Michela Ropele

Data Mining: banca

1. Quesito d'analisi

Una banca che opera nel Credito al Consumo delle famiglie decide di saggiare la proposta del proprio Prestito Personale attraverso il canale telefonico outbound.

Considerato che il costo di ogni telefonata è di €2,50, ed avendo stabilito un costo massimo mensile per le attività di Direct di €100.000, la proposta può essere fatta solo a 40.000 nominativi.

La Direzione Marketing decide di estrarre casualmente i clienti a cui fare la proposta telefonica (40.000 su 200.000).

Il mese successivo vengono osservati i risultati delle vendite ottenute attraverso la proposta telefonica e la Direzione chiede al proprio ufficio analisi di elaborare un modello per poter ripetere l'operazione non in modo aleatorio ma mirato ad ottimizzare l'investimento.

L'ufficio di Analytics utilizza dunque i 40.000 clienti oggetto del test (considerati come rappresentativi dell'intero portafoglio clienti) per creare un modello per prevedere la variabile target "accetta la proposta di finanziamento".

Per modellare tale fenomeno, l'ufficio ha a disposizione diverse variabili che caratterizzano ogni cliente e che possono essere ricondotte a 3 macro categorie:

- variabili socio-demografiche

- variabili di "equipaggiamento"

- variabili storico-comportamentali

Per la nuova campagna online vengono stanziati €25.000.

E' dunque necessario selezionare 10.000 clienti, tra i 160.000 clienti (cioè il 6.25%) che non sono stati chiamati la volta precedente, in modo da ottimizzare l'investimento.

Si identifichi il miglior modello che selezioni i 10.000 clienti con più elevata probabilità di accettare la proposta di finanziamento.

I dati sono nel file bancafamiglie.csv e le variabili in appendice.

2. Dati

Sono inizialmente presenti 40 mila osservazioni non ripetute e 74 variabili. Le unità statistiche sono i singoli clienti oggetto del test. La risposta TARG_TOT è dicotomica e presenta una percentuale di 1 appena superiore al 5%. Si tratta quindi di un caso di dati sbilanciati.

L'obiettivo dell'analisi è individuare il modello più efficace per identificare i 10.000 clienti con la maggiore probabilità di accettare la proposta di finanziamento.

Eseguo le seguenti operazioni di pulizia:

- Rimuovo l'identificativo di riga;
- La variabile FIND_PPQ18SS_MONTH_DAT_MAX_FIN (mesi dalla richiesta di finanziamento senza seguito) presenta il 97% di valori mancanti. La rendo dicotomica per chi ha avuto questa richiesta negata e chi no;
- Per le variabili intere ANZ_BAN, ANZ_RES e ANZ_PROF creo 4 classi, in cui l'ultima contiene sia i valori mancanti che quelli codificati con 98 e 99, poiché, anche se la causa del dato mancante varia a seconda della raccolta, per questa analisi possono essere trattati allo stesso modo e accorpati;
- Rimuovo un totale di 50 osservazioni con valori mancanti in corrispondenza delle variabili FIND_NUM_MEN_RES, PPQ_NUM_MEN_RES, IMP_FAM e IMP_RED perché osservo che sono tutte osservazioni con risposta pari a 0, quindi non perdo informazione sulle unità di interesse;
- Le variabili numeriche FIND_PPQ18_IMP_FIN e PPQ_18_IMP_FIN, invece, vengono convertite in categoriali con modalità '0', '1' e 'Mancante', per evitare di perdere 14 osservazioni con risposta positiva, che altrimenti verrebbero escluse eliminando i valori mancanti;
- Accorpo la modalità 'Società/Associazioni' della variabile PROF in 'Sconosciuto' perché contiene due sole osservazioni.

Tutte le variabili quantitative presentano più di quattro modalità numeriche. Se la suddivisione in stima e verifica sarà equilibrata, non dovrei incontrare problemi di stima nei modelli che includono spline di lisciamento di grado 3.

Ora il dataset presenta 39950 osservazioni e 73 variabili, con uno sbilanciamento della risposta pari a 5,9% di osservazioni che presentano modalità 1 e 94,1% di osservazioni con modalità 0.

2. Bilanciamento

Suddivido il dataset in due sottoinsiemi: stima e verifica, assegnando le unità statistiche in modo casuale (70% stima e 30% verifica) e impostando un seme a 25 per la riproducibilità.

Campiono ulteriormente il 30% delle osservazioni con risposta pari a 0 nell'insieme di stima, in modo da ridurre lo sbilanciamento tra le due modalità della variabile risposta. Dopo il campionamento, la proporzione di osservazioni con risposta pari a 1 sale al 17%. L'insieme di verifica viene invece mantenuto invariato, per garantire una valutazione imparziale delle prestazioni del modello.

3. Operazioni preliminari

Sono presenti 30 coppie di variabili con correlazione assoluta compresa tra 0,70 e 0,90. Scelgo di non rimuoverle, lasciando che siano i modelli a selezionarle e privilegio la scelta di modelli che la gestiscono.

Eseguo la standardizzazione separata delle variabili quantitative nei due sottoinsiemi, in modo da rendere i coefficienti confrontabili in termini di una deviazione standard (ad esempio nel modello lineare senza interazioni) ed evitare che alcune variabili influenzino eccessivamente l'analisi nei modelli sensibili alla scala (come la PCA).

4. Prestazioni del modello

La metrica usata per confrontare le prestazioni di modelli diversi è la curva Lift in corrispondenza del valore 0,0625 perché il problema richiede di selezionare esattamente il 6,25% della popolazione. In quel punto, la curva quantifica quanto l'uso del modello è più efficace rispetto alla selezione casuale, ed è quindi il criterio corretto per valutare se il modello consente di "ottimizzare l'investimento" nelle chiamate.

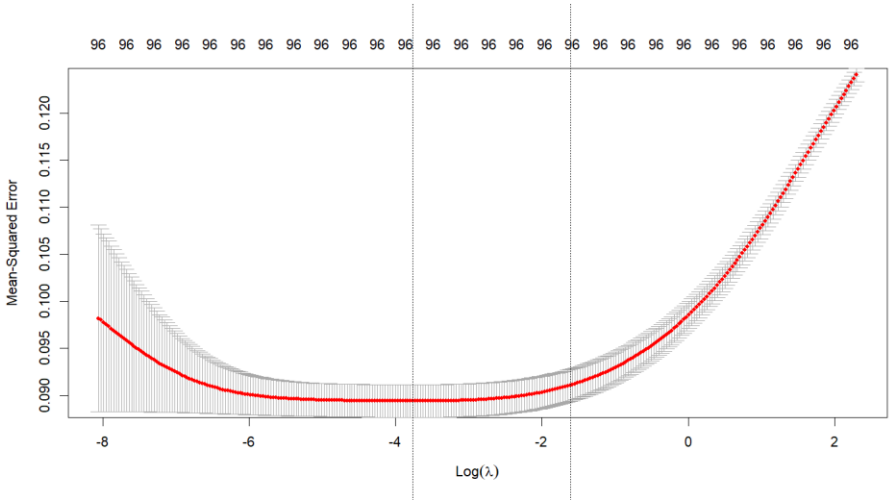
La scelta del parametro di regolazione verrà effettuata confrontando i modelli all'interno della stessa classe tramite curva lift (SVM e albero di classificazione) oppure errore quadratico medio. Se dovrò effettuare la scelta di un criterio di informazione utilizzerò l'AIC perché tende a minimizzare la perdita di informazione.

Ogni volta che otterrò il modello migliore, all'interno di ciascuna classe di modelli, lo stimerò su tutti i dati di stima e ne calcolerò la metrica su quelli di verifica. Al termine si eleggerà il modello migliore confrontando le prestazioni così ottenute.

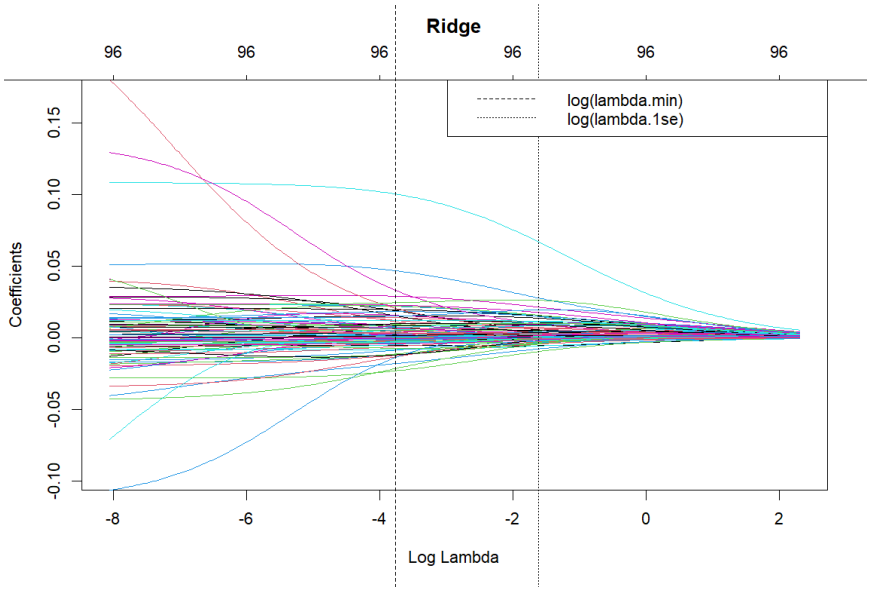
5. Modelli

5.1 Modello Ridge

Lo stimo tramite convalida incrociata i 5 sottoinsiemi dell'insieme di verifica. Uso una griglia di 300 valori lambda compresi tra 10^-3,5 e 10. Il lambda minimo è 0.0232 come si può osservare dal seguente grafico dell'errore quadratico medio al variare log(lambda).



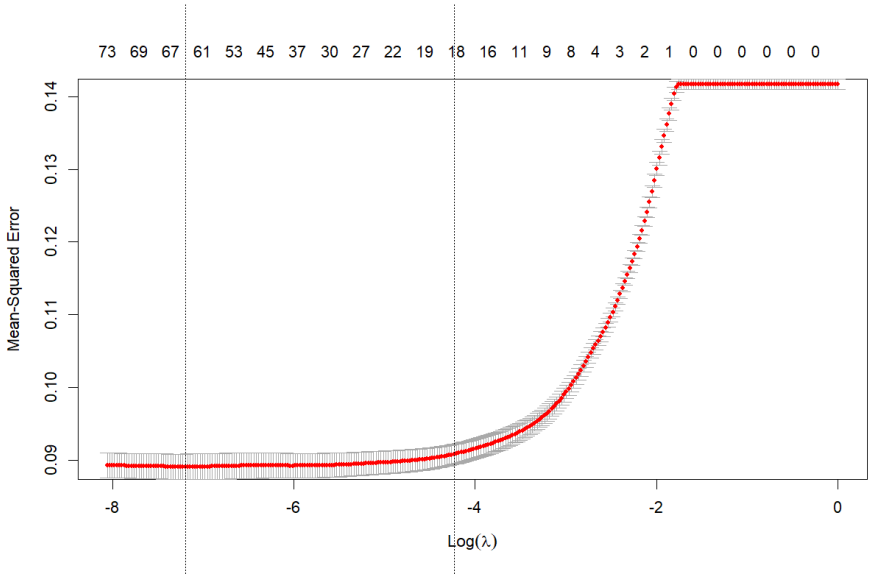
Segue anche il grafico del profilo dei coefficienti con due linee tratteggiate in corrispondenza del lambda minimo e del lambda 1 standard error.



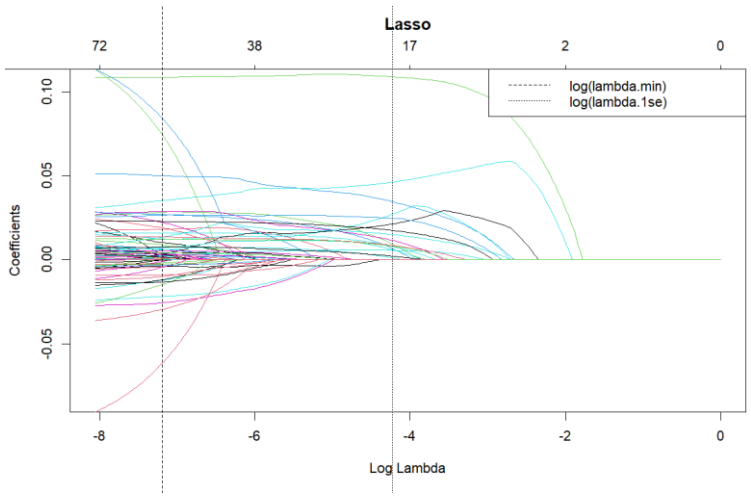
Metrica: 8.1473

5.2 Modello Lasso

Anche qui uso la convalida incrociata in 5 sottoinsiemi e una griglia di 300 valori di lambda compresi tra 10^-3.5 e 0. Il lambda minimo è 0.0232 come si può osservare dal seguente grafico dell'errore quadratico medio al variare log(lambda). Il lambda minimo è 0.000749, come si può osservare dal seguente grafico dell'errore quadratico medio al variare log(lambda). Vengono selezionate 64 coefficienti su 97.



Segue anche il grafico del profilo dei coefficienti con due linee tratteggiate in corrispondenza del lambda minimo e del lambda 1 standard error.



Metrica: 8.1473

5.3 GAM con variabili selezionate da Lasso

Lo stimo solo sulle variabili selezionate dal lasso e uso 3 gradi di libertà per ogni variabile quantitativa. Non includo lisciatori bivariati.

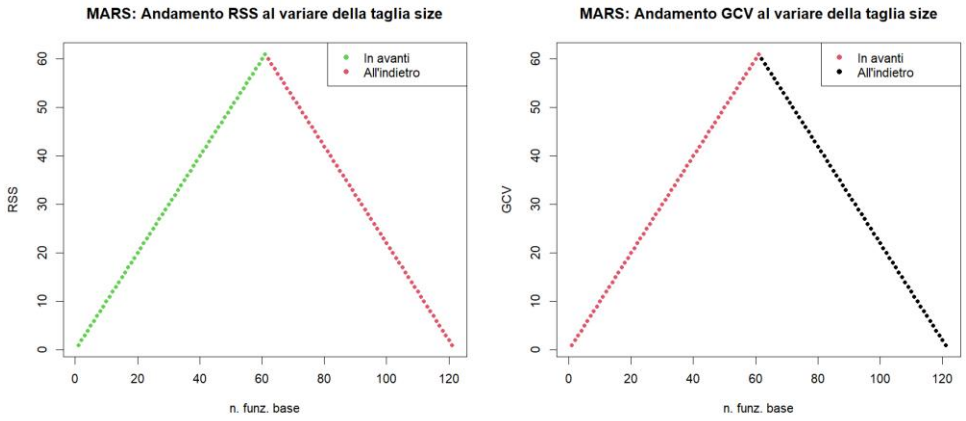
Metrica: 8.4627

5.4 MARS

Uso il numero di funzioni di base massimo pari a 100 ed un numero di nodi calcolato di default pari a 20. Inserisco nel modello solo gli effetti principali ma lascio la possibilità di introdurre interazioni. Il criterio di arresto è impostato con una tolleranza di 0,001.

Il risultato restituisce 19 interazioni tra coppie di variabili ed un valore lift buono.

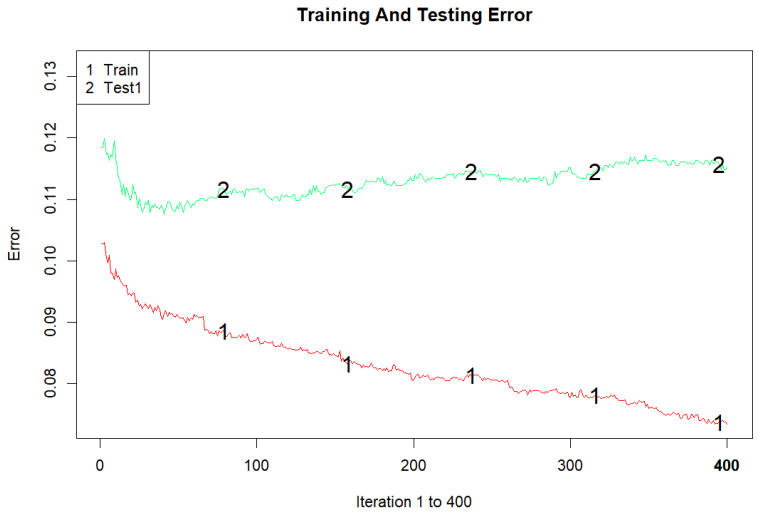
Tuttavia, il grafico di crescita e potatura mostra un andamento troppo lineare nell'errore e questo andrebbe indagato prima di, eventualmente, scegliere questo modello.



Metrica: 8.3313

5.5 Boosting

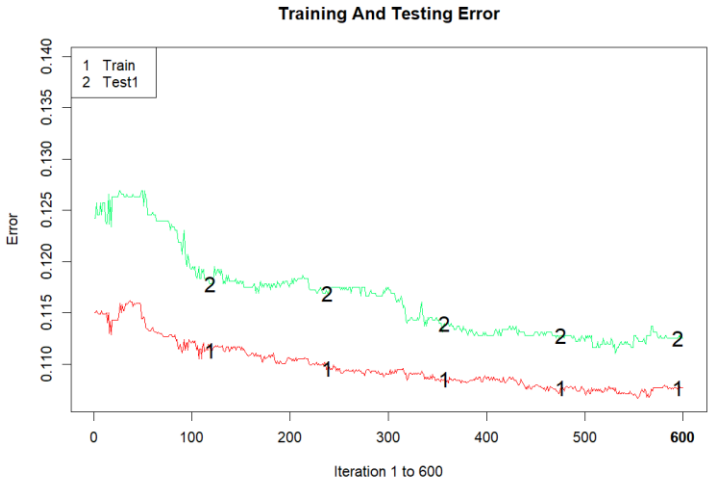
Stimo il modello con 400 iterazioni (anche se vedo che c'è una stabilizzazione dell'errore di test già a 350). Uso i parametri di default, dunque peso ciascuna classificazione errata di 0.1, e una frazione di dati da usare nel bagging di 0.5 e una profondità massima degli alberi pari a 3.



Metrica: 7.4114

5.6 Boosting con stump

Provo a stimare anche un boosting con solo alberi ad un unico split. Inizio utilizzando 400 iterazioni ma, data l'instabilità degli alberi poco profondi, non osservo una stabilizzazione dell'errore nell'insieme di test. Aumento le iterazioni a 600.

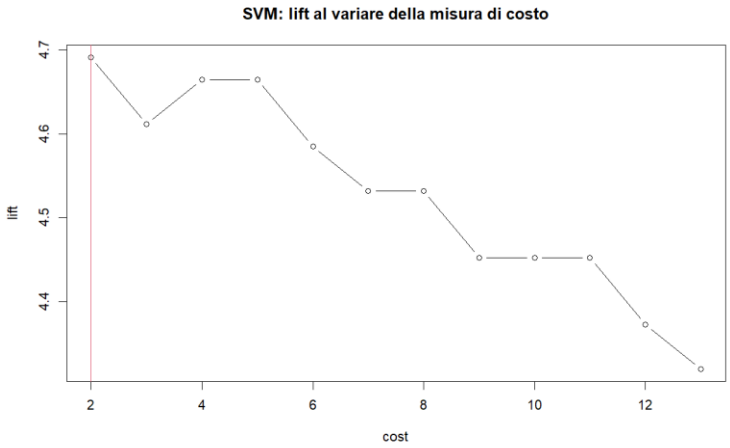


Noto che la metrica è superiore a prima. Potrebbe indicare che non sono presenti interazioni particolarmente forti, altrimenti ci sarebbe peggioramento evidente nei risultati.

Metrica: 7.8056

5.7 Support Vector Machine

Calcolo una Support Vector Machine testando valori del parametro di costo compresi tra 2 e 13 tramite il metodo di stima-convalida e sulla base del valore lift in 0,0625. Il parametro di regolazione selezionato è 2. Stimo il modello su tutto l'insieme di stima.

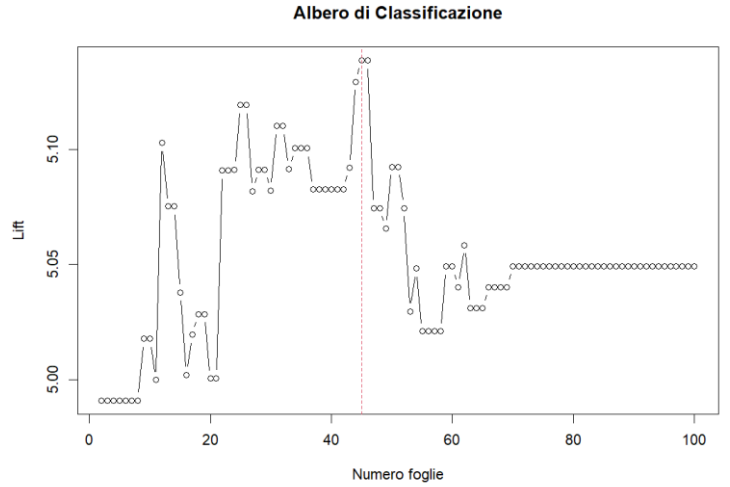


Metrica: 8.1473

5.8 Albero di classificazione

Uso la convalida incrociata su 5 sottoinsiemi. Imposto l'arresto dell'algoritmo ad un numero di osservazioni minimo per foglia pari a 15 e una devianza minima di 0,001. Il test viene eseguito su alberi con un numero di foglie j variabile da 2 a 100. Dal grafico di andamento del fattore lift in funzione della taglia dell'albero osservo particolare instabilità tra i parametri di regolazione 9 e 21, e poi una stabilizzazione da 25, con un picco a 44.

Raggiungo il massimo in corrispondenza di j=44, dunque un numero di foglie pari a 45.



Stimo l'albero finale potato in corrispondenza di j sull'insieme con tutti i valori di stima. Non inserisco il grafico dell'albero finale perché non è leggibile.

Metrica: 6.4915

5.9 Analisi Discriminante Lineare

Provo a stimarlo utilizzando solo le variabili quantitative, ma compare un avviso di collinearità tra alcune di esse.

Metrica: 8.1210

Stimando il modello sulle sole variabili quantitative selezionate dal lasso, la metrica cala a 8.0947. Me lo aspettavo perché sono passata ad avere da 74 variabili di partenza, ad averne 59 contando solo quelle quantitative, a 45 se quantitative e selezionate da lasso. Ho comunque voluto testare anche questa configurazione e, nonostante tutto, il modello più ridotto non mostra prestazioni scadenti.

5.10 Analisi Discriminante Quadratica

Non posso stimare la QDA con le sole variabili quantitative perché nel gruppo con risposta 0 manca rango sufficiente e la matrice di covarianza risulta singolare (variabili troppo collineari).

Lo ristimo utilizzando le variabili quantitative selezionate dal Lasso e ottengo la metrica più bassa tra tutti i modelli. Ho comunque voluto includere anche questa prova per completezza.

Metrica: 6.3601

6. Conclusioni

Il modello migliore per la selezione dei clienti risulta il Modello Additivo con le sole variabili selezionate dal Lasso. In particolare, usare il modello risulta 8.42 volte più efficace rispetto alla selezione casuale. Il secondo modello migliore risulta essere il MARS, ma a seguito di una revisione dovuta allo strano andamento di devianza e GCV.

Modello	**Valore Lift 6,25%**
GAM selezione Lasso	8.462668
MARS	8.331261
Lineare Ridge	8.147289
Lineare Lasso	8.147289
SVM	8.147289
LDA quantitative	8.121008
LDA quantitative + selezione Lasso	8.094726
Boosting con stump	7.804629
Boosting	7.411405
Albero	6.491550
QDA quantitative + selezione Lasso	6.360142

Appendice: descrizione variabili

TARG_TOT 1 se il cliente accetta la proposta telefonica e 0 se il cliente non accetta.

Caratteristiche socio-demografiche

Codice_Cliente Codice identificativo cliente

AGE EtÃ

ANZ_BAN Anzianità Banca (espressa in anni)

ANZ_RES Anzianità Residenza (espressa in anni)

ANZ_PROF Anzianità professionale (espressa in anni)

COD_RES Tipologia Residenza

COD_STA_CIV Stato civile (X: sconosciuto, D: divorziato, V: vedovo, M: sposato, K: convivente, C: celibe/nubile)

NUM_FIGLI Numero figli

FLG_SEX Sesso

IMP_RED Reddito Richiedente

IMP_FAM Reddito Famiglia

PROF Professione

Equipaggiamento del cliente

FIND_PPQ_C_NUM_PRA Prestiti Personali in corso- numero pratiche

ALTR_PPQ_C_NUM_PRA Altri Prestiti in corso- numero pratiche

NUM_PPQ_C Totale prestiti in corso- numero pratiche

FIND_PPQ_C_IMP_RES Prestiti personali in corso - importo residuo al saldo

ALTR_PPQ_C_IMP_RES Altri prestiti in corso - importo residuo al saldo

IMP_PPQ_C Totale prestiti in corso - importo residuo a saldo

FIND_NUM_MEN_RES Prestiti personali in corso - numero mensilità residue al saldo

ALTR_PPQ_C_NUM_MEN_RES Altri prestiti in corso - numero mensilità residue al saldo

PPQ_NUM_MEN_RES Totale prestiti in corso - numero mensilità residue al saldo

FIND_PPQ_C_IMP_MEN Prestiti personali in corso - importo rata

ALTR_PPQ_C_IMP_MEN Altri prestiti in corso - importo rata

PPQ_C_IMP_MEN Totale prestiti in corso - importo rata

FIND_CC_C_NUM_PRA_GRA Prestiti finalizzati gratuiti in corso- numero pratiche

FIND_CC_C_NUM_PRA_TAS Prestiti finalizzati a tasso in corso - numero pratiche

ALTR_CC_C_NUM_PRA Altri prestiti finalizzati in corso - numero pratiche

CC_C_NUM Totale prestiti finalizzati in corso - numero pratiche

FIND_CC_C_IMP_RES_TAS Prestiti finalizzati gratuiti in corso- importo residuo al saldo

FIND_CC_C_IMP_RES_GRA Prestiti finalizzati a tasso in corso - importo residuo al saldo

ALTR_CC_C_IMP_RES Altri prestiti finalizzati in corso - importo residuo al saldo

CC_C_IMP_RES Totale prestiti finalizzati in corso - importo residuo al saldo

FIND_CC_C_IMP_MEN_TAS Prestiti finalizzati gratuiti in corso - importo rata

FIND_CC_C_IMP_MEN_GRA Prestiti finalizzati a tasso in corso - importo rata

ALTR_CC_C_IMP_MEN Altri prestiti finalizzati in corso - importo rata

CC_C_IMP_MEN Totale prestiti finalizzati in corso - importo rata

CRT_PRE_C_FLG_PRE Carta - Cliente in possesso di carta

CRT_TODU_REV Carta - Esposizione Carta di credito

Storico del cliente

FIND_PPQ18SS_MONTH_DAT_MAX_FIN Richieste Prestito personale negli ultimi 18 mesi senza seguito - mesi dalla richiesta di finanziamento

FIND_PPQ18SS_NUM_PRA Richieste Prestito personale negli ultimi 18 mesi senza seguito - numero richieste finanziamento

FIND_PPQ18_NUM_PRA Prestiti personali saldati negli ultimi 18 mesi - numero pratiche

ALTR_PPQ_18_NUM_PRA Altri prestiti saldati negli ultimi 18 mesi - numero pratiche

PPQ_18_NUM_PRA Totale prestiti saldati negli ultimi 18 mesi - numero pratiche

FIND_PPQ18_IMP_FIN Prestiti personali saldati negli ultimi 18 mesi - importo finanziato

ALTR_PPQ_18_IMP_FIN Altri prestiti saldati negli ultimi 18 mesi - importo finanziato

PPQ_18_IMP_FIN Totale prestiti saldati negli ultimi 18 mesi - importo finanziato

FIND_PPQ18_AVG_NUM_MEN Prestiti personali saldati negli ultimi 18 mesi - durata media per cliente

ALTR_PPQ_18_AVG_NUM_MEN Altri prestiti saldati negli ultimi 18 mesi - durata media per cliente

PPQ_18_AVG_NUM_MEN Totale prestiti saldati negli ultimi 18 mesi - durata media per cliente

FIND_PPQ18_AVG_IMP_MEN Prestiti personali saldati negli ultimi 18 mesi - importo medio rata per cliente

ALTR_PPQ_18_AVG_IMP_MEN Prestiti personali saldati negli ultimi 18 mesi - importo medio rata per cliente

PPQ_18_AVG_IMP_MEN Prestiti personali saldati negli ultimi 18 mesi - importo medio rata per cliente

FIND_CC18_NUM_PRA_GRA Prestiti finalizzati gratuiti saldati negli ultimi 18 mesi - numero pratiche

FIND_CC18_NUM_PRA_TAS Prestiti finalizzati a tasso saldati negli ultimi 18 mesi - numero pratiche

ALTR_CC_18_NUM_PRA Altri prestiti finalizzati saldati negli ultimi 18 mesi - numero pratiche

CC_18_NUM Totale prestiti finalizzati saldati negli ultimi 18 mesi - numero pratiche

FIND_CC18_IMP_FIN_TAS Prestiti finalizzati gratuiti saldati negli ultimi 18 mesi - importo finanziato

FIND_CC18_IMP_FIN_GRA Prestiti finalizzati a tasso saldati negli ultimi 18 mesi - importo finanziato

ALTR_CC_18_IMP_FIN Altri prestiti finalizzati saldati negli ultimi 18 mesi - importo finanziato

CC_18_IMP_FIN Totale prestiti finalizzati saldati negli ultimi 18 mesi - importo finanziato

CRT_PRE_18_FLG_PRE Carta ultimi 18 mesi - Cliente in possesso di carta

CRT_REV18_NUM_FIN_OCF Carta ultimi 18 mesi - Numero finanziamenti rimborso fine mese

CRT_REV18_IMP_FIN_OCF Carta ultimi 18 mesi - Importo finanziato rimborso fine mese

CRT_REV18_NUM_FIN_REV Carta ultimi 18 mesi - Numero finanziamenti rimborso revolving

CRT_REV18_IMP_FIN_REV Carta ultimi 18 mesi - Importo finanziato rimborso revolving

CRT_PRO18_NUM_FIN_DIR Carta ultimi 18 mesi - Numero finanziamenti rimborso promo diretto

CRT_PRO18_IMP_FIN_DIR Carta ultimi 18 mesi - Importo finanziato rimborso promo diretto

CRT_PRO18_NUM_FIN_DIS Carta ultimi 18 mesi - Numero finanziamenti rimborso promo distribuzione

CRT_PRO18_IMP_FIN_DIS Carta ultimi 18 mesi - Importo finanziato rimborso promo distribuzione

FIND_PPQ10Y_NUM_PRA Prestiti personali saldati prima di 18 mesi - numero pratiche

FIND_PPQ10Y_AVG_IMP_FIN Prestiti personali saldati prima di 18 mesi - importo medio finanziato per cliente

FIND_PPQ10Y_AVG_NUM_MEN Prestiti personali saldati prima di 18 mesi - durata media per cliente

FIND_PPQ10Y_AVG_IMP_MEN Prestiti personali saldati prima di 18 mesi - importo medio rata per cliente