Università degli studi di Milano-Bicocca,
Dipartimento di Informatica, Sistemistica e Comunicazione,
Facoltà di Data Science
Anno Accademico 2019/20

ELABORATO FINALE DI DIGITAL MARKETING

Simone Tufano, matricola 816984

AGENDA e BUSINESS QUESTIONS

01

ANALISI ESPLORATIVE

Trasformazione delle variabili, gestione delle date, comprensione generale dei dataset a disposizione.

02

Come posso ottenere una visione generale del comportamento della clientela?

Creazione di un modello di segmentazione basato su ‘Recency’, ‘Frequency’ e ‘Monetary’.

03

Come posso prevedere la probabilità di abbandono di un cliente?

Creazione di un classificatore attraverso tecniche di machine learning.

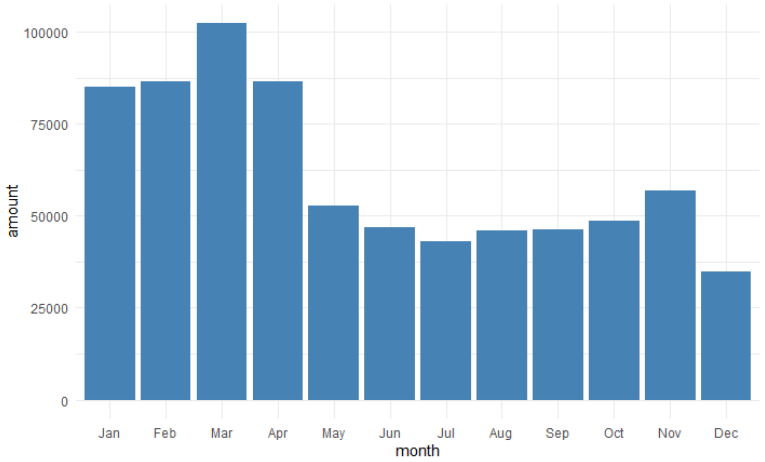
04

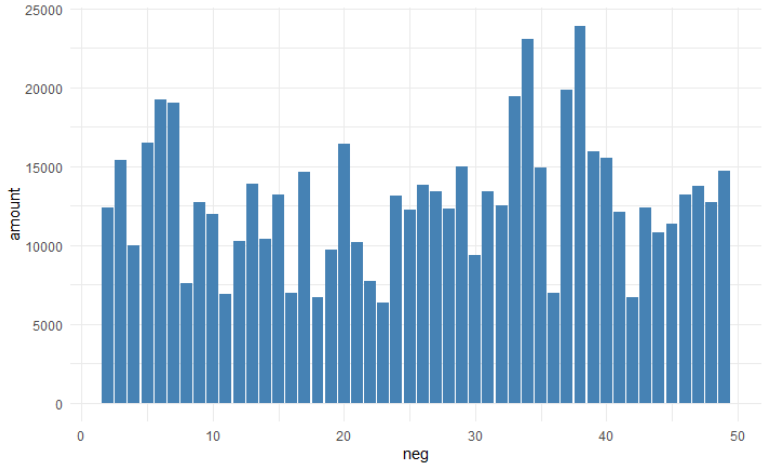
Quali sono le tendenze di acquisto?

Studio delle transazioni dei prodotti attraverso market basket analysis e regole associative.

01

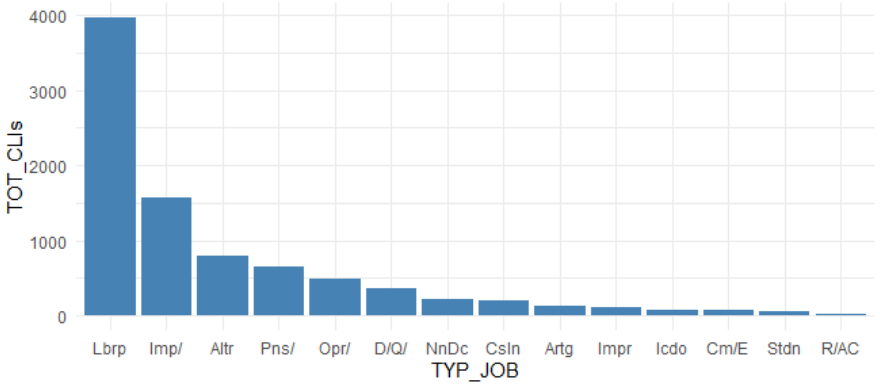
Il primo passo relativo alle analisi esplorative è stato quello di comprendere la struttura del dataset: chiavi primarie e secondarie per join future, distribuzioni delle variabili, prodotti più venduti, spese medie dei clienti ecc…

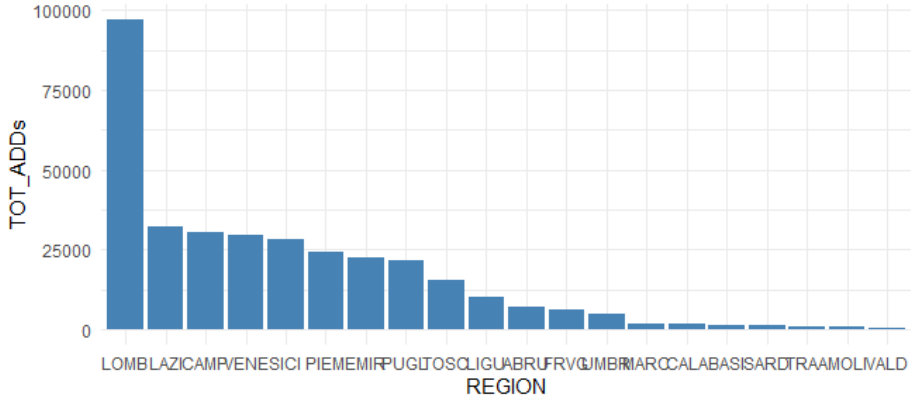




I due grafici mostrano la distribuzione degli account attivi a seconda dei mesi e la distribuzione dei negozi che vendono più prodotti. Il primo evidenzia come ci sia una riduzione nei mesi estivi e nel mese di dicembre.

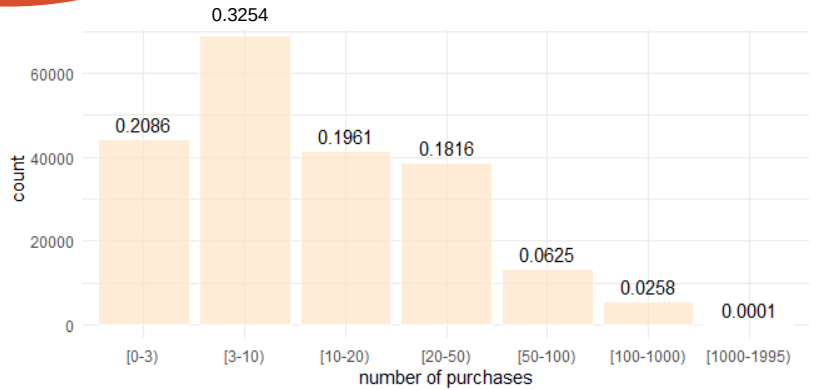
01

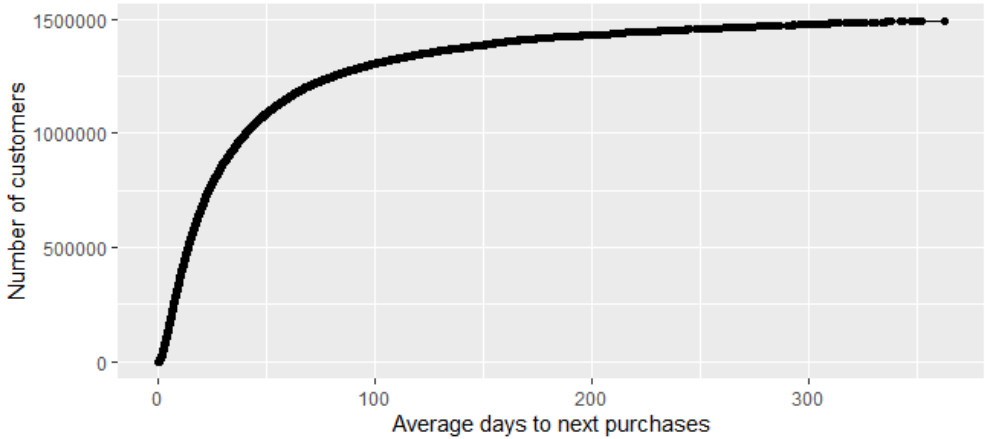




I dati fanno riferimento a 15 tipologie di lavoratori differenti, la stragrande maggioranza è un libero professionista e la regione con il maggior numero di acquisti è la lombardia, seguita da lazio e campania.

01





Il dataset relativo alle transazioni ha permesso di creare la distribuzione dei clienti per numero di acquisti. La maggioranza dei clienti, considerando tutto il periodo di riferimento compra dai 3 ai 10 articoli. Esiste anche una piccola percentuale (2.5%) che ha acquistato oltre 100 articoli.

Inoltre, lo stesso dataset ha permesso di creare la curva di riacquisto, ossia la distribuzione dei clienti in base alla differenza media in giorni tra un acquisto e il successivo (importante per la creazione della variabile target futura).

02

Dopo le analisi iniziali, si procede con la creazione delle variabili necessarie per segmentare i clienti sulla base di numero di acquisti, frequenza e spesa.

STEP

Ripulire il dataset dai duplicati: le transazioni a cui era stato applicato uno sconto erano presenti 2 volte;

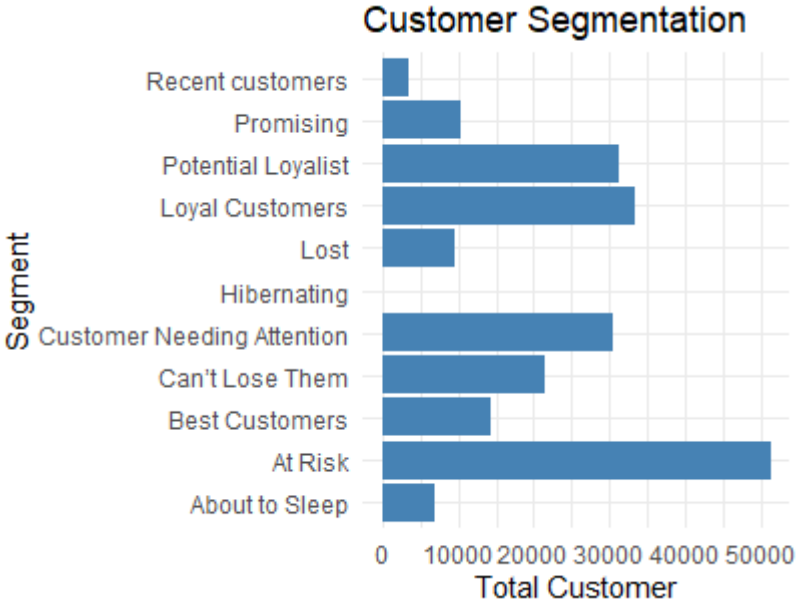
Creazione delle variabili:

Recency: differenza in giorni tra l'ultima data presente (30/04/2019) e la data dell'ultimo acquisto;
Frequency: numero di articoli per ogni cliente;
Monetary: somma degli importi per ogni cliente;

Scoring: Creazione di punteggi (R_score, F_score e M_score) basati sulla distribuzione delle variabili create in precedenza e sui rispettivi quartili;

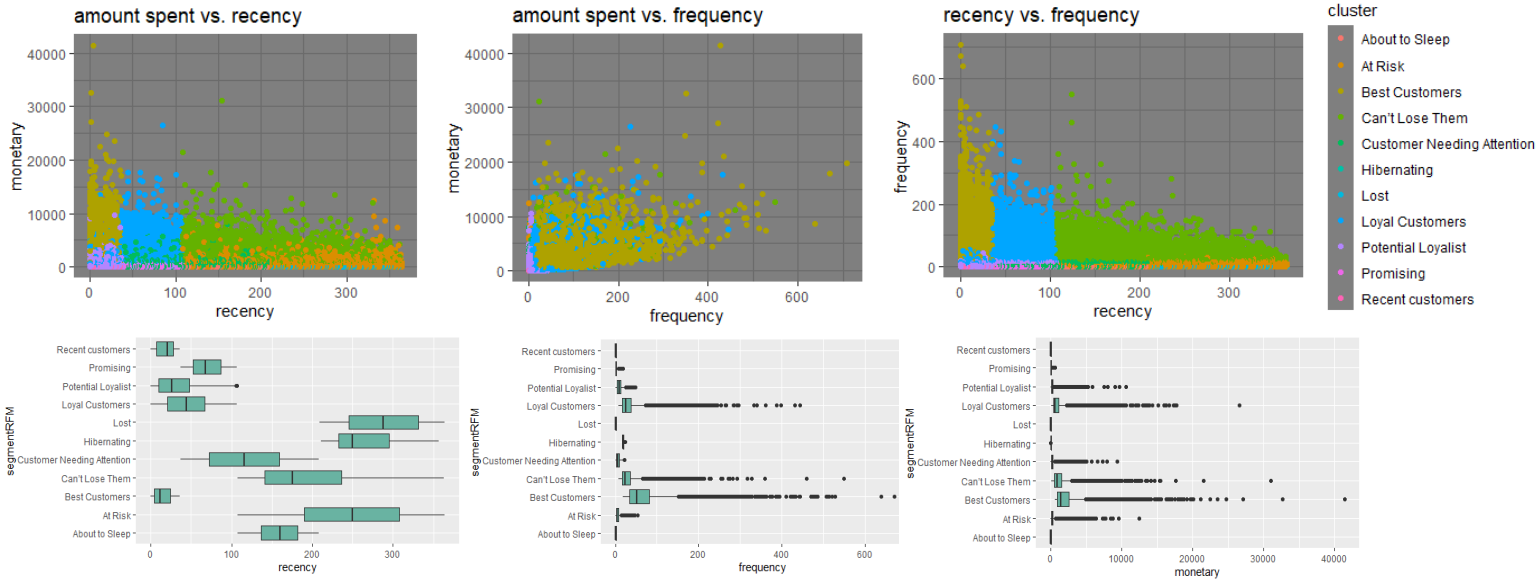
Segmentazione sulla base del punteggio creato in precedenza: 444-> 'best customer' / 111-> 'lost'.

RISULTATI MODELLO RFM 1



I risultati dell'analisi hanno portato a stabilire che la maggior parte dei clienti è a rischio, quindi presentano bassi punteggi di Recency e Frequency. Significa che sono clienti che non comprano da tanto ed è necessario riacquisirli attraverso messaggi personalizzati o offerte mirate.
Fortunatamente è presente anche una buona fetta di clienti loyal, o potenziali loyal, ossia clienti con punteggi medio-alti su tutte le dimensioni. Questi clienti sono tra i più importanti ed è necessario fidelizzarli e richiedere recensioni sull'azienda e sui prodotti acquistati per creare engagement.

RISULTATI MODELLO RFM 2



Gli scatterplot e i relativi boxplot dei cluster mostrano la presenza di molti outliers per le dimensioni di 'Frequency' e 'Monetary', ma evidenziano anche differenze significative ragionevoli tra i boxplot per la dimensione 'Recency'.

03

Si procede con il modello di churn, utile per stimare la probabilità di abbandono di un cliente.

STEP

Analisi esplorative e creazione della variabile target: Considerando la curva di riacquisto, si è stabilito un periodo di holdout pari a 60 giorni (circa l'80% dei clienti riacquistava entro 2 mesi), quindi la 'reference date' è il 01/03/2019;

Unione dei dataset attraverso join per avere un'analisi più dettagliata;

Analisi delle correlazioni e creazione delle partizioni con modalità bilanciate della variabile target;

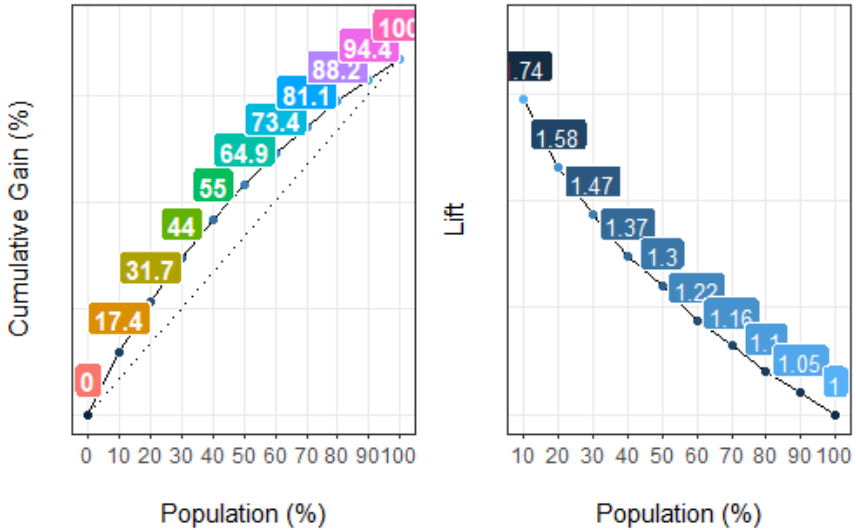
Addestramento dei seguenti modelli: Albero Decisionale (lanciato inizialmente per stabilire importanza delle variabili), Regressione Logistica, Random Forest, Naive Bayes, Bagging;

Assessment attraverso curve ROC e Gain Lift.

ASSESSMENT PER I MODELLI DI CHURN

Poiché non esiste una curva ROC che domini sulle altre, ci si è basati sulle gain lift per scegliere il modello vincente, ossia Random Forest.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    c1    c0
##         c1 25967  8700
##         c0 14574 14396
##
##                Accuracy : 0.6343
##                  95% CI : (0.6305, 0.638)
##     No Information Rate : 0.6371
##     P-Value [Acc > NIR] : 0.9294
##
##                   Kappa : 0.2501
##
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##             Sensitivity : 0.6405
##             Specificity : 0.6233
##          Pos Pred Value : 0.7490
```



Il modello presenta un'accuratezza del 63,4% e riesce a coprire mediamente più del 30% dei successi con il 20% della popolazione.

04

Infine, si svolge una Market Basket Analysis sui 100 items più acquistati.

STEP

Analisi esplorativa e identificazione dei prodotti più acquistati sul periodo di riferimento;

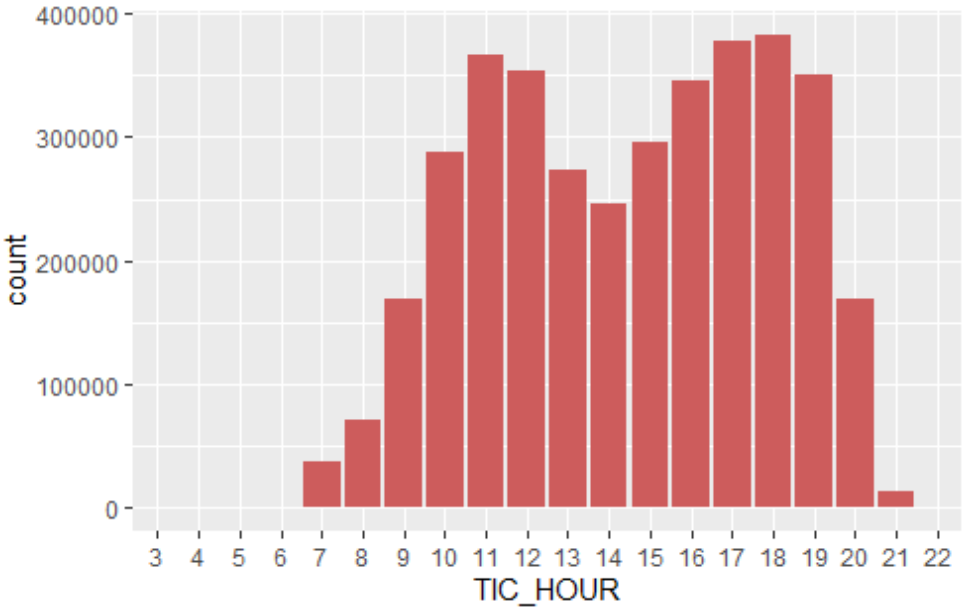
Trasformare il dataset in modo che ogni osservazione rappresentasse una transazione;

Analisi descrittive sulle transazioni

Creazione di regole associative;

Interpretazione dei risultati

RISULTATI MBA1

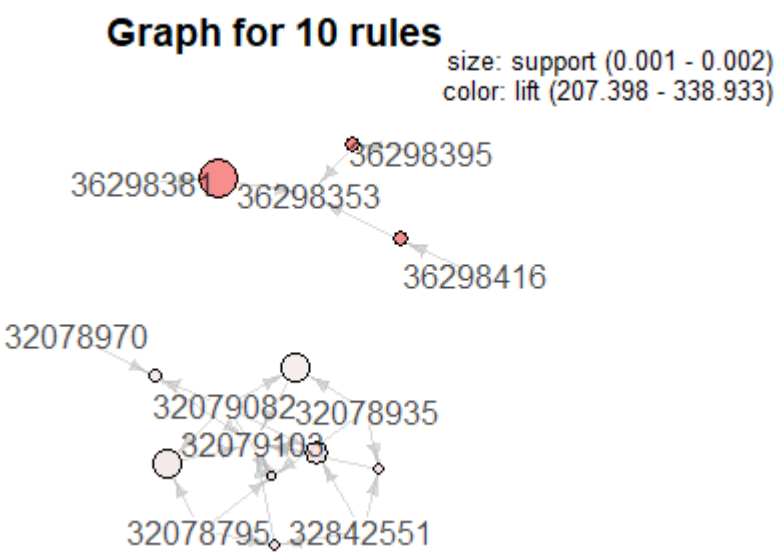


Dalle analisi esplorative iniziali sul dataset si può notare come le fascie orare di acquisto siano la tarda mattinata e il tardo pomeriggio (10:00-12:00, 16:00-19:00).

Transazioni totali presenti nel dataset: 885865
ID Items più frequenti: 33700716, 33817091, 34843564, 32882024, 34252904

RISULTATI MBA 2

```
> inspect(rules[1:10])
     lhs                                rhs         support     confidence coverage
[1]  {32078795,32842551}             => {32079103} 0.001069012 0.9185257  0.001163834
[2]  {32078935,32842551}             => {32079103} 0.001057723 0.9123661  0.001159319
[3]  {32078795,32078935,32079082}    => {32079103} 0.001009183 0.9048583  0.001115294
[4]  {32079082,32842551}             => {32079103} 0.001356866 0.8748180  0.001551026
[5]  {32078795,32079082}             => {32079103} 0.001543125 0.8575910  0.001799371
[6]  {32078970,32079082}             => {32079103} 0.001098361 0.8438855  0.001301553
[7]  {36298395}                      => {36298353} 0.001184153 0.8325397  0.001422339
[8]  {36298416}                      => {36298353} 0.001143515 0.8316913  0.001374927
[9]  {36298381}                      => {36298353} 0.001797114 0.8304643  0.002163987
[10] {32078935,32079082}             => {32079103} 0.001552155 0.8283133  0.001873875
     lift     count
[1]  229.9858  947
[2]  228.4435  937
[3]  226.5637  894
[4]  219.0420 1202
[5]  214.7286 1367
[6]  211.2970  973
[7]  338.9328 1049
[8]  338.5874 1013
[9]  338.0879 1592
[10] 207.3979 1375
```



Esempi intepretativi: Considerando la prima regola, si ha che se gli items con ID 32078795 e 32842551 vengono acquistati, allora si ha una probabilità pari al 91.8% che l'item con ID 32079103 venga acquistato anch'esso.

GRAZIE PER L'ATTENZIONE