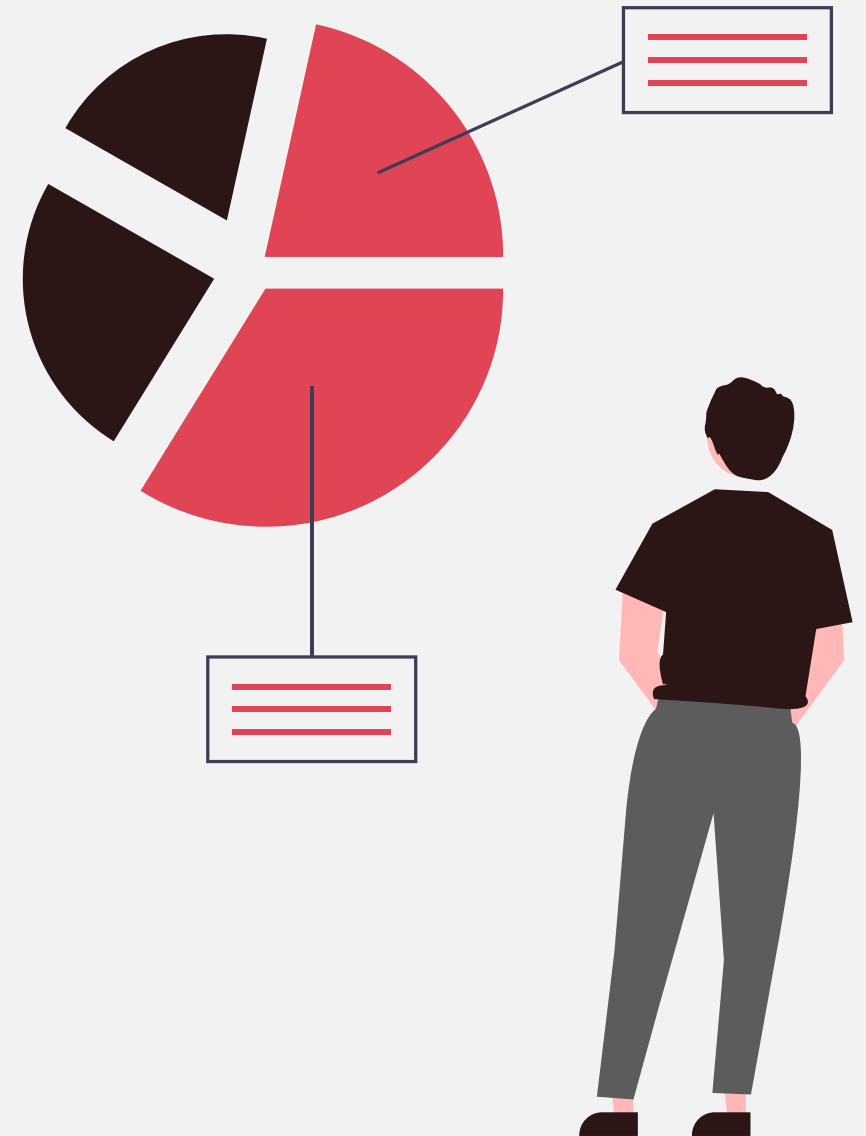


DIGITAL MARKETING

Costruzione di una strategia data driven per il mantenimento della clientela ad alto valore

Alfredo Galli & Riccardo Rubini
A.A 2020/2021



OBIETTIVO DELL'ANALISI: LA CUSTOMER RETENTION

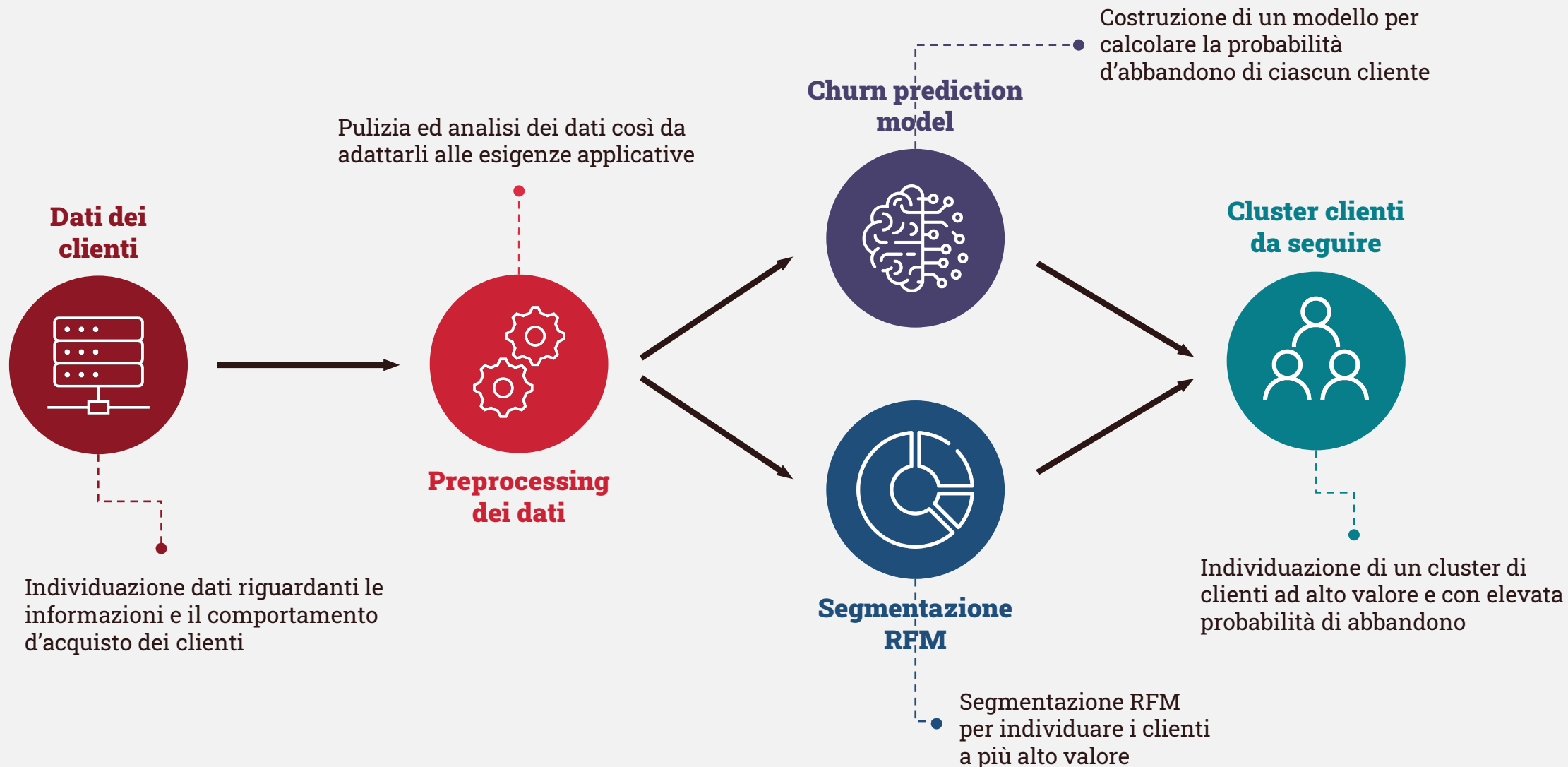
Il mantenimento della clientela già acquisita è una delle principali sfide per ogni azienda, perché ha costi inferiori rispetto a ottenere nuovi clienti e benefici maggiori in termini di guadagno. È infatti più semplice vendere ad un cliente già fidelizzato rispetto ad uno nuovo.

In particolar modo l'aspetto veramente importante è cercare di mantenere con sé quei clienti che nel corso del tempo hanno dimostrato di avere un alto valore per il business aziendale, ovvero coloro i quali hanno acquistato molto e spesso.

Per questa ragione si è deciso di porre come obiettivo di questa analisi la costruzione di una strategia data driven atta al mantenimento della clientela ad alto valore, individuando un cluster ideale di consumatori sul quale incentrare le forze allo scopo di mantenerli.



WORKFLOW ANALISI PROCESSO GENERALE



CUSTOMER SEGMENTATION



DIVISIONE CLIENTI

CON SEGMENTAZIONE RFM

Per dividere i clienti in maniera tale da stabilire quali fossero quelli a più alto valore, si è scelto di affidarsi all'analisi RFM. Questa tecnica permette di segmentare la clientela tramite lo storico delle loro transizioni, in base a quando e quanto hanno acquistato. Nella pratica ad ogni cliente viene assegnato un punteggio da 1 a 3 per i valori di recency, frequency e monetary del suo storico.

I clienti sul quale si è eseguita questa analisi sono stati quelli risultanti attivi al 30 aprile 2019, ovvero coloro nel trimestre precedente (febbraio-aprile 2019) avevano effettuato almeno un acquisto.






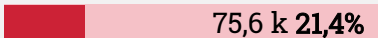


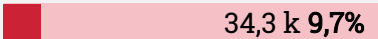












Per quanto riguarda lo storico delle transazioni si è scelto di utilizzare come range temporale il semestre precedente alla data di riferimento scelta, dunque il periodo novembre 2018 – aprile 2019.

Basandosi sui punteggi RFM ottenuti, la clientela è stata divisa in sette diversi gruppi: diamond, gold, silver, bronze, copper, tin e cheap.

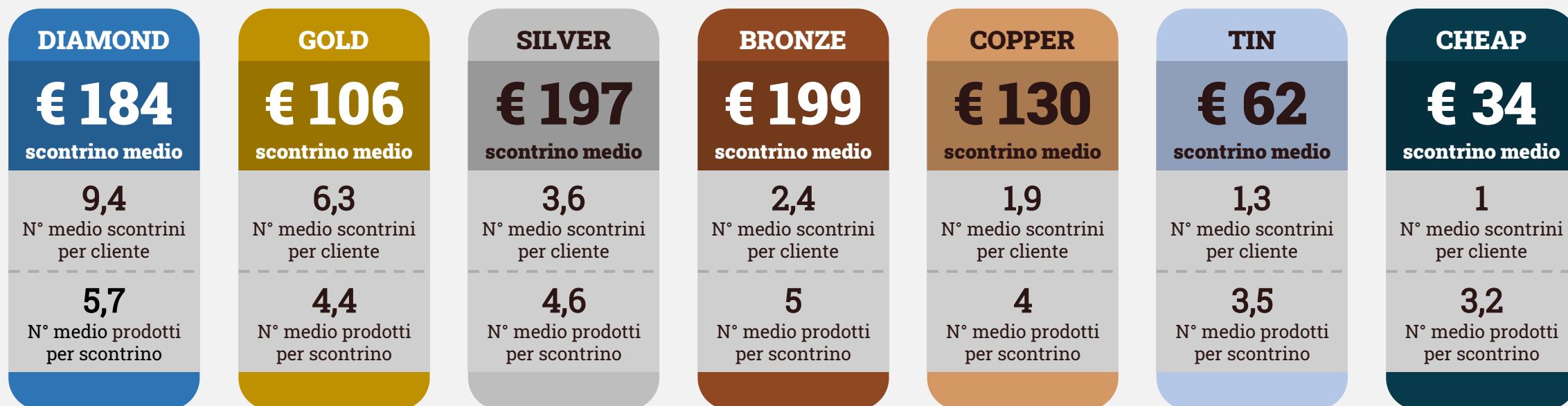
I clienti appartenenti al gruppo diamond sono quelli con più alto valore per l'azienda, poiché spendono di più di tutti, molto frequentemente e con pochi giorni intercorsi tra la data di riferimento e l'ultimo loro acquisto. A seguire vi sono i gruppi gold e silver, composti da clienti con valore attuale/potenziale medio-alto, con alcuni dei quali però che non acquistano da parecchi giorni. Infine i restanti cluster presentano consumatori che per un motivo o per l'altro non risultano così determinanti per il business aziendale.

I clienti considerati ad alto valore sono quelli appartenenti ai gruppi diamond, gold e silver.

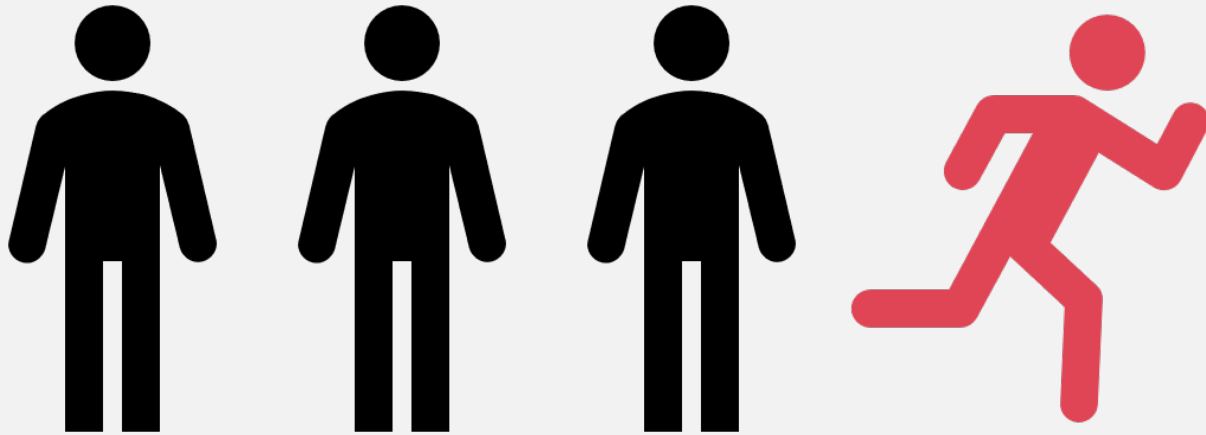
DIVISIONE DEI CLIENTI CON SEGMENTAZIONE RFM

Classe	Descrizione	Tot Clienti	Tot Spesa	Tot Scontrini
DIAMOND	Spendono complessivamente più di tutti e acquistano molto spesso	 15,5 k 16%	 € 26,7 mln 49,6%	 145 k 41,1%
GOLD	Spendono abbastanza e acquistano frequentemente	 11,9 k 12%	 € 8 mln 14,9%	 75,6 k 21,4%
SILVER	Spendono parecchio ma non molto frequentemente	 9,5 k 10%	 € 6,8 mln 12,6%	 34,3 k 9,7%
BRONZE	Spendono parecchio ma di rado	 13,4 k 14%	 € 6,4 mln 11,9%	 32,2 k 9,1%
COPPER	Spendono abbastanza ma acquistano raramente	 16,6 k 17%	 € 4,1 mln 7,6%	 31,2 k 8,9%
TIN	Spendono poco e acquistano raramente	 17,7 k 18%	 € 1,4 mln 2,7%	 2,3 k 6,5%
CHEAP	Spendono poco e acquistano generalmente una volta sola	 13,3 k 13%	 € 0,4 mln 0,7%	 1,1 k 3,3%

DIVISIONE DEI CLIENTI CON SEGMENTAZIONE RFM

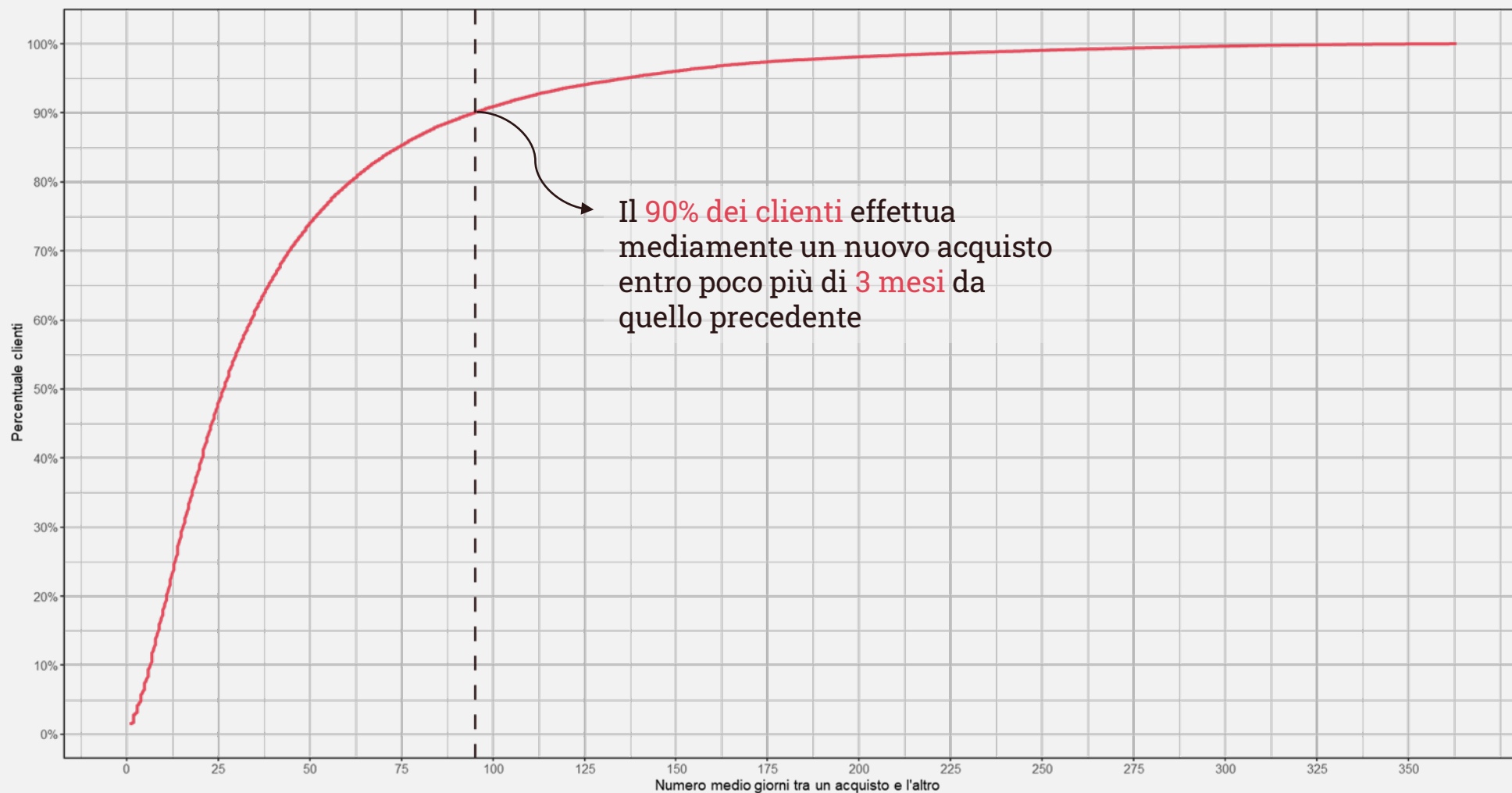


CHURN PREDICTION MODEL



PURCHASE TIME SCALE

Il grafico mostra l'andamento cumulato del numero di giorni entro i quali un cliente effettua un nuovo acquisto nei negozi dell'azienda



calcolato nel periodo maggio 2018- aprile 2019

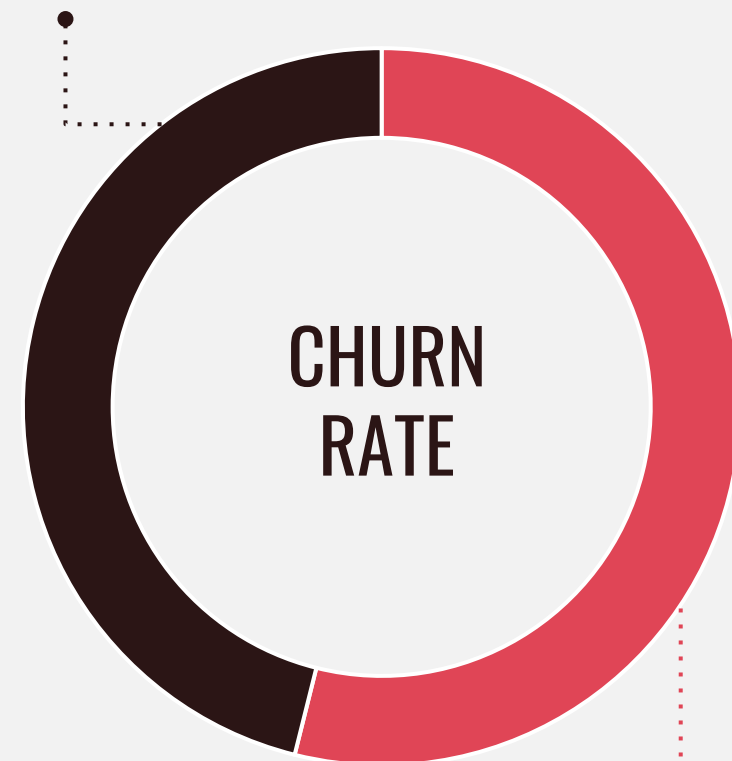
IDENTIFICAZIONE DEI CHURNER

Nella costruzione del dataset di allenamento per il modello di churn prediction, si sono considerati churner quei clienti che a una certa data di riferimento non hanno acquistato nei tre mesi successivi. La scelta di questa scala di tempo pone le basi sul calcolo del purchase time scale visto in precedenza.

Nel nostro specifico caso si è scelta come data di riferimento il 1 febbraio 2019 e si sono considerati nell'analisi solo i clienti che avessero effettuato almeno un acquisto nel trimestre precedente, ovvero da novembre 2018 a gennaio 2019. In totale i consumatori considerati sono stati 95'248.

I dati hanno evidenziato come 51'347 di coloro che avevano acquistato nel trimestre gennaio-novembre non ha più acquistato nei tre mesi successivi, ponendo il churn rate ad un valore pari al 53.9%. Nell'ultimo trimestre l'azienda è comunque riuscita a mantenere 43'901 clienti. Considerando solo i clienti top, il churn rate si abbassa a quota 36%.

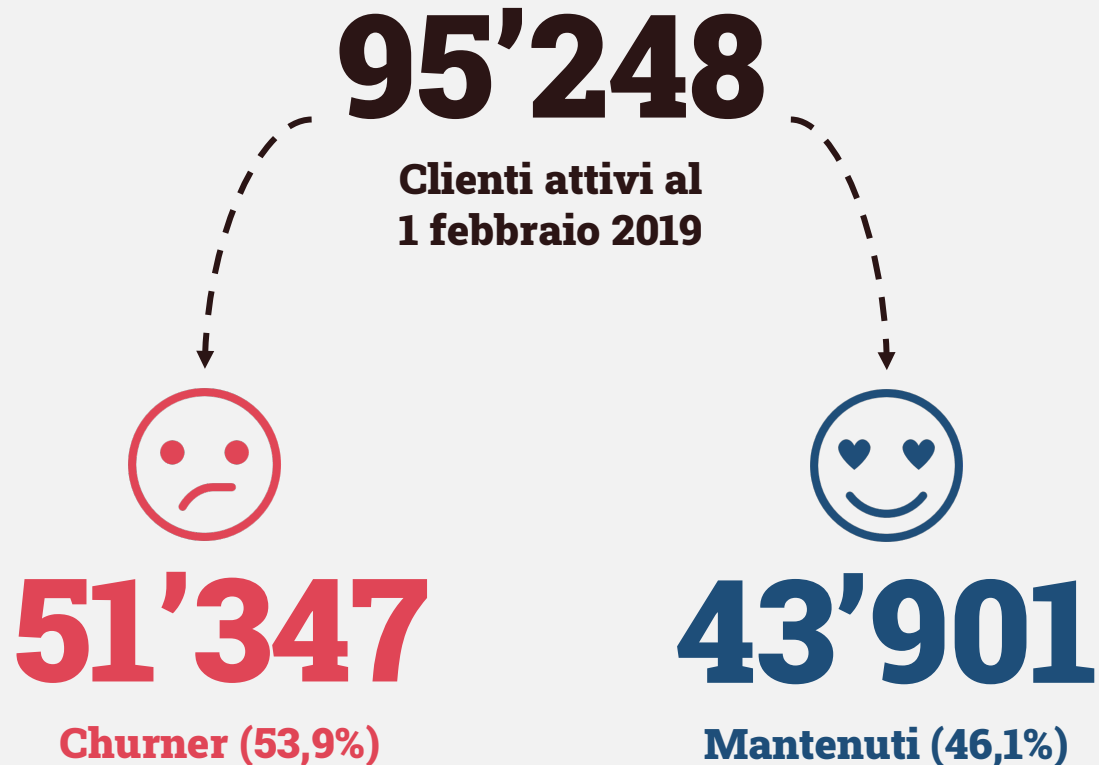
Il 46.1% dei clienti attivi nel trimestre novembre-gennaio ha acquistato tra febbraio e aprile



Il 53,9% dei clienti attivi nel trimestre novembre-gennaio **NON** ha più acquistato tra febbraio e aprile

STRUTTURA DATASET PER ALLENAMENTO MODELLO

UNITÀ STATISTICHE



VARIABILI



Comportamento d'acquisto:

Totale prodotti acquistati/resi; totale importo speso/reso; % prodotti/spesa per singolo reparto ecc...



Informazioni generali:

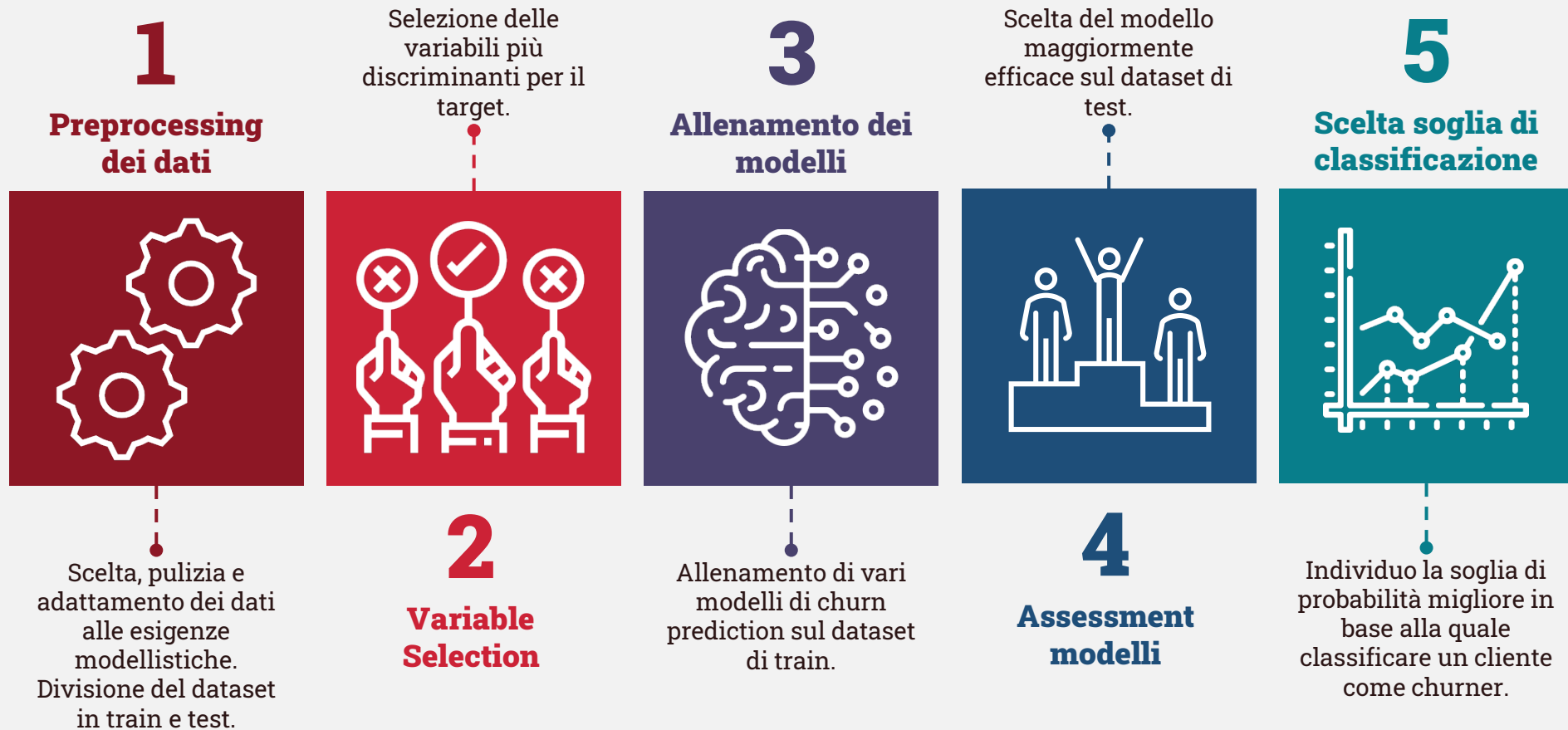
Lavoro, mail provider, programma di fidelizzazione, negozio di riferimento, preferenze privacy e marketing ecc...



Status abbandono cliente:

Variabile binaria 1/0 che indica se il cliente ha o meno abbandonato l'azienda, dunque se ha o meno fatto churn

WORKFLOW ANALISI CHURN PREDICTION MODEL



FEATURE SELECTION & TRAINING DEI MODELLI

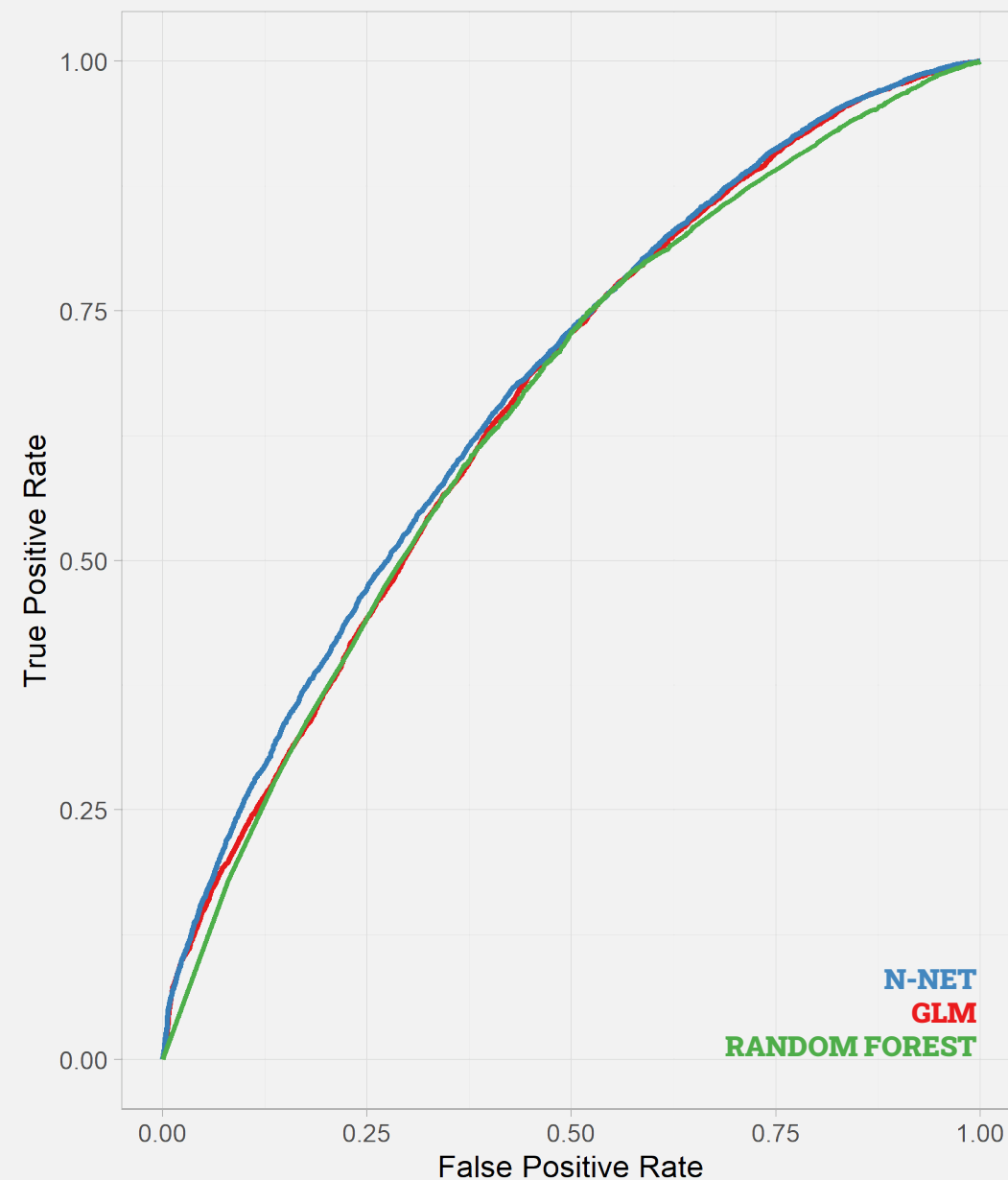
In seguito al preprocessing dei dati (brevemente descritto nella slide precedente), si è proceduto con la selezione del sottoinsieme (auspicabilmente) ottimale delle covariate di partenza sul quale eseguire l'addestramento del set di modelli scelti. Tale step viene indicato usualmente col nome di *feature selection*. Esso è stato eseguito applicando dapprima un albero di classificazione, per poi mantenere quelle variabili predittive che hanno avuto maggior importanza nella creazione dell'albero. A seguito di questa fase è stata poi effettuata un'ulteriore scrematura andando ad omettere quelle variabili affette da *near zero variance*.

Una volta effettuata la *feature selection*, nella fase di training si è scelto addestrare tre tipologie di modelli: MLP, Random Forest e Logistic. I primi due modelli, disponendo di iper-parametri, sono stati sottoposti alla cosiddetta fase di *tuning* per selezionare la miglior «versione» (tra quelle considerate), sulla base della loro performance in termini di *Accuracy*.

ASSESSMENT DEI MODELLI

In seguito all'addestramento, i tre modelli sono stati valutati congiuntamente per eleggere il «miglior modello» da utilizzare per la classificazione dei clienti. La metrica principe in questa fase (detta di *assessment*) è la curva ROC.

Come si può osservare nel grafico a destra, la curva ROC in azzurro associata al MLP (rete neurale) domina, seppur leggermente, le restanti curve; per tale ragione, dunque, tale modello è eleggibile come «miglior modello» da adottare per la classificazione dei clienti.

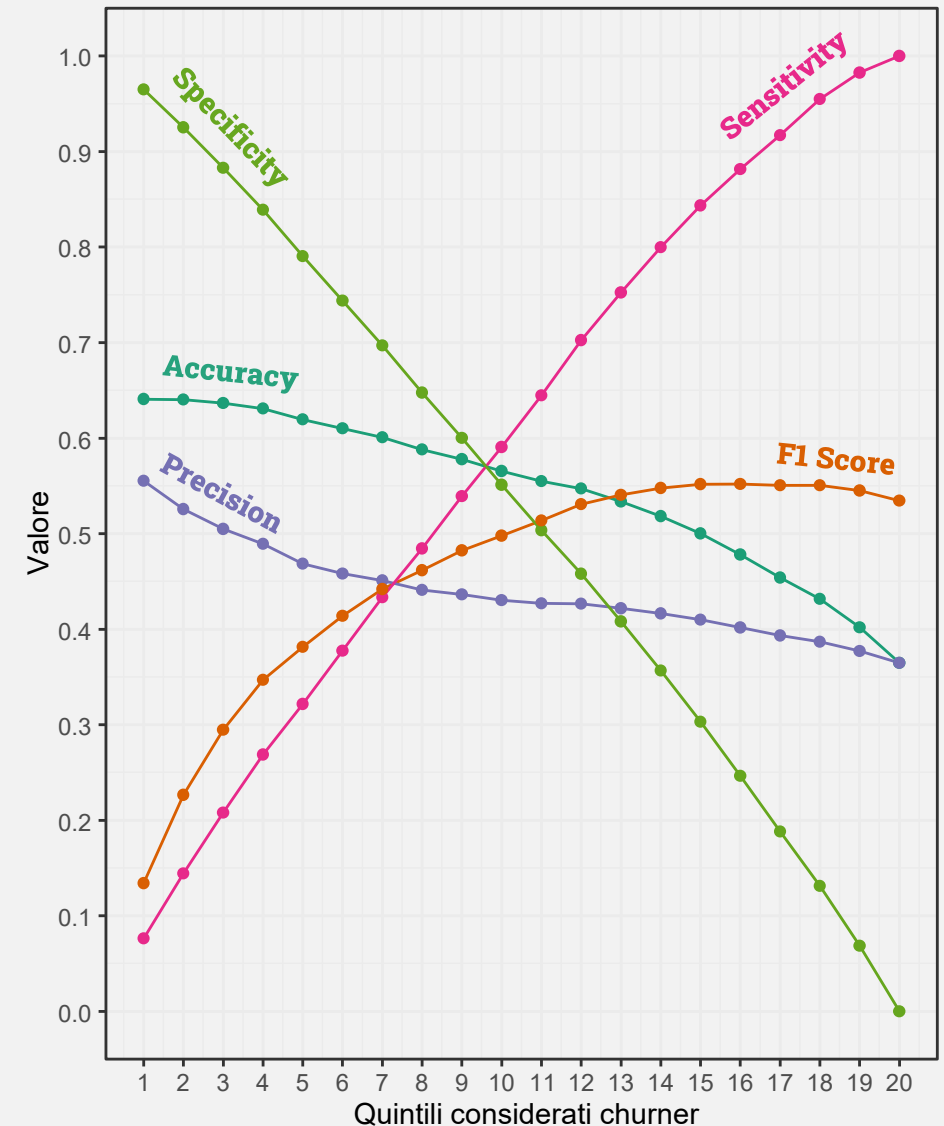


SCELTA DEL QUANTILE SOGLIA

La classificazione dei clienti è basata sui percentili (o meglio ventili) relativi alla probabilità di churn ordinata in senso decrescente, quindi il 1° ventile contiene il 5% dei clienti con probabilità di abbandono prevista più alta.

In pratica, la classificazione avviene progressivamente in questo modo: le osservazioni appartenenti ai primi n ventili vengono classificate come churmer, con $n=1,...,20$. Al variare del «ventile soglia» vengono calcolate le metriche di performance del modello classiche (accuracy, sensitivity...).

Dall'analisi del grafico a destra si è ritenuta una buona strategia classificare i churmer appartenenti ai primi 10 ventili (in pratica il primo 50%). Questo perché in quel punto tutte le metriche assumevano valori considerati soddisfacenti in relazione all'obiettivo d'analisi.



DIVISIONE CLIENTI

CON SEGMENTAZIONE RFM

Per dividere i clienti in maniera tale da stabilire quali fossero quelli a più alto valore, si è scelto di affidarsi all'analisi RFM. Questa tecnica permette di segmentare la clientela tramite lo storico delle loro transizioni, in base a quando e quanto hanno acquistato. Nella pratica ad ogni cliente viene assegnato un punteggio da 1 a 3 per i valori di recency, frequency e monetary del suo storico.

I clienti sul quale si è eseguita questa analisi sono stati quelli risultanti attivi al 30 aprile 2019, ovvero coloro nel trimestre precedente (febbraio-aprile 2019) avevano effettuato almeno un acquisto.

Per quanto riguarda lo storico delle transazioni si è scelto di utilizzare come range temporale il semestre precedente alla data di riferimento scelta, dunque il periodo novembre 2018 – aprile 2019.

Basandosi sui punteggi RFM ottenuti, la clientela è stata divisa in sette diversi gruppi: diamond, gold, silver, bronze, copper, tin e cheap.

I clienti appartenenti al gruppo diamond sono quelli con più alto valore per l'azienda, poiché spendono di più di tutti, molto frequentemente e con pochi giorni intercorsi tra la data di riferimento e l'ultimo loro acquisto.

A seguire vi sono i gruppi gold e silver, composti da clienti con valore attuale/potenziale medio-alto, con alcuni dei quali però che non acquistano da parecchi giorni. Infine i restanti cluster presentano consumatori che per un motivo o per l'altro non risultano così determinanti per il business aziendale.

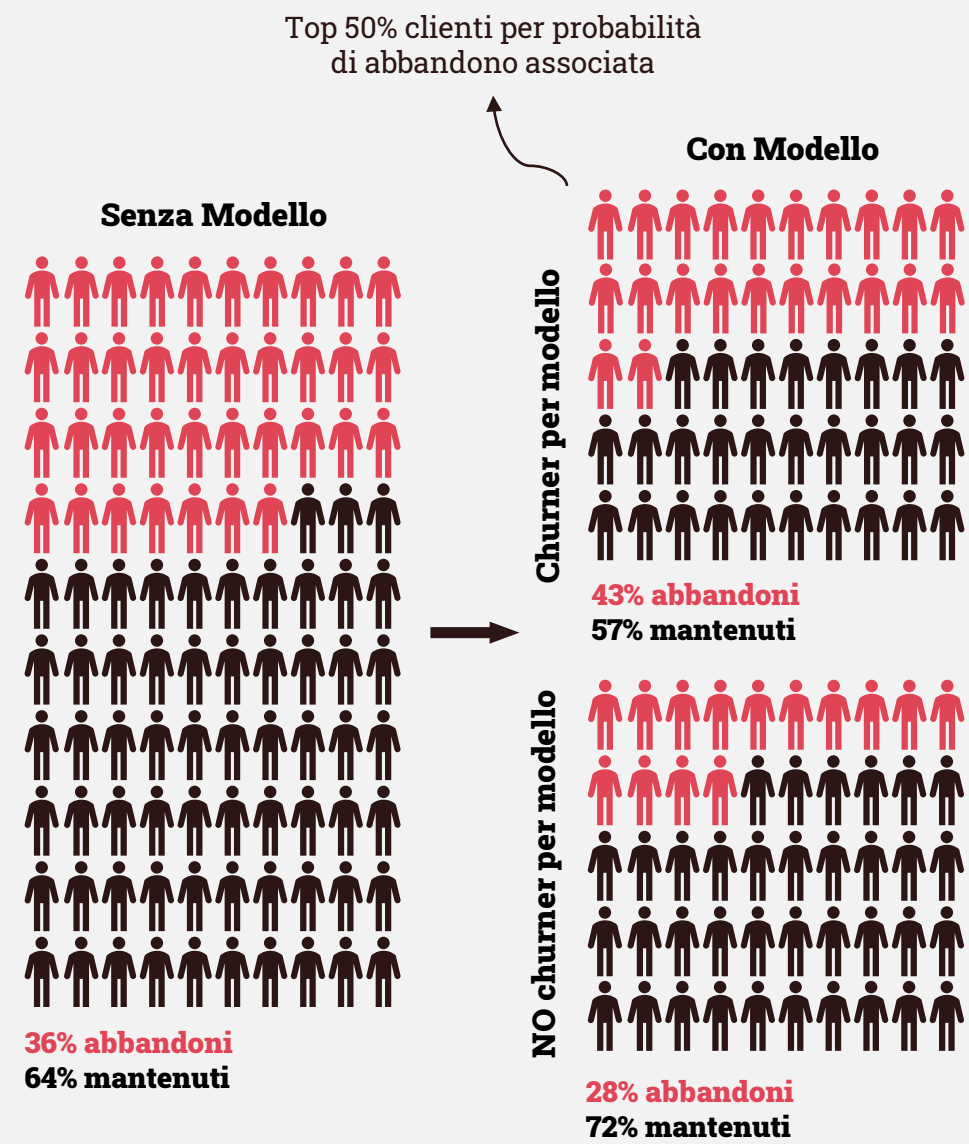
PERFORMANCE MODELLO PER CLIENTI AD ALTO VALORE

Come si può vedere dalle analisi effettuate sul dataset di set, il modello classifica correttamente circa il 57% dei clienti ad alto valore. Nel particolare coloro realmente abbandonanti correttamente classificati sono il 55%, mentre i non churner individuati con esattezza sono il 59%.

Questi risultati portano a considerare le performance del modello come discretamente soddisfacenti, in quanto quest'ultimo permette di individuare con più facilità i clienti potenzialmente persi. Infatti, pensando di voler sottoporre una strategia anti churn a 100 clienti ad alto valore, procedendo senza modello (ovvero casualmente) troveremmo 36 clienti realmente lascianti (valore corrispondente al churn rate per questa fascia di clientela). Invece, utilizzando il modello si andrebbe ad individuare un target migliore, in quanto tra i 100 clienti il numero di reali churner si alzerebbe a 43 (+19/20%), permettendo dunque una migliore applicazione della strategia.

	Accuracy	TPR	TNR	Precision
Valore	57%	55%	59%	43%

Recap metriche su dataset test clienti ad alto valore



INDIVIDUAZIONE CLUSTER STRATEGICO



CLUSTER CLIENTI DA MONITORARE

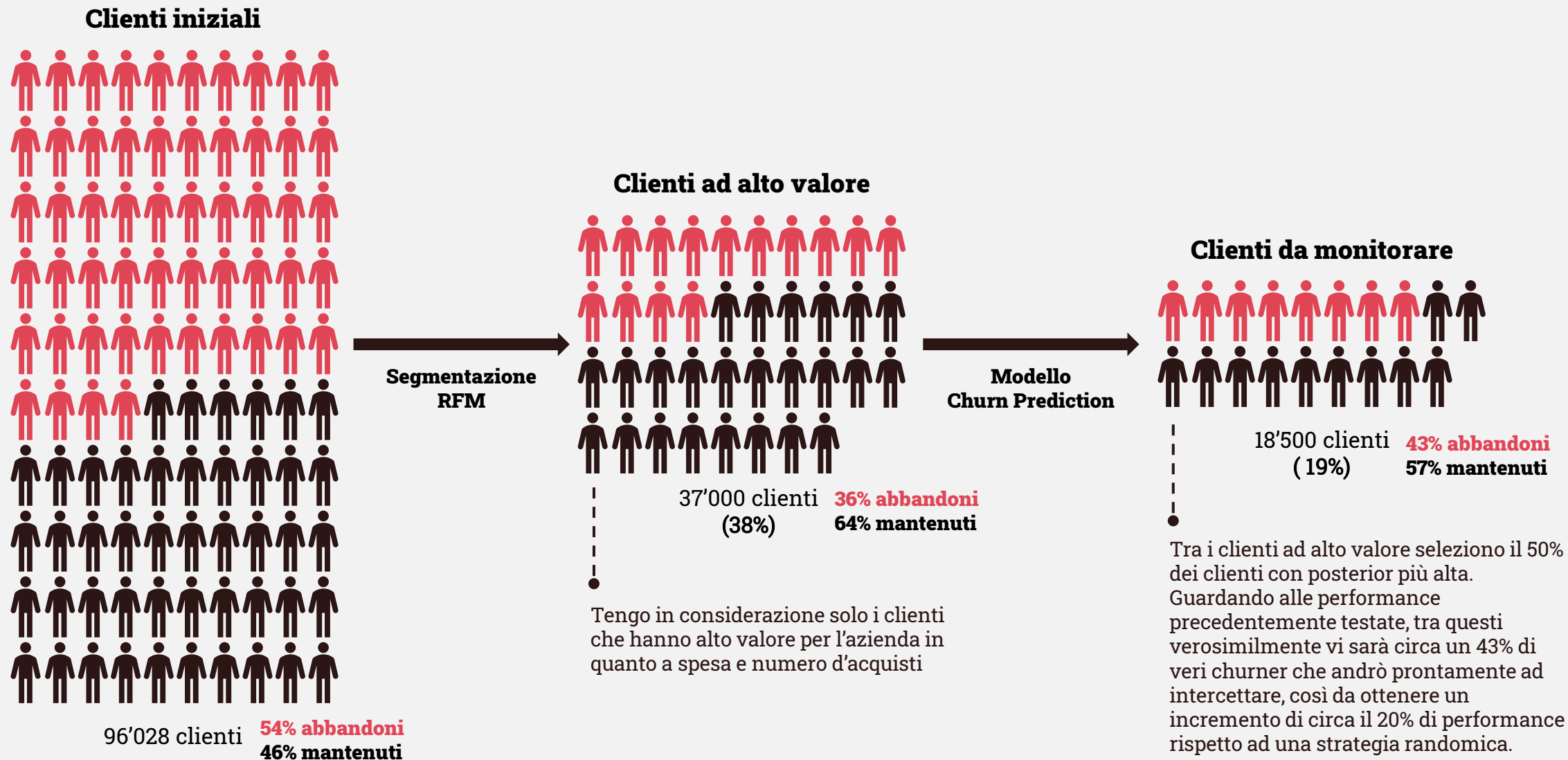
Una volta costruito il modello di churn prediction ed eseguita la segmentazione RFM sui clienti attivi tra febbraio e aprile, il passaggio successivo è stato unire le due strategie, così da individuare un cluster di clienti attivi a fine aprile ad alto valore e con elevata probabilità di abbandono, stabilendo dunque un target strategico per l'azienda.

Per quanto riguarda la segmentazione RFM, come detto in precedenza si è deciso di considerare clienti ad alto valore quelli appartenenti ai gruppi diamond, gold e silver, in quanto questi sono responsabili del 77,1% degli introiti, del 72,2% degli scontrini, essendo però il 38% del totale di consumatori. In quanto al modello di churn prediction si sono considerati ad alta probabilità d'abbandono i top 50% dei clienti per posterior assegnata dal modello.

Questa strategia applicata ai clienti del trimestre febbraio-aprile ha portato all'individuazione di un cluster target composto da 18'500 clienti, il 19% del totale e il 50% di quelli ad alto valore.



CLUSTER CLIENTI DA MONITORARE



ANALISI MAIL MARKETING

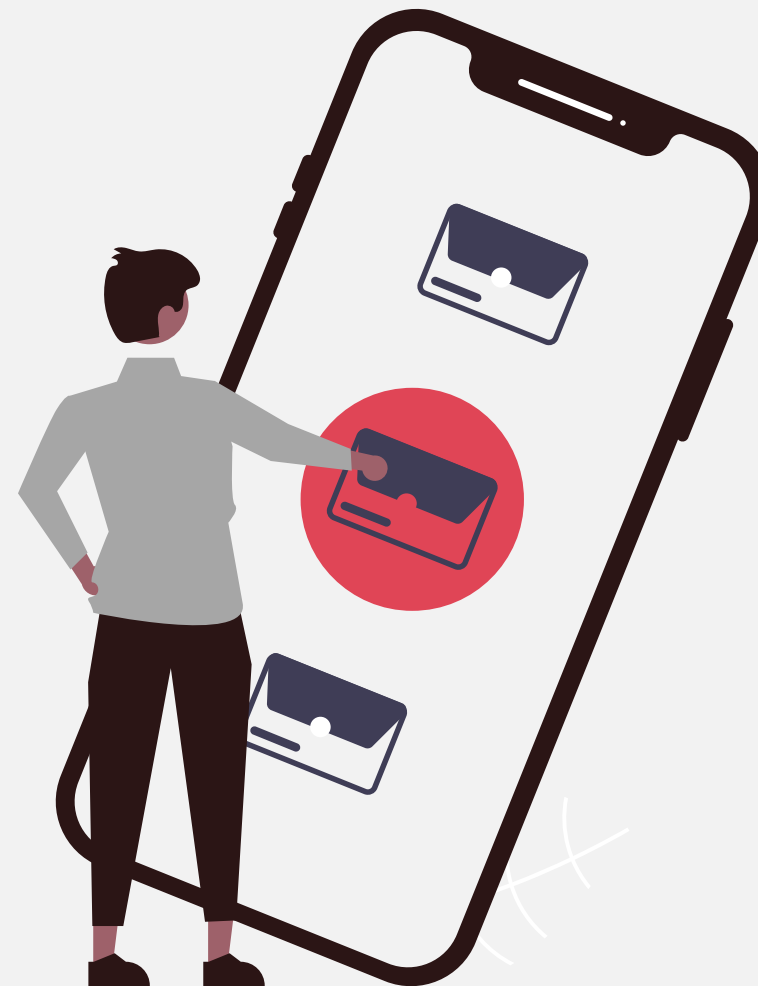


MAIL MARKETING E PROBABILITA' D'ABBANDONO

Tra i dati in nostro possesso erano presenti quelli relativi alla campagna di mail marketing condotto nel periodo tra gennaio e aprile 2019, la quale ha riguardato poco più di 190 mila clienti.

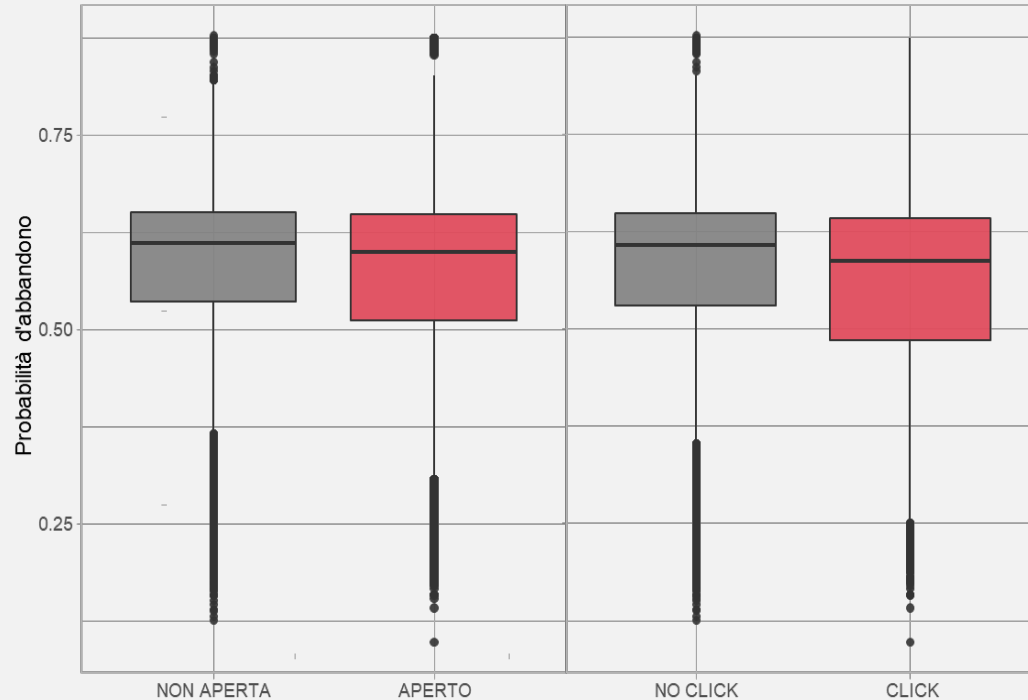
Si è deciso di analizzare la possibile efficacia del mail marketing sul contenimento del churn rate, studiando la relazione tra la probabilità d'abbandono dei clienti attivi nell'ultimo trimestre, calcolata dal modello precedentemente presentato, e la risposta di questi ultimi alla campagna mail (ovvero se e quanto hanno aperto/cliccato le mail).

Sia i grafici che i test statistici condotti sul campione di clienti sembrano evidenziare un effetto positivo, seppur limitato, della campagna marketing rispetto al customer churn, in quanto i consumatori che hanno cliccato (o anche solo aperto) una di queste mail tendono ad avere una probabilità d'abbandono leggermente inferiore rispetto agli altri.



MAIL MARKETING E PROBABILITÀ D'ABBANDONO

Distribuzioni condizionate



Commento: La distribuzione relative alla probabilità d'abbandono degli utenti che hanno cliccato appare concentrata maggiormente su posterior più basse rispetto agli altri. Stesso fenomeno, seppur in maniera meno marcata, succede tra coloro che hanno anche solo aperto la mail rispetto a quelli che non lo hanno fatto.

Regressione sulle posterior

Variabile	β (C.I. 95%)	<i>P value</i>
Il cliente ha aperto la mail? Sì	-0,4% (-0,6%, -0,2%)	~ 0
Il cliente ha cliccato il link nella mail? Sì	-2,2% (-2,5%, -1,9%)	~ 0

Commento: Lo studio del modello lineare avente come dipendente le probabilità d'abbandono e come esplicative le variabili dicotomiche relative all'apertura e al click della mail ha portato a rifiutare le ipotesi nulle sulla significatività sia generale che sui singoli coefficienti. Questo per tanto rileva una possibile efficacia della campagna marketing, in quanto il click o anche solo l'apertura della mail porta ad una statisticamente significativa riduzione della probabilità d'abbandono. I valori puntuali dei coefficienti sono da ritenersi puramente indicativi e tutt'altro che affidabili.

CONCLUSIONI

RISULTATI

1. **CHURN PREDICTION AFFIDABILE:**

Il modello di churn prediction addestrato ha garantito nella parte di test un'accuratezza nell'identificare i churner effettivamente migliore rispetto ad un approccio casuale.

2. **INDIVIDUAZIONE CLUSTER EFFICACE:**

La strategia realizzata si è rivelata in grado di identificare efficacemente un cluster di clienti ad alto valore e possibili churner, permettendo di targettizzare in modo migliore eventuali contromisure verso l'abbandono dei clienti.

3. **MAIL MARKETING UTILE:**

La campagna marketing eseguita tra gennaio e aprile ha dimostrato un'efficacia statisticamente significativa nella riduzione della probabilità d'abbandono prevista dal modello. Se un cliente clicca o anche solo apre una mail risulta meno propenso all'abbandono rispetto agli altri.

SUGGERIMENTI

1. **AFFINAMENTO DEL MODELLO:**

Per migliorare il modello è possibile andare a ricercare tra i dati dei clienti altre variabili potenzialmente influenti sulla probabilità d'abbandono, così da migliorare le performance generali del modello.

2. **NUOVA STRATEGIA ANTI CHURN:**

Sebbene la campagna di mail marketing sia risultata in qualche modo efficace nel contenimento dei churner, le sue performance sono senz'altro migliorabili. Restano poi da studiare possibili differenze tra le varie tipologie di mail inviate.