Unblind Consulting

KNOWLEDGE BEHIND CLASSIFYING

MODELIZE CAMPAIGNS

Belardinelli Marco, Cicalè Juri,
Lepore Marco Antonio, Manzo Manuel

GOAL SETTING

Elaborazione di un **modello** volto a classificare la clientela della banca in relazione alla scelta di sottoscrivere una determinata tipologia di contratto.

Il dataset contiene caratteristiche dello specifico cliente, informazioni sulle politiche commerciali adottate e sul contesto macroeconomico di riferimento.

Il modello permetterà di migliorare sensibilmente la **strategia commerciale** della banca, ottimizzando le risorse interne in relazione alla qualità dei risultati.

ANALYSIS & FINAL RESULTS

Di seguito un'analisi descrittiva delle **variabili** relativamente alla clientela che ha deciso di sottoscrivere un contratto (nello specifico quelle relative alle caratteristiche del contatto):

(Tabelle: i valori indicano la percentuale di sottoscrittori rispetto al totale di persone contattate che presentano quella caratteristica. ES: il 10% dei divorziati contattati ha sottoscritto un contratto)

AGE

Age	<23	23-60	>60
Contratti	20%	10.5%	40%

JOB

Job	admin	technician	retired	management	student
Contratti	13%	10.8%	25%	11%	31%

EDUCATION

Education	university	high	professional
Contratti	13.7%	10.8%	11.3%

PREVIOUS DEFAULT

Previous Default	no	unknown	yes
Contratti	13%	0.05%	0%

MARITAL STATUS

Marital Status	divorced	married	single
Contratti	10%	10%	14%

CONTACT

Contact	cellular	phone
Contratti	14.7%	0.05%

HOUSE OWNER

House Owner	no	yes
Contratti	10.9%	11.6%

EXISTING LOANS

Existing Loans	no	yes
Contratti	11.3%	10.9%

DAYS

Days	monday	tuesday	wednsday	thursday	friday
Contratti	10%	11.7%	11.7%	12%	10%

PREVIOUS

Previous	0	1	2	3	4
Contratti	8%	21%	46%	60%	54%

MONTH

Month	april	may	june	july	august	september	october	november	december
Contratti	20%	6%	10.5%	9%	10.6%	45%	44%	10%	49%

È possibile rilevare dunque alcuni **pattern di potenziale interesse:**

- Acquistano di più le due fasce di età e categorie sociali agli antipodi ovvero studenti e pensionati;
- All'aumentare del livello educativo c'è una tendenza incrementale nella sottoscrizione;
- Gli ultimi 3 mesi dell'anno vedono la concentrazione maggiore dei contratti (con il giovedì che sembra essere il giorno della settimana più favorevole alle conversioni positive);
- Il possesso di un immobile e prestiti in essere non sembrano avere particolare rilevanza;
- I contatti mediante cellulare risultano sensibilmente più efficaci;
- I single tendono a sottoscrivere maggiormente;
- I potenziali clienti già contattati in precedenti campagne manifestano un tasso di conversione maggiore. Quando si superano i 3 contatti questo effetto sembrerebbe diminuire progressivamente.
- Coloro che hanno alle spalle un fallimento non sottoscrivono mai.

IL MODELLO

Lo strumento della **regressione logistica** risulta ideale quando si tratta di classificare in maniera coerente i dati a disposizione; nello specifico, sulla base di variabili selezionate è possibile ottenere un modello che individua le caratteristiche comuni dei clienti in relazione alle loro scelte di sottoscrizione.

Da questo punto di vista è fondamentale rilevare che, un modello basato solo sulle caratteristiche degli utenti a disposizione, sarebbe stato ideale per gli scopi dell'analisi ma purtroppo non adeguato ad una classificazione di buona qualità.

È stato quindi necessario includere anche un importante **indicatore macroeconomico**, ovvero il tasso Euribor in essere al momento del contatto (risultato come prevedibile dirimente per l'influenza nelle scelte dei potenziali clienti) insieme alle variabili MONTH, PDAYS, CAMPAIGN ed AGE.

		Actual class: 0	Actual class: 1
Predicted class	0	7877	379
	1	3087	1013

Immagine 1 - Confusion Matrix

Come evidenziato nella **confusion matrix** di cui sopra, il modello permette di dimezzare le chiamate necessarie per ottenere un esito positivo.

Maggiore focus è stato dedicato alla sensitivity, con l'obiettivo primario di classificare correttamente i sottoscrittori, riducendo al minimo il rischio potenziale della perdita clienti.

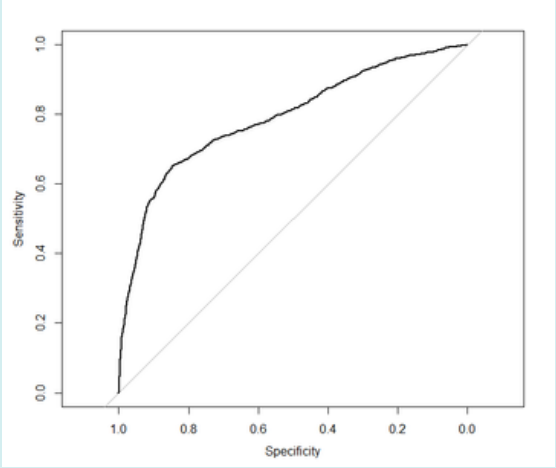

Immagine 2 - Curva ROC

Come possiamo osservare dalle risultanze di due indicatori tipici (la curva ROC e l'AUC, ovvero il valore dell'area sotto di essa, rappresentano la capacità di classificazione mettendo in rapporto il true positive rate con il false positive rate) otteniamo dei risultati estremamente positivi.

Si ipotizzi ad esempio un utente di 44 anni, già contattato 20 giorni prima (per la prima volta nella campagna in corso), durante il corrente mese (marzo), con tasso Euribor attuale (2,753%); il cliente in questione verrebbe classificato come sottoscrittore del contratto proposto.

APPENDICE TECNICO METODOLOGICA

Si può suddividere il lavoro effettuato in due fasi principali:

ANALISI ED ELABORAZIONE PRELIMINARE DEL DATASET

IMPLEMENTAZIONE DEL MODELLO DI REGRESSIONE LOGISTICA

Abbiamo verificato e corretto l'eventuale presenza delle seguenti criticità:

- duplicati
- outliers
- valori nulli

Si è provveduto poi a scalare i dati (ove necessario) con la **metodologia min/max**.

È stata implementata una **matrice di correlazione**, una **funzione stepwise** ed un **algoritmo random forest** per determinare le variabili da utilizzare nel modello.

Si è diviso il dataset in **train e test** per poi andare ad utilizzare una **regressione logistica** pesata così da compensare il problema dello sbilanciamento del dataset e tener conto dell'importanza relativa dei dati di entrambe le classi durante l'addestramento.

Per quanto riguarda la scelta del cut off si è deciso di utilizzare il parametro soglia dello 0,4 per migliorare la **sensitivity**.

La significatività dei coefficienti è stata verificata attraverso un **test di Wald** mentre un **McFadden** è stato calcolato per valutare la bontà generale del modello.