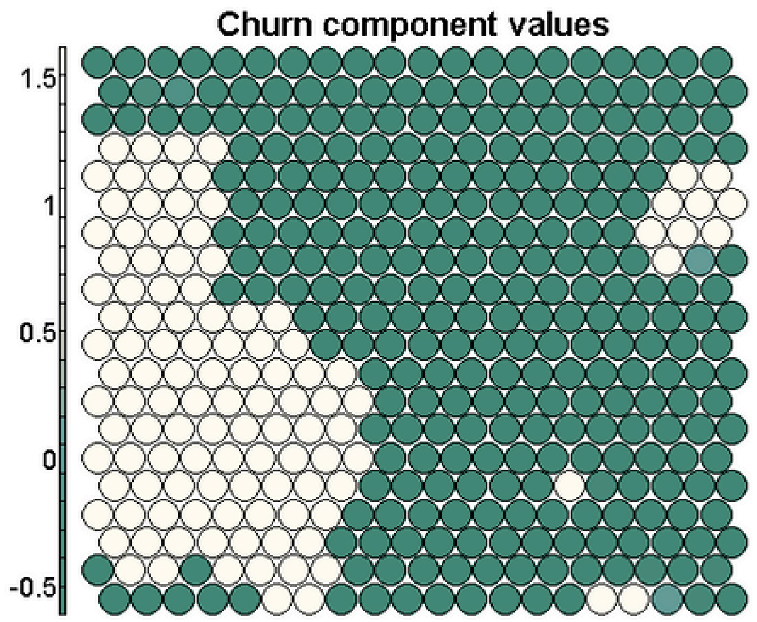
Cluster 1
- Low values for num_visite, quota_risparmiata, and musei_unici → low card engagement.
- Represents a less active segment, more likely to churn.

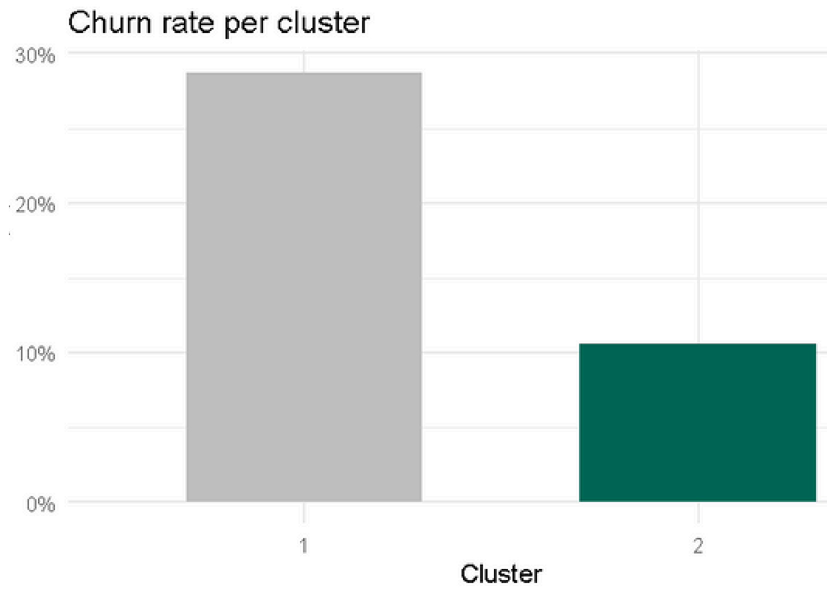
Cluster 2
- High values for num_visite, quota_risparmiata, musei_unici → highly active card users.
- More recent last visit (ultimo_ingr_mese), higher usage → greater retention.

Cluster 1: Churn rate around 29% → these customers are at higher risk of abandoning the card.
Cluster 2: Churn rate around 11% → more loyal customers with lower churn risk.

Business takeaways
- Focus marketing campaigns on Cluster 1: these customers need incentives (e.g. targeted promotions, reminders, usage stimulation).
- Reward Cluster 2: maintain engagement with loyalty programs or exclusive offers to keep them renewing.


Medie variabili per cluster


Churn component values

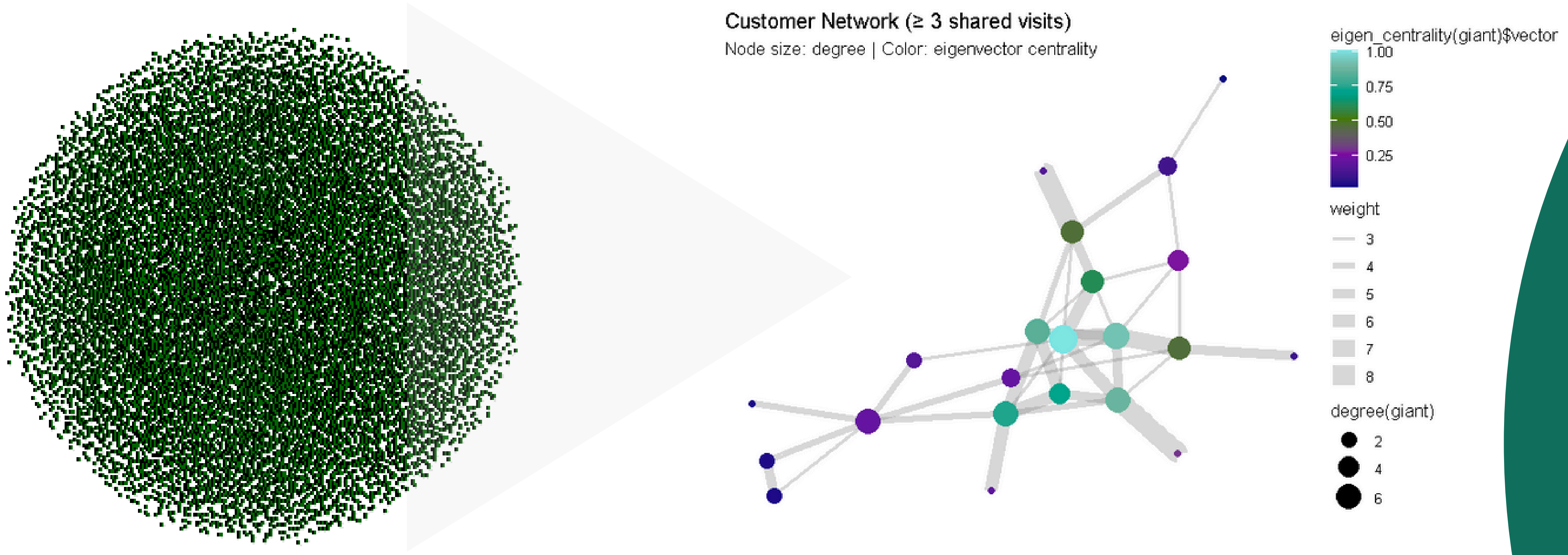

Churn rate per cluster

MUSEUM PASS

Customer Network Analysis

Objective: Build a customer network connecting users who visited the same museum at the same time at least 3 times

Rete clienti con almeno 3 visite condivise



Nodes: Customers (30,337)
Edges: 17,550 edges
Edge weight: Number of shared visits

The 10 most central customers:

- 80% are women
- Born on average between 1975 and 1982
- 90% have Standard reduction
- 90% prefer Bancomat as payment method
- 60% did not apply any discount

The network highlights small clusters of customers with shared cultural interests. Targeting these micro-groups could enhance campaign efficiency.

MUSEUM PASS

Causal analysis

Objective: Assess whether gender has a causal impact on churn probability, controlling for observed characteristics.

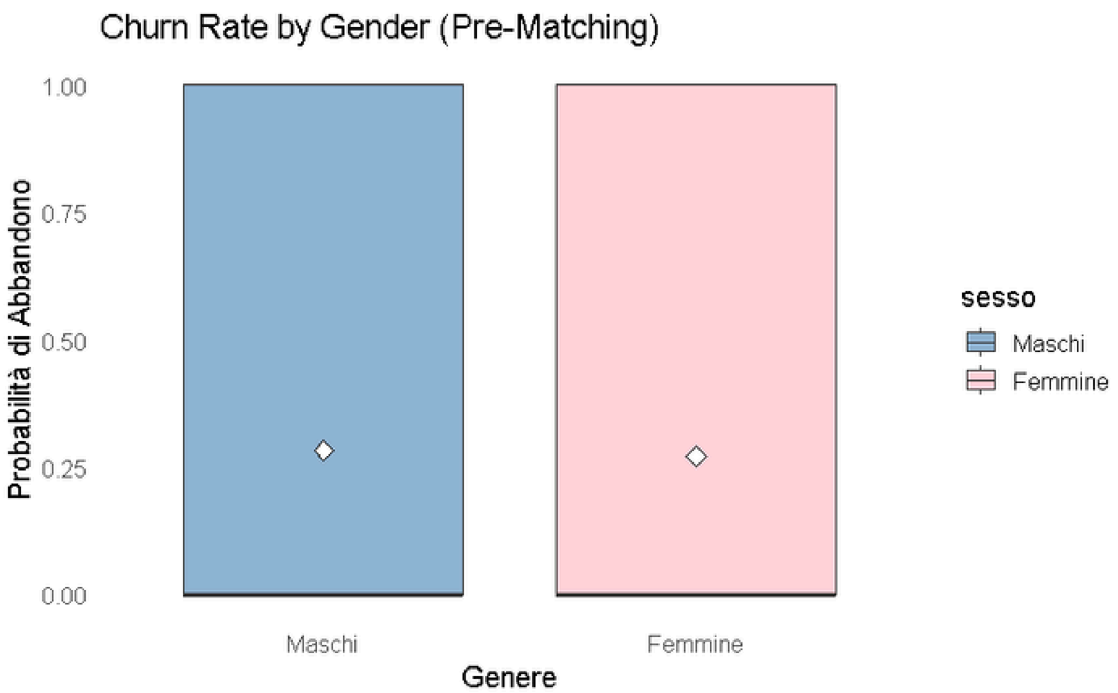
Methodology

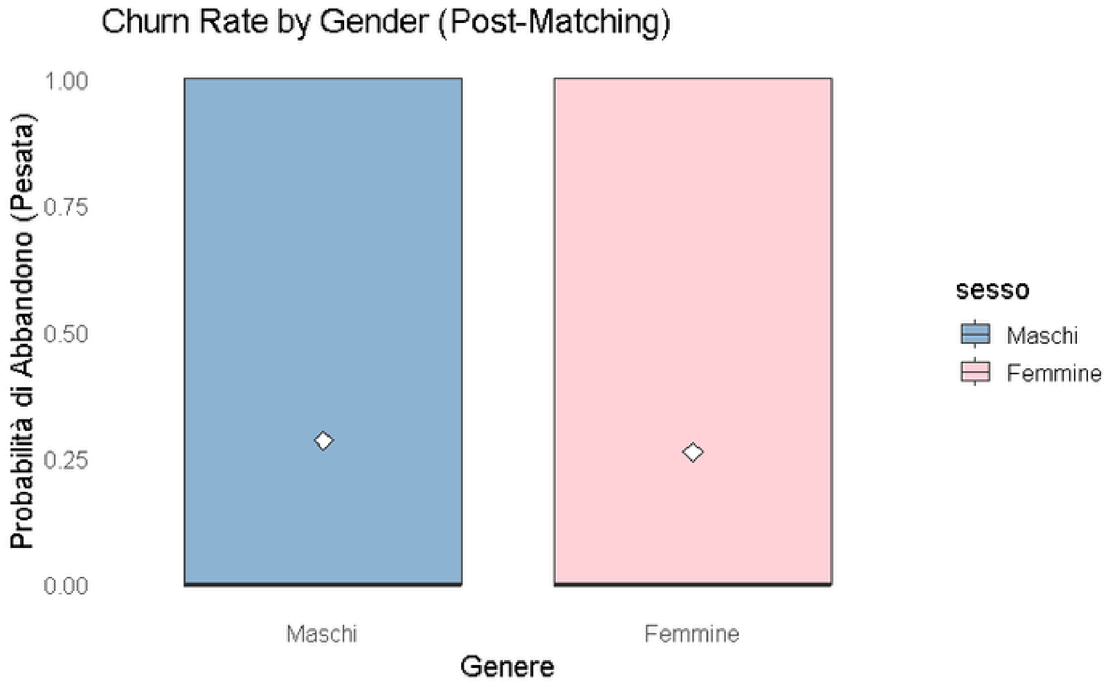
Used Propensity Score Matching to create comparable groups of males and females.
Checked covariate balance post-matching
→ good balance achieved (all differences < 0.1 threshold).

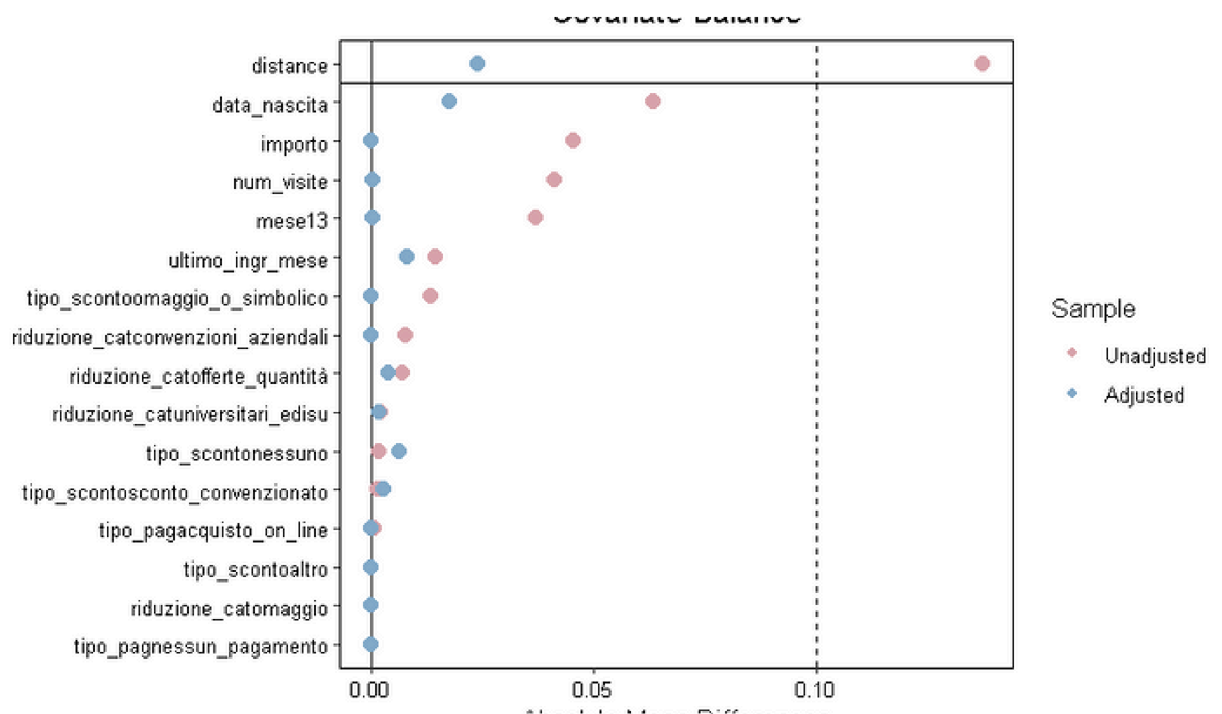
Result

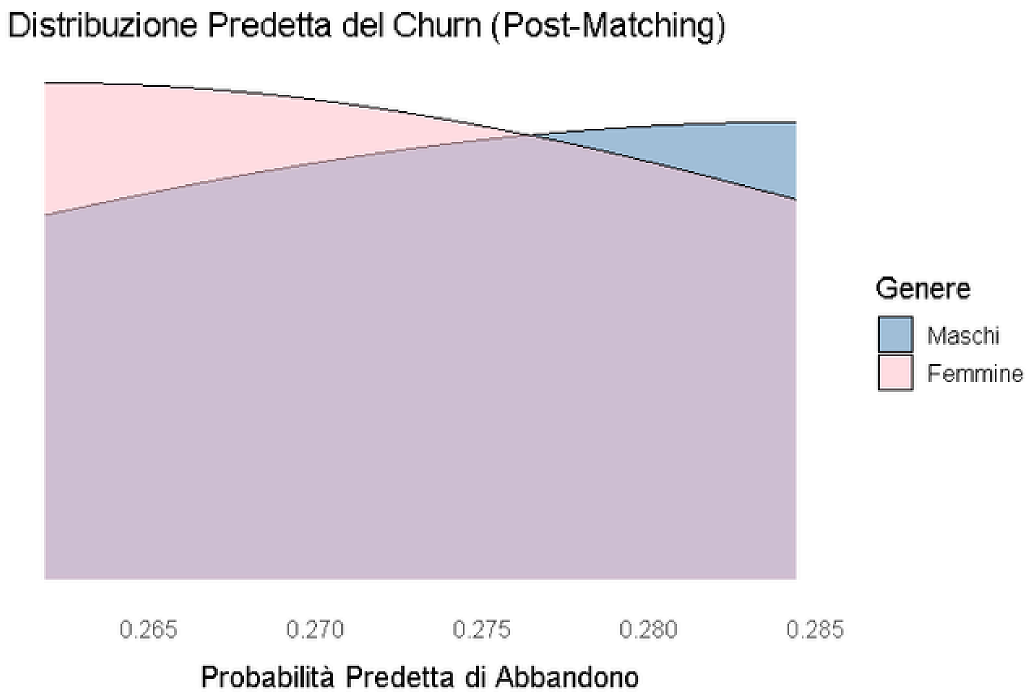
- LPM: Being female → churn probability ↓ ~2% (p < 0.001)
- Logit: Being female → lower churn likelihood (p < 0.001)

Gender shows a small but significant causal effect: women are less likely to churn.









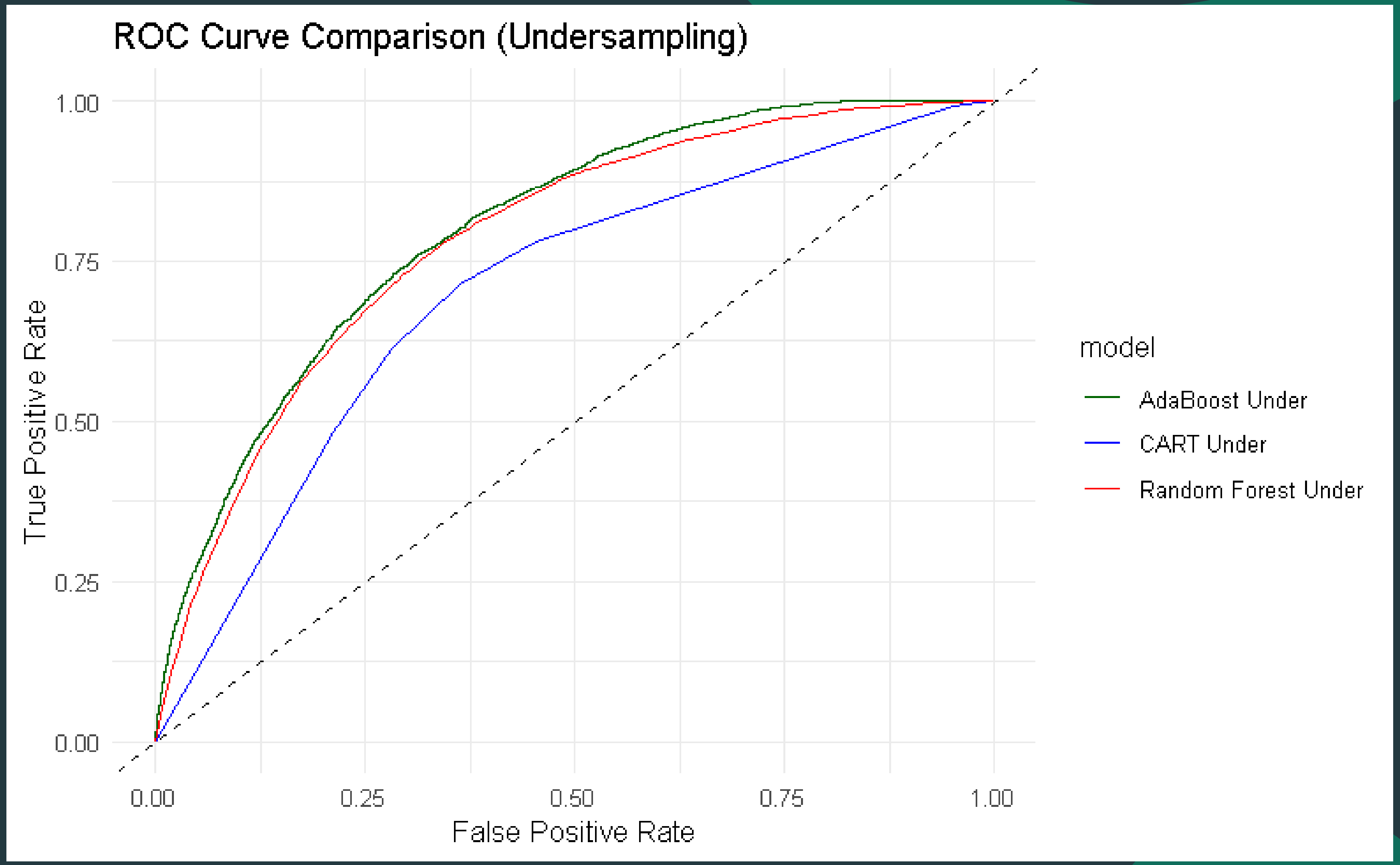
MUSEUM PASS

Churn Prediction

Objective: Build predictive models to identify potential churners and support targeted marketing campaigns.

Model	AUC (Standard)	AUC (Und.Sampl)
CART	64.89%	70.59%
Random Forest	77.67%	78.38%
KNN	73.53%	x
AdaBoost	**79.85%**	**79.71%**

AdaBoost is the preferred model for campaign targeting, given its superior discriminatory power.



THYNK UNLIMITED

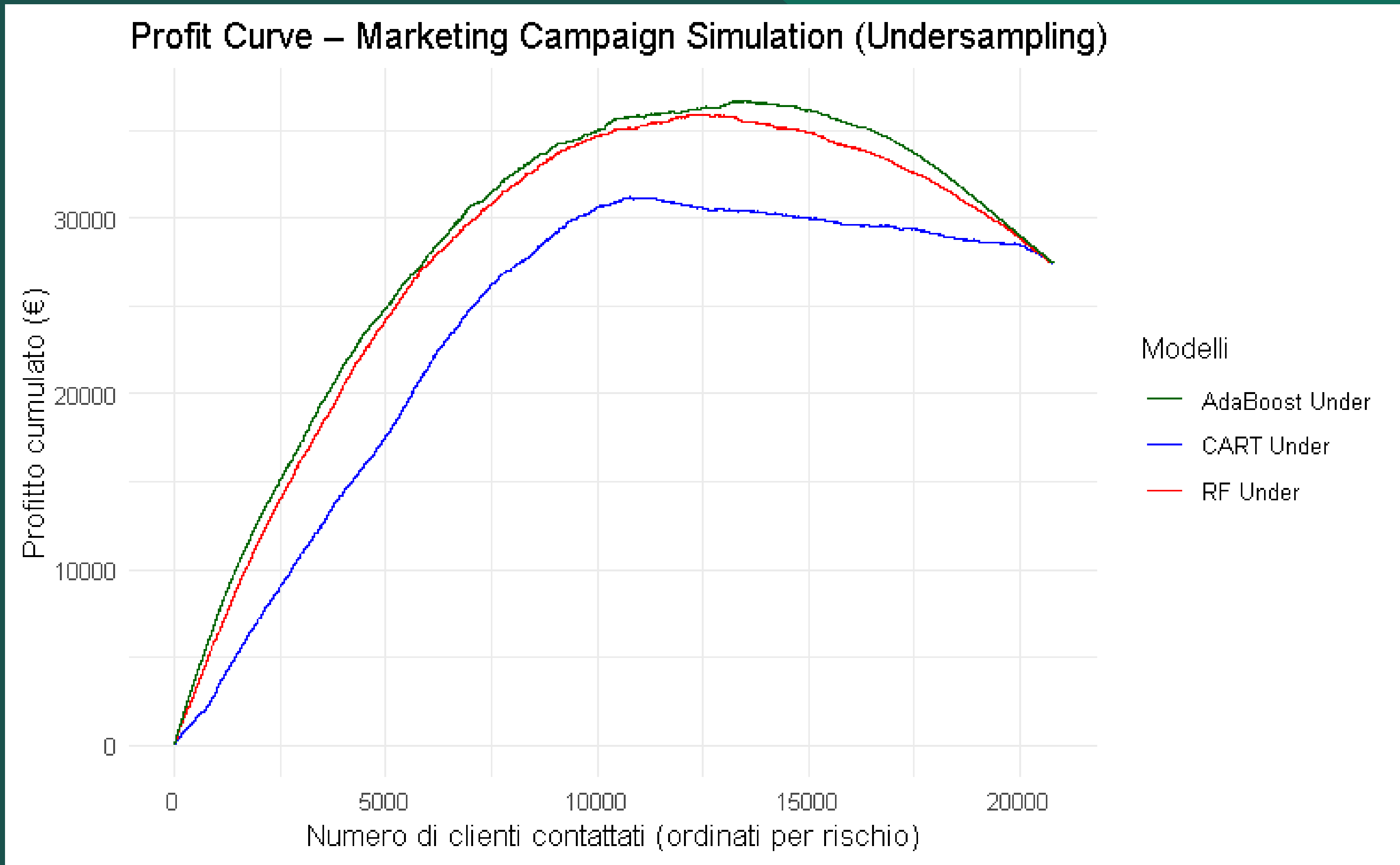
Marketing Campaign

Objective: Evaluate how predictive models support a cost-effective marketing campaign by contacting customers at risk of churn

Scenario:

- Each contacted churner who renews → +10€ net profit
- Each contacted non-churner → -2€ net cost

Model	Max Profit (€)	Optimal
CART Under	31968	13202
RF Under	36200	11207
AdaBoost Under	**36544**	**14188**



AdaBoost yields the highest cumulative profit.
Random Forest Under achieves nearly the same profit but requires fewer contacts, making it more cost-efficient under budget constraints

Recap

CHURN DRIVERS IDENTIFIED

Low card engagement (fewer visits, low savings, few unique museums) is a strong predictor of churn.

CLUSTERING

SOM clustering distinguished loyal vs. at-risk customers → marketing can target low-engagement cluster.

CASUAL INSIGHT

Gender has a small but significant impact, women are less likely to churn.

PREDICTIVE MODEL

:

AdaBoost delivers the highest cumulative profit, but Random Forest achieves similar profit with fewer contacts (cost-efficient for tighter budgets).

Business recommendations:

- Focus retention efforts on low-engagement customers through tailored promotions.
- Maintain loyalty of high-engagement customers with exclusive offers.
- Leverage network micro-groups to design group-based campaigns.
- Prioritize cost-effective models like Random Forest when budget is limited.