UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

DIPARTIMENTO DI SCIENZE POLITICHE

CORSO DI LAUREA MAGISTRALE IN SCIENZE STATISTICHE PER LE DECISIONI

TESI DI LAUREA

IN

METODI STATISTICI PER DATI COMPLESSI

Customer Churn Prediction e analisi del Lifetime Value: un'applicazione delle Random Survival Forests.

Customer Churn Prediction and Lifetime Value analysis: an application of Random Survival Forests.

Relatore

Prof. Giancarlo Ragozini

Correlatore

Prof. Domenico Vistocco

Candidato

Flavio Canonico

Matr. M10000235

ANNO ACCADEMICO 2018-2019

INDICE

Introduzione

Il mercato dei servizi energetici e delle telecomunicazioni è caratterizzato da una crescente concorrenza a causa della liberalizzazione del mercato[1]. Per le aziende che operano in questo settore appare determinante adottare delle strategie efficaci per trattenere i propri clienti, differenziandosi attraverso delle modalità di azione mirate. Adottare delle specifiche metodologie statistiche diventa cruciale nello studio del fenomeno dell'abbandono dei clienti (*customer churn*). In questo modo si può stimare la probabilità di abbandono dei clienti e individuare delle strategie che mirino alla sua minimizzazione, ottenendo un grosso vantaggio per l'azienda.

Questo lavoro di tesi mira ad illustrare un progetto di data science portato avanti in una nota società che opera principalmente nel settore energetico e delle telecomunicazioni. Il progetto è stato svolto in collaborazione con il team di Business Intelligence, Customer Operation Analytics & Big Data nel settore Operation della società, di cui chi scrive è attualmente uno dei tre componenti. Il progetto in tutte le sue parti è durato diversi mesi e non riguarda esclusivamente le analisi statistico-matematiche che saranno l'oggetto di questo lavoro di tesi, seppur esse costituiscono il fulcro dell'intero progetto. Le diverse fasi del progetto verranno discusse nel corso dell'elaborato.

Gli scopi dell'analisi sono diversi ma tutti strettamente collegati. Il primo obiettivo è quello di stimare la probabilità di abbandono (*churn probability*) dei clienti dell'azienda in un arco temporale che copre 42 mesi. A partire da questo risultato si otterrà una stima del tempo di permanenza di ciascun cliente (*lifetime*), che permetterà il calcolo del valore economico potenziale (*lifetime value*) dei clienti attuali per l'azienda. Questa misura

[1] Si vedano, ad esempio, i dati pubblicati sul sito dell'Autorità di Regolazione per Energia Reti e Ambiente (ARERA) e su quello dell'Autorità per le Garanzie nelle Comunicazioni (AGCOM).

sarà utilizzata per segmentare il parco clienti e adottare delle specifiche strategie nell'ambito della *customer care* e del marketing aziendale, con lo scopo di aumentare la *fidelity* (riducendo quindi il *churn rate*) e massimizzare il *lifetime value* per i singoli clienti e per l'intero parco clienti.

Per portare a termine gli obiettivi indicati è fatto utilizzo dell'analisi statistica della sopravvivenza (tra i moltissimi contributi si veda Cox, Oakes, 2018, Collett, 2013) . Questa viene utilizzata nell'accezione di analisi dei fenomeni *time-to-event*, laddove l'evento, detto *terminale*, è rappresentato dal momento in cui il cliente cessa i propri rapporti commerciali con la società. Nello specifico la metodologia impiegata è quella delle *Random Survival Forest* (Ishwaran et. al, 2008). Si tratta di una estensione della metodologia *Random Forest* (Breiman, 2001) per dati di sopravvivenza.

Andando più nello specifico, nel primo capitolo verrà trattata l'analisi statistica della sopravvivenza nelle sue principali declinazioni. Verranno discussi i concetti fondamentali dell'analisi della sopravvivenza, con particolare attenzione al concetto di tempo di sopravvivenza. Verrà fatta poi una breve rassegna dei principali modelli parametrici e semi-parametrici per modellare i fenomeni *time-to-event*. Si passerà poi a una trattazione più approfondita delle metodologie non parametriche, facendo riferimento al metodo delle curve di Kaplan e Meier (Kaplan e Meier, 1958) e al log-rank test (Mantel e Haenszel, 1959), per poi passare alla trattazione dell'algoritmo *Random Survival Forest* (RSF) (Ishwaran et. al, 2008). Nel trattare le RSF verrà fatta una breve descrizione della metodologia *Random Forest* (Breiman, 2001), poiché come accennato quella RSF ne rappresenta un'estensione.

Nel secondo capitolo si descriverà il contesto aziendale di riferimento e verranno discusse le principali strategie adottate nella costruzione della base dati che verrà poi utilizzata nell'analisi. Si farà

riferimento alle tecnologie utilizzate e verrà fatta una descrizione delle variabili che vanno a costituire la base dati stessa. Inoltre un paragrafo sarà dedicato alla discussione le precauzioni adottate per la tutela della privacy dei soggetti coinvolti nello studio e per il rispetto degli accordi contrattuali che legano chi scrive all'azienda. Si farà poi riferimento alle risorse *hardware* a disposizione e che sono state necessarie per portare a termine lo studio. Infine verrà fatta una descrizione d'insieme del modo in cui l'analisi verrà inserita nel progetto aziendale generale, si spiegherà a grandi linee come verrà portata a termine la messa in produzione del modello e i relativi risultati nei sistemi aziendali.

Nel terzo capitolo si passerà alla discussione dell'analisi statistica vera e propria. Verrà fatto un breve riferimento alle librerie R utilizzate per poi passare alla parte delle operazioni preliminari sulle variabili presenti nella base dati (*data preprocessing*). Dopo una descrizione della distribuzione della (o meglio delle) variabili di risposta si passerà a discutere le strategie di costruzione della RSF e i risultati ottenuti in termini di performance. Nel paragrafo successivo verranno discussi i risultati facendo riferimento alle misure di Variable Importance. Verrà poi trattata la costruzione della stima del tempo di permanenza dei clienti (*lifetime*) e della stima del valore economico potenziale (*lifetime value*), analizzandone le distribuzioni. Una volta esposti i risultati più strettamente tecnico-statistici, si proverà a fare un'analisi dei risultati che fornisca strumenti interpretativi ai vari *stakeholders*. Nello specifico verranno in primo luogo studiate le relazioni tra le variabili indipendenti inserite nella RSF e il tempo di permanenza stimato. Nell'ultimo paragrafo si discuterà delle strategie di segmentazione del parco clienti sulla base del *lifetime value* stimato.

Infine verranno esposte le conclusioni a cui si è giunti attraverso l'analisi, rimarcando gli obiettivi preposti ed evidenziando eventuali criticità ed esigenze.

In appendice verrà riportata una descrizione di tutte le variabili inserite nella RSF che, a causa della loro numerosità, si è deciso di non riportare direttamente nel corpo principale di questo lavoro di tesi. Verrà infine riportato per esteso il codice R utilizzato per l'analisi.

CAPITOLO I

L'ANALISI DELLA SOPRAVVIVENZA

SOMMARIO: 1.1 L'analisi statistica della sopravvivenza – 1.2 Tempo di sopravvivenza – 1.3 Modelli parametrici e semi-parametrici – 1.4 Stimatore di Kaplan-Meier – 1.5 Log-rank test – 1.6 Random Survival Forest – 1.6.1 Random Forest – 1.6.2 Random Survival Forest (RSF): l'algoritmo – 1.6.3 Predizione per il nodo terminale – 1.6.4 Errore di previsione – 1.6.5 Variable importance (VIMP) – 1.6.6 Dati mancanti

1.1 L'analisi statistica della sopravvivenza

Con il termine Analisi della Sopravvivenza si fa riferimento ad una Tecnica statistica di analisi di dati, ottenuti da una coorte di unita osservate longitudinalmente, che consente di stimare la probabilità del verificarsi di un determinato evento in funzione del tempo (tra i molti manuali disponibili, si veda ad esempio Cox, Oakes, 2018, Collet, 2013 o Lawless, 2002). Nelle prime applicazioni l'evento in studio era quasi sempre la morte del paziente, tanto è vero che si parla spesso di dati di sopravvivenza. Successivamente questo termine venne utilizzato per tutti i tipi di eventi in studio. In ogni caso l'evento può essere considerato come una transizione da uno stato ad un altro.

I campi di applicazione sono i più disparati, come medicina, sanità pubblica, scienze sociali, ingegneria, e di recente anche in ambito aziendale. In medicina il tempo prima che si verifichi l'evento terminale può essere il tempo fino alla morte del paziente o il tempo fino al verificarsi di un'infezione (Collett, 2013). Nelle scienze sociali si può considerare il tempo fino a che non si verifica un evento di rilevanza per un individuo o un'intera comunità, come il cambiamento di un lavoro, la nascita di un bambino e così via. In ingegneria si può valutare il tempo finché non si verifica un guasto di un macchinario. Non per ultimo in campo aziendale si può utilizzare per studiare il tempo fino alla risoluzione di un ticket, il tempo fino alle dimissioni di un dipendente, il tempo finché un cliente rescinde un contratto o un abbonamento, e così via. Questi fenomeni sono stati di forte interesse nelle scienze statistiche e attuariali e hanno portato allo sviluppo di un intero campo di studi detto spesso proprio analisi della sopravvivenza.

Da questo punto in poi si parlerà indistintamente di morte di un individuo ed evento terminale, così come di tempo di sopravvivenza piuttosto che di tempo prima che si verifichi l'evento terminale per un'osservazione, anche se questo lavoro non ha nulla a che fare con gli

studi sulla morte biologica. Seppur queste espressioni potrebbero apparire fuori luogo in un lavoro in cui si studia un fenomeno che ha a che fare con la conclusione di un contratto, verranno usate solo per semplicità e continuità con i termini classicamente utilizzati nel campo degli studi sulla sopravvivenza.

Durante il periodo di osservazione, però, solo alcuni individui sperimentano l'evento terminale, mentre altri soggetti alla fine del *follow up* non hanno ancora sperimentato l'evento, pertanto non si conosce il loro tempo di sopravvivenza. Questi dati vengono detti censurati (Collett, 2013). Il tipo di censura più diffusa negli studi di sopravvivenza è la censura a destra. Esistono anche casi in cui si ha a che fare con dati censurati a sinistra o censurati in un intervallo, tuttavia sono meno diffusi. La censura a destra può verificarsi per due ragioni principali: nel primo caso il soggetto in analisi si ritira dallo studio o non si può più osservare il suo stato prima della fine del periodo di osservazione; nel secondo caso per il soggetto in osservazione non si verifica l'evento terminale entro la fine dello studio. In entrambi i casi non sarà possibile sapere quando si verifica l'evento terminale per il soggetto in questione. Chiaramente con i metodi tradizionali non è possibile analizzare dati del genere, invece il grosso vantaggio dell'analisi della sopravvivenza è quello di riuscire a trattare anche casi in cui non si conosce il tempo di sopravvivenza finale (Collett, 2013).

Le principali metodologie statistiche sviluppare per modellare questi fenomeni *time-to-event* possono essere divise in tre grandi categorie:

- Modelli parametrici
- Modelli semi-parametrici
- Modelli non parametrici

In questo lavoro si utilizzerà un approccio non parametrico sviluppato recentemente. Si tratta della metodologia *Random Survival Forest* (Ishwaran et al., 2008) di cui si parlerà ampiamente nel corso di questo capitolo.

Nel discutere le specificità dell'analisi della sopravvivenza, si farà invece solo un breve cenno ad alcuni dei principali modelli di sopravvivenza parametrici e semi-parametrici in quanto la loro trattazione approfondita esula dagli scopi di questo lavoro di tesi.

1.2 Tempo di sopravvivenza

Nell'analizzare i dati di sopravvivenza le funzioni di interesse centrale sono: la funzione di sopravvivenza, la funzione di densità di probabilità e la funzione di rischio (*hazard function* o HF) (per una trattazione approfondita, si veda ad esempio Cox, Oakes, 2018).

L'effettivo tempo *t* di sopravvivenza di un individuo, può essere considerato come il valore di una variabile aleatoria *T*, che assume solo valori non negativi. Si supponga che la variabile casuale *T* abbia una distribuzione di probabilità *F*(*t*) con funzione di densità di probabilità *f*(*t*). La distribuzione di probabilità di *T* e data da:

$$F(t) = Pr(T < t) = \int_0^t f(u)du \tag{1.1}$$

e rappresenta la probabilità che il tempo di sopravvivenza sia inferiore ad un dato valore *t*. La funzione di sopravvivenza, *S*(*t*), e definita come la probabilità che il tempo di sopravvivenza *T* sia maggiore o uguale al valore *t*:

$$S(t) = P(T \geq t) = 1 - F(t) = 1 - P(T < t) \tag{1.2}$$

La funzione di sopravvivenza può quindi essere usata per rappresentare la probabilità che un individuo sopravviva oltre il tempo *t*. La funzione di rischio è la probabilità che un individuo muoia al tempo *t*, condizionata al fatto che sia sopravvissuto fino a quell'istante:

$$\lambda(t) = \lim_{\Delta t \to 0} \left\{ \frac{(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \right\} \qquad (1.3)$$

La funzione di rischio quindi rappresenta la probabilità istantanea dell'evento in studio per un individuo che non ha sperimentato l'evento fino al tempo *t*. C'è una relazione tra la funzione di sopravvivenza *S*(*t*) e la funzione di rischio λ(t), espressa dalla formula seguente:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log S(t) \qquad (1.4)$$

$$S(t) = \exp\left[-\int_0^t \lambda(u) du \right] = \exp[-\Lambda(t)], \quad t \geq 0 \qquad (1.5)$$

dove $\Lambda(t) = \int_0^t \lambda(u) du$ è detta funzione di rischio cumulato (*Cumulative Hazard Function* o CHF), la quale può essere ottenuta dalla funzione di sopravvivenza poiché $\Lambda(t) = -log\, S(t)$. E la funzione di densità di probabilità di *T* si può scrivere come:

$$f(t) = \lambda(t) \exp\left[-\int_0^t \lambda(u) du \right], \quad t \geq 0 \qquad (1.6)$$

Queste tre funzioni forniscono una specificazione equivalente della distribuzione del tempo di sopravvivenza *T* e conscendo una delle funzioni si possono determinare le altre. È sufficiente, quindi, sceglierne una come base per l'analisi statistica a seconda degli obiettivi finali dello studio. Ad esempio la funzione di sopravvivenza è maggiormente utile per confrontare la sopravvivenza di due o più gruppi di soggetti, mentre la funzione di

rischio fornisce una descrizione del rischio che si verifichi l'evento terminale per ognuno degli istanti temporali osservati.

1.3 Modelli parametrici e semi-parametrici

I modelli di regressione parametrici vengono usati quando si assume una distribuzione teorica per F(t), con $t_i, i = 1, \dots, n$ tempi di sopravvivenza considerati (Collett, 2013). Di solito vengono distinti in base al comportamento che assume la loro funzione di rischio. I metodi parametrici per l'analisi dei dati di sopravvivenza consentono la stima della funzione di sopravvivenza e della funzione di rischio, adattando ai dati una funzione di cui si assume una certa forma e che dipende da uno o più parametri. Naturalmente è necessario tenere presente che la loro applicabilità è limitata dal fatto che la funzione di rischio assuma una ben definita forma matematica, mentre in molte situazioni pratiche non ci sono sufficienti giustificazioni per l'adozione di un particolare modello. Come per la regressione lineare si procede alla stima dei coefficienti di regressione, che in questo caso stimeranno l'effetto esercitato da ciascuno dei *p* regressori inseriti nel modello sul rischio che si verifichi l'evento terminale.

Il *modello esponenziale* presuppone che il rischio istantaneo di morte non vari al trascorrere del tempo (Collett, 2013). Questo implica, ad esempio, che il rischio di abbandono nei primi mesi in cui il cliente sottoscrive il contratto è uguale a quello che ha al terzo anno dalla contrattualizzazione. Chiaramente questa assunzione non può essere sostenuta. Infatti nella fase di acquisizione del cliente e attivazione delle utenze tendono a verificarsi maggiori criticità, per cui il rischio di abbandono dopo questo periodo risulta essere inferiore.

Il *modello di Weibull* presuppone, invece, che la funzione rischio possa essere costante nel tempo, monotona crescente o monotona decrescente, a seconda del valore assunto dal parametro di forma α della distribuzione (Collett, 2013).

Il *modello log-normale* è l'unico ad avere una funzione rischio non monotona, ma che cresce dal valore iniziale nullo (t=0) fino a raggiungere un punto di massimo, per poi decrescere verso zero (per $t \to \infty$) (Collett, 2013).

Questi modelli possono tutti essere ricondotti a una classe di modelli caratterizzati dal presupporre la proporzionalità della funzione di rischio (*proportional hazard models*). Infatti le loro funzioni di rischio (qui non esplicitate per evitare di dover entrare in una trattazione troppo approfondita) possono essere scomposte in due fattori, di cui uno dipendente solo dal tempo e uno solo dalle variabili indipendenti:

$$\lambda(t; x) = \lambda_0(t)h(x) \tag{1.7}$$

In base alle differenti forme parametriche assunte da $\lambda_0(t)$ e $h(x)$ si ottengono diversi modelli.

I *modelli a tempi di evento accelerati* o modelli AFT (*Accelerated Failure Time*): sono modelli generali per dati di sopravvivenza, in cui si assume che le variabili esplicative misurate su un soggetto agiscano moltiplicativamente sulla scala temporale e così hanno il ruolo di aumentare o diminuire la velocita con cui un individuo procede lungo l'asse dei tempi, accelerando o rallentando il verificarsi dell'evento terminale (Collett, 2013). Un modello AFT può essere scritto come:

$$\log T_i = x_i^T \beta + \alpha\epsilon \tag{1.8}$$

in cui α $(con\ \alpha > 0)$ è un parametro di scala e ϵ è una variabile casuale che si assume abbia una particolare distribuzione. Per ogni distribuzione di ϵ si ha una corrispondente distribuzione di *T.* Ad esempio se si assume che ϵ si distribuisca come una v.c. normale, *T* si distribuirà come una v.c. log-normale. Si tratta di modelli più "flessibili" in situazioni reali rispetto a

quelli parametrici classici, tuttavia si presuppone comunque un effetto moltiplicativo dei regressori sul tempo di sopravvivenza. Il tentativo di utilizzo di questi modelli per l'analisi oggetto di questa tesi portava infatti in una parte dei casi a risultati troppo ottimistici nella stima del tempo prima che un cliente rescinda il proprio contratto, superando ampiamente il range di mesi verosimile prima del *churn* del cliente.

I modelli parametrici consentono di perseguire due obiettivi simultaneamente. Il modello deve descrivere la distribuzione sottostante il tempo di sopravvivenza (componente dell'errore), ma deve anche caratterizzare i cambiamenti della distribuzione in funzione delle esplicative (componente sistematica). In alcuni casi però è sufficiente che un modello persegua soltanto il secondo degli obiettivi descritti, ad esempio si vuole studiare se una combinazione di variabili migliora il tempo di sopravvivenza di un paziente con una determinata patologia. Lo scopo quindi non è la descrizione del tempo di sopravvivenza, ma la descrizione dell'impatto delle variabili indipendenti sulla sopravvivenza del paziente. Si potrebbero quindi evitare tutte le assunzioni stringenti richieste dai modelli parametrici sulla componente erratica, poiché l'inferenza riguarderà soltanto i parametri della parte sistematica del modello. Per queste ragioni sono stati introdotti i modelli semi-parametrici.

Il *modello di Cox* (Cox, 1972; Cox, Oakes, 2018) si colloca tra le procedure semi-parametriche. Infatti non assume alcuna forma specifica circa la distribuzione sulla forma della funzione di sopravvivenza, ma modella l'effetto delle variabili indipendenti sulla sopravvivenza in modo parametrico. In altre parole viene specificata solo la forma funzionale per valutare l'influenza delle covariate sui tempi di sopravvivenza e la forma dell'andamento del rischio viene lasciata non specificata. Si tratta del modello più diffuso nella classe dei modelli a rischi proporzionali. Così come nella 1.7 si può scrivere la funzione di rischio come funzione del tempo e delle variabili indipendenti. La funzione $\lambda_0(t)$ è la funzione di

rischio quando $h(x, \beta) = 1$ ed è detta funzione *baseline hazard*. Essa indica come il rischio cambia in funzione del tempo ma non dipende dal vettore delle variabili esplicative. Invece $h(x, \beta)$ indica come la funzione di rischio cambia in funzione delle variabili esplicative, ma non dipende dal tempo. Nel modello di Cox si definisce $h(x, \beta) = \exp(x\beta)$, per cui la funzione di rischio diventa:

$$\lambda(t, x, \beta) = \lambda_0(t) \exp(x\beta) \tag{1.9}$$

Che può essere ovviamente ripensata anche nella sua forma log-lineare. La stima dei coefficienti di regressione di questo modello permette di studiare l'effetto moltiplicativo delle variabili indipendenti sulla differenza di sopravvivenza a un certo tempo *t*. Si potrà quindi dire, ad esempio, che per un certo valore delle variabili indipendenti la probabilità di sopravvivenza del soggetto *i* è maggiore di una certa percentuale di quella del soggetto *j* ad un certo tempo *t* considerato. Per quanto utile non è quindi adatto a perseguire gli scopi di questa analisi.

In conclusione tutti questi modelli si sono ritenuti meno adatti di una formulazione non parametrica. Lo scopo infatti è quello di adottare una soluzione di *machine learning* il più flessibile possibile. Tra i motivi principali per i quali si è optato per un algoritmo Random Survival Forest c'è l'esigenza di una metodologia all'avanguardia che possa gestire in modo efficace un grosso numero di covariate, che sia robusta e in grado di tenere conto delle complesse interazioni tra le variabili e della forma non lineare delle relazioni tra queste ultime e la variabile di risposta. I principali svantaggi invece hanno a che fare con la maggiore difficoltà di interpretazione dei risultati e con la necessità di una grossa potenza computazionale, che ha reso necessario in alcuni casi il ricorso al calcolo distribuito e alla saturazione di molta memoria della macchina utilizzata.

1.4 Stimatore di Kaplan-Meier

Il metodo non parametrico maggiormente diffuso nell'analisi della sopravvivenza è quello di Kaplan - Meier (Kaplan e Meier, 1958). Questo criterio è spesso definito per piccoli campioni in quanto sfrutta meglio l'informazione disponibile. Consiste nello stimare la probabilità condizionata di sopravvivenza in corrispondenza di ciascuno dei tempi in cui si verifica almeno un evento terminale. Esso non presuppone la suddivisione dell'asse temporale in intervalli di ampiezza predefinita.

Si supponga di avere n soggetti e di osservare i tempi in cui si verifica almeno un evento terminale, in ordine crescente, per gli n individui: $t(1) < t(2) < \ldots < t(J)$, con $J \leq n$. Sia d_j il numero di soggetti che presentano l'evento d'interesse al tempo $t(j)$, per $j = 1,\ldots, J$. Nel caso in cui d_j sia maggiore di uno, significa che più soggetti presentano tempi di evento uguali. Sia, inoltre, n_j il numero di individui a rischio al tempo $t(j)$. È possibile, a questo punto, definire la probabilità condizionata che si verifichi l'evento terminale nell'istante $t(j)$, dato che per il soggetto in questione questo non si è ancora verificato all'istante immediatamente precedente. La sua stima è definita come:

$$\hat{q}_j = \frac{d_j}{n_j} \tag{1.10}$$

Quindi la probabilità di sopravvivere al tempo $t(j)$ è data dal complementare di $\hat{q}_j$ ed è stimata da:

$$\hat{p}_j = 1 - \hat{q}_j = \frac{n_j - d_j}{n_j} \tag{1.11}$$

Moltiplicando tra loro le stime delle probabilità condizionate di sopravvivere, si ottiene la stima della probabilità cumulata di sopravvivere oltre l'istante $t(j)$, data da:

$$\widehat{S(t)} = \prod_{j|t(j)\leq t} \hat{p}_j \qquad (1.12)$$

Che è detto stimatore di Kaplan-Meier. Dal momento che gli istanti di tempo, in cui non si verifica l'evento terminale, assumono probabilità pari a 1, si ha che lo stimatore di Kaplan-Meier coinvolge solo i tempi *t* in cui si verifica almeno un evento terminale. Esso è non distorto ed è asintoticamente normale.

La stima di Kaplan-Meier della probabilità di sopravvivenza viene rappresentata con una curva a gradini decrescente. Essa parte da 1 in quanto la probabilità di sopravvivenza al tempo $t(0)$ della coorte considerata è certa dato che non si è ancora verificato nessun evento terminale. L'altezza dei gradini dipende dal numero di eventi e dal numero dei soggetti a rischio.

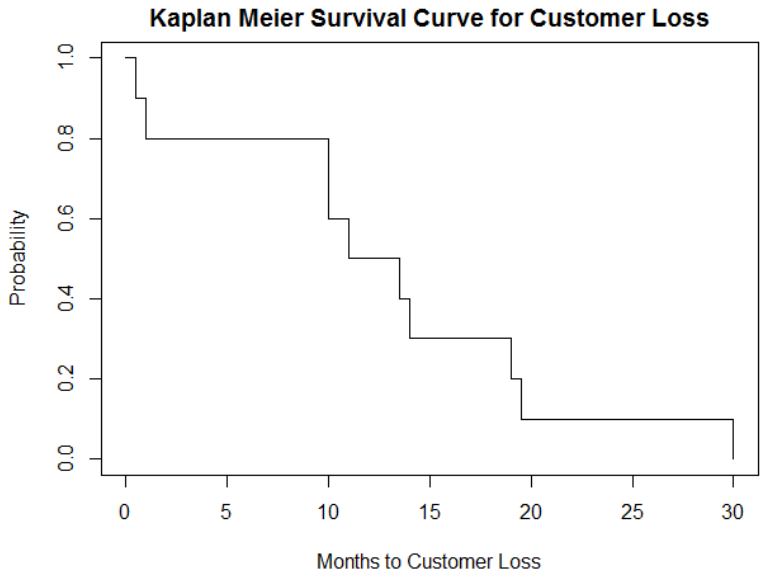

Figura 1 - esempio di curva di Kaplan-Meier

1.5 Log-rank test

È necessario anche un approfondimento sul log-rank test, in quanto anch'esso sarà usato nella metodologia cardine di questo lavoro di tesi. Il log-rank test, proposto da Mantel nel 1966 come estensione del test di Mantel e Haenszel (Mantel e Haenszel, 1959), è un test non parametrico, basato sui ranghi, che permette di confrontare curve di sopravvivenza di due (o più) insiemi di soggetti sottoposti a differenti trattamenti o esposti a differenti rischi. Non è necessario quindi ipotizzare la forma delle sottostanti funzioni teoriche di sopravvivenza. Si costruisce calcolando il numero osservato e quello atteso degli eventi in uno dei gruppi in ciascun momento degli eventi osservati e poi sommando questi ultimi per ottenere una sintesi complessiva lungo tutti i punti temporali in cui si verifica un evento.

Sia N il numero di soggetti sotto osservazione. Si supponga di aver attribuito con una procedura casuale (randomizzazione) la metà dei pazienti al trattamento A e l'altra metà al trattamento B. Si otterranno due insiemi di soggetti, di numerosità rispettivamente n_A ed n_B.

Il log-rank test viene utilizzato saggiare l'ipotesi nulla $H_0 : \lambda_A(t) = \lambda_B(t)$, per ogni istante t. Si vuole, pertanto, verificare l'uguaglianza delle probabilità di sopravvivenza sui due differenti gruppi sottoposti ai diversi trattamenti. Siano j=1,…,J i distinti tempi degli eventi osservati in ciascuno dei due gruppi e d_{Aj} il numero di pazienti del gruppo A per i quali si è verificato l'evento terminale al tempo $t(j)$. Sotto l'ipotesi nulla, la distribuzione di d_{Aj} risulta essere ipergeometrica, per cui essa ha media e varianza date, rispettivamente, da:

$$E(d_{Aj}) = n_{Aj}.\frac{d_j}{n_j} \tag{1.13}$$

e

$$var(d_{Aj}) = \left[n_{Aj} \cdot \frac{d_j}{n_j}\left(1 - \frac{d_j}{n_j}\right)\right]\left(\frac{n_j - n_{Aj}}{n_j - 1}\right) \tag{1.14}$$

La statistica test quindi è data da:

$$Q = \frac{\sum_{j=1}^{J}\left[d_{Aj} - E(d_{Aj})\right]^2}{\sum_{j=1}^{J} Var(d_{Aj})} \qquad (1.15)$$

Che asintoticamente è distribuita come una v.c. χ^2 con un grado di libertà. Si opera quindi un confronto tra il numero di eventi terminali osservati nei vari tempi *t*(*j*) e quelli attesi sotto H_0. Si rifiuta l'ipotesi nulla di uguaglianza delle curve di sopravvivenza per valori elevati di Q.

1.6 *Random Survival Forest*[2]

Sono state introdotte per la prima volta da Ishwaran e Kogalur in un articolo nel 2007 (Ishwaran, Kogalur, 2007), i quali hanno provveduto anche allo sviluppo di un apposito pacchetto per l'implementazione in R (Ishwaran, Kogalur, 2008). Questa metodologia non parametrica ha tutta una serie di vantaggi, tra i quali:

- Non sono necessarie assunzioni distribuzionali né assunzioni di altro tipo e si possono utilizzare efficacemente predittori discreti, continui e variabili categoriali.
- È robusta nel caso ci siano molte variabili di rumore, quindi riesce a estrarre efficacemente i pattern significativi tra la moltitudine delle variabili inserite e non è troppo sensibile all'inserimento o all'esclusione di singole variabili dal modello.
- È robusta rispetto al problema dell'*overfitting*, per cui si possono inserire un gran numero di variabili indipendenti.

[2] Per la trattazione degli aspetti metodologici delle Random Survival Forest in questo paragrafo il riferimento principale è stato Ishwaran et al., 2008.

- Riesce a tenere conto di un gran numero di interazioni tra le variabili e riesce a modellare efficacemente relazioni non lineari e molto complesse tra le variabili.
- Grazie alla tecnica del *bagging* (che verrà rapidamente trattata oltre) tutti i dati a disposizione possono essere utilizzati per l'addestramento, non essendo necessario escludere una parte del dataset per la fase di test.

Le RSF sono un'estensione della metodologia Random Forest (RF) (Breiman, 2001) per l'analisi dei dati di sopravvivenza. È necessario quindi una breve discussione della metodologia RF e dei concetti fondamentali a essa connessi.

1.6.1 Random Forest

Si tratta di una metodologia di *machine learning* che si sta diffondendo in modo sempre più marcato nel campo dell'analisi dei dati. Questo perché permette di raggiungere un'elevata accuratezza in una vasta gamma di applicazioni. Le RF utilizzano un insieme di alberi decisionali o alberi di regressione per risolvere i più svariati problemi di classificazione o regressione. Di solito gli alberi sono costruiti utilizzando la metodologia CART (Breiman et al., 1984). Le RF si basano sulla stima di una moltitudine di alberi binari (in genere dalle centinaia alle migliaia) e sulla tecnica del *bagging*. Infatti gli alberi decisionali, se sono troppo grandi, hanno varianza elevata. Se sono troppo piccoli hanno *bias* elevato. Per provare a risolvere il problema, si utilizza una tecnica definita *bagging* (*bootstrap aggregating*) che fa utilizzo della tecnica del bootstrap.

Il *bootstrap* è una tecnica statistica di ricampionamento per approssimare la distribuzione campionaria di una statistica. Permette cioè, di approssimare media e varianza di uno stimatore, costruire intervalli di confidenza e calcolare *p-value* di test statistici in particolare quando non si conosce la distribuzione di uno stimatore. I dati di training sono

ricampionati *B* volte con *B* nell'ordine delle migliaia. Se si hanno *n* unità nei dati di training il campionamento prevede di selezionarne *n* a caso. Per evitare di ottenere sempre lo stesso campione il campionamento è con ripetizione. A questo punto si avranno a disposizione molti dataset, tutti di dimensione *n* da utilizzare per stimare il modello. Un secondo vantaggio del *bagging*, direttamente dipendente dal fatto di avere a disposizione molti set di dati è che si ottiene una riduzione della varianza della risposta. Infatti i risultati sono calcolati come media dei risultati ottenuti dai molteplici dataset ottenuti tramite bootstrap. La varianza si riduce perché intuitivamente se si ha un insieme di *n* osservazioni indipendenti $Y_1, \ldots, Y_n$ ognuna con varianza σ^2, la varianza della media $\bar{Y}$ delle osservazioni è $\frac{\sigma^2}{n}$. Detto in altri termini i campioni bootstrap avranno varianza osservata s^2 differente, ma sono generati dagli stessi dati, per cui sono tutte realizzazioni di v.c. indipendenti e identicamente distribuite (i.i.d.). Di conseguenza le v.c. che le generano hanno tutte la stessa varianza σ^2. Ora, la v.c. media campionaria generata da queste v.c. i.i.d. per definizione avrà varianza uguale a $\frac{\sigma^2}{n}$ per cui ne risulta una riduzione della varianza.

Riassumendo il bagging è strutturato sulle seguenti fasi: vengono generati *B* dataset di training tramite bootstrap; su ogni dataset viene stimato un albero di regressione o classificazione; viene effettuata una previsione attraverso la media nel caso di regressione. Nel caso di problemi di classificazione si possono usare due approcci per ottenere la previsione: viene registrata la classe che ogni dataset bootstrap predice e si fornisce una previsione aggregata usando la classe più frequente tra le *B* previsioni; se il classificatore produce stime di probabilità si può calcolare la media delle *B* probabilità e si sceglie la classe con probabilità media più elevata.

Dato che il bootstrap implica la selezione casuale di *sottoinsiemi* di osservazioni dal dataset originario, per ogni campione bootstrap le unità non selezionate possono essere utilizzate come unità di test. Si parla infatti

di stime *Out-Of-Bag* (OOB), poiché ogni albero *bagged* si avvale di circa 2/3 delle osservazioni totali, così il restante 1/3 può essere utilizzato direttamente per fare la *cross-validation*. Il grosso vantaggio è quello di seguire la procedura diffusa nel *machine learning* di validazione mediante split del dataset originario in *training* e *test set* senza dover rinunciare a parte del dataset nella fase di training. Inoltre la stima del tasso di errata classificazione è computata in modo più robusto come media su tutti gli alberi della foresta.

Nelle RF viene poi inserito un secondo elemento di randomizzazione. Ad ogni split nella costruzione di ognuno degli alberi, solo un campione casuale di *m* predittori è utilizzato per decidere lo split, di solito $m \sim \sqrt{p}$ oppure $m = p/3$. Il motivo è che se si ha un predittore (o un gruppo di predittori) molto importante si rischia che nella maggior parte degli alberi questo venga utilizzato nel primo split, per cui tutti gli alberi tenderanno ad essere simili. Di conseguenza tutte le previsioni saranno altamente correlate. Invece, utilizzando ad ogni passo un sottoinsieme casuale di predittori, si otterranno alberi diversi tra loro e quindi de-correlati. In questo modo si possono sfruttare i vantaggi derivanti dalla stima di un grande numero di alberi di regressione o classificazione.

1.6.2 Random Survival Forest (RSF): l'algoritmo

Le principali fasi dell'algoritmo possono essere riassunte nei seguenti punti:

- Vengono costruiti *B* campioni bootstrap dai dati originari. Ogni campione bootstrap in media esclude il 37% delle osservazioni, che andranno a formare i dati *out-of-bag*.
- Per ogni campione bootstrap viene addestrato un albero di sopravvivenza. A ogni nodo dell'albero vengono selezionate *p* variabili in modo casuale. Lo split di ogni nodo avviene utilizzando tra le variabili candidate quella che massimizza la differenza di sopravvivenza tra i nodi figli.

- L’albero viene fatto crescere fino alla grandezza massima sotto il vincolo secondo cui un nodo terminale non deve avere meno di $d_0 > 0$ morti.
- Si calcola la funzione cumulata di rischio (*cumulative hazard function* o CHF) per ogni albero, la cui media permette il calcolo della CHF generale.
- Usando i dati OOB, si calcola l’*error rate* per la CHF generale.

Così come nella metodologia CART, gli alberi di sopravvivenza sono alberi binari ottenuti attraverso split ricorsivi dei nodi dell’albero. Il primo nodo contiene tutte le osservazioni ed è detto nodo radice. Quest’ultimo viene suddiviso in due nodi figli utilizzando un certo criterio proprio dell’analisi della sopravvivenza. A loro volta i nodi figli vengono suddivisi nei rispettivi nodi figli di destra e sinistra. Il processo si ripete ricorsivamente fino al raggiungimento di un certo criterio di stop.

Lo split è ricercato secondo il criterio della massimizzazione della differenza nello *score* di sopravvivenza tra i nodi figlio. Nello specifico lo split migliore per un nodo viene individuato cercando tra tutte le variabili *x* a disposizione e tutti i possibili valori di split *c* (o categorie nel caso di variabili fattoriali) e selezionando quella variabile x^* e quello split c^* che massimizza la differenza nello *score* di sopravvivenza. In questo modo i nodi figlio diventano sempre più omogenei al proprio interno in termini di caratteristiche e relativo score di sopravvivenza e diversi all’esterno. Di solito il criterio di default utilizzato per lo split è quello della massimizzazione della statistica del log-rank test.

1.6.3 Predizione per il nodo terminale

Siano $t_{1,h} < t_{2,h} < \cdots < t_{N(h),h}$ gli N(h) tempi distinti per gli eventi. Siano $d_{l,h}$ e $Y_{l,h}$ il numero di eventi terminali e gli individui a rischio al tempo $t_{l,h}$. Siano τ l’insieme dei nodi terminali e h un generico nodo

terminale di un albero di sopravvivenza. La funzione cumulata di rischio stimata per *h* è la stima di Nelson-Aalen:

$$\widehat{H}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}} \tag{1.16}$$

Tutti i soggetti nel nodo terminale *h* hanno la stessa CHF.

Ogni osservazione *i* è caratterizzata da un vettore *d*-dimensionale di covariate x_i. Sia $H(t|x_i)$ la CHF dell'osservazione *i*. Per determinare questo valore si fa scorrere x_i lungo l'albero. Essendo un albero binario x_i finirà in un solo nodo terminale $h \in \tau$. La CHF per *i* è lo stimatore di Nelson-Aalen per il nodo terminale in cui giunge x_i:

$$H(t|x_i) = \widehat{H}_h(t) \qquad \text{se } x_i \in h \tag{1.17}$$

Per calcolare la CHF generale è necessario fare una media su tutti i *B* alberi di sopravvivenza. Sia $I_{i,b} = 1$ se *i* è un osservazione OOB per l'albero *b* e $I_{i,b} = 0$ altrimenti e sia $H_b^*(t|\boldsymbol{x})$ la CHF per un albero addestrato con il *b*-esimo campione bootstrap. La CHF generale (ensemble CHF) della foresta per l'*i*-esima osservazione è:

$$H_e^{**}(t|x_i) = \sum_{b=1}^{B} I_{i,b} H_b^*(t|x_i) \tag{1.18}$$

Si tratta di una media calcolata per i campioni bootstrap in cui *i* è un OOB. In altre parole in questo modo si ottiene la probabilità di sopravvivenza al tempo *t* per *i*, calcolata come media delle probabilità ottenute facendo scivolare *i* nei soli alberi di sopravvivenza in cui *i* è OOB e quindi non è parte del train set. Alternativamente è possibile calcolare la CHF generale su tutti gli alberi di sopravvivenza stimati:

$$H_e^*(t|x_i) = \frac{1}{B}\sum_{b=1}^{B} H_b^*(t|x_i) \tag{1.19}$$

Si noti che nella formula vengono utilizzati tutti gli alberi di sopravvivenza e non i soli in cui *i* è OOB. In questo modo è possibile stimare la probabilità che si verifichi un evento per ognuno dei tempi *t* considerati.

Può essere utile un piccolo accenno al principio di conservazione degli eventi. Secondo questo principio, in condizioni di regolarità la somma dei valori stimati dalla CHF a un certo tempo *t* è uguale al numero di morti stimate per quel tempo *t*. In questo modo si può stimare ad esempio il numero di clienti per i quali si registrerà il *churn* a un certo tempo *t*.

1.6.4 Errore di previsione

Per la stima dell'errore di previsione viene utilizzato l'indice di concordanza di Harrell (Harrell et al., 1982), noto come indice C. Il calcolo dell'indice C necessita la stima di un valore predetto. Per ottenerlo si usa la stima della CHF generale calcolata sulle osservazioni OOB. Di conseguenza l'indice C viene calcolato sulle osservazioni OOB, cosa che ha senso in quanto si tratta di un tasso di errore e avrebbe poco senso stimare usando le previsioni ottenute sulle osservazioni del training set.

Inoltre, considerata una coppia di osservazioni *i* e *j*, necessita che sia stabilito quale delle due abbia il risultato (*rank*) stimato peggiore. Siano $t_1^o, \dots , t_m^o$ i tempi distinti preselezionati. Si potrà dire che *i* ha un risultato stimato peggiore di *j* se:

$$\sum_{l=1}^{m} H_e^{**}(t_l^o|x_i) > \sum_{l=1}^{m} H_e^{**}\left(t_l^o\middle|x_j\right) \tag{1.20}$$

L'indice è calcolato nel seguente modo:

1. Vengono formate tutte le possibili coppie del dataset.
2. Si escludono le coppie per le quali il tempo più corto è censurato. Le coppie *i* e *j* vengono escluse se $T_i = T_j$ a meno che per uno dei due non si sia verificato l'evento terminale. Il numero totale di coppie che possono essere formate seguendo questi criteri verranno dette *Possibili* (*Permissible*).
3. Per ogni coppia possibile si assegna 1 se il più piccolo tempo di sopravvivenza ha un risultato predetto peggiore e 0.5 se i due risultati di previsione sono uguali. Per ogni coppia in cui $T_i = T_j$ se per entrambe le osservazioni si è verificato l'evento terminale si assegna 1 se i due risultati di previsione sono uguali, altrimenti 0.5. Se invece si è verificato l'evento terminale solo per una delle due osservazioni, si assegna 1 se il soggetto per il quale si è verificato l'evento terminale ha un risultato predetto peggiore, altrimenti 0.5. La somma di questi valori è detta *Concordanza* (*Concordance*).
4. L'indice C è quindi così definito: $C = \frac{Concordance}{Permissible}$.

Il tasso di errore (*Prediction Error* o *Error Rate*) è definito come $PE = 1 - C$ con $0 \leq PE \leq 1$. Per quel che riguarda l'interpretazione, un valore di $PE = 0.5$ può essere interpretato come se la RSF preveda i valori in modo del tutto casuale, paragonabile alla situazione del lancio della moneta. Se $PE = 0$ l'accuratezza previsionale è massima.

1.6.5 Variable importance (VIMP)

Per la selezione delle variabili e l'interpretazione dei risultati bisogna calcolare una misura di importanza delle variabili (VIMP). Per calcolare la VIMP della variabile *x*, si inseriscono le osservazioni OOB nei rispettivi alberi di sopravvivenza. Quando si giunge a uno split fatto sulla variabile *x* in questione, invece di seguire lo split scelto dall'albero originario si

assegnano in modo casuale le osservazioni ai nodi figlio. A questo punto si calcola la CHF generale. La VIMP per *x* è la differenza tra il tasso di errore originario e il tasso di errore calcolato attraverso l'assegnazione casuale delle osservazioni per gli split di *x* (Ishwaran, 2007; Ishwaran et al. 2008; Breiman, 2001).

Valori elevati indicano che la variabile ha capacità predittiva, mentre valori negativi o pari a zero indicano che la variabile va esclusa. Ricordando che l'indice C misura la probabilità che due osservazioni vengano correttamente classificate, la VIMP di *x* è una misura dell'incremento (o il decremento) dell'errore di classificazione sui dati di test dipendente dall'inserimento o meno della variabile *x* nel modello.

Bisogna però essere cauti nell'interpretazione, perché la VIMP non misura *direttamente* quanto cambia l'errore di predizione in due foreste addestrate con e senza la variabile *x* in questione. Se ad esempio due variabili sono altamente correlate ed entrambe hanno VIMP elevata, rimuovere una delle variabili può influenzare la VIMP dell'altra variabile (che crescerebbe) senza che però l'errore di predizione cambi. VIMP e errore di predizione sono quindi concettualmente diversi, ma nella pratica nella maggior parte dei casi la VIMP finisce per misurare proprio il cambio nell'errore di predizione di cui poco sopra. Il calcolo della VIMP può essere molto oneroso a livello computazionale.

Un'altra misura dell'importanza delle variabili può essere la misura di *minimal depth*. Si basa sulla considerazione che se una variabile viene mediamente utilizzata nei primi split degli alberi di sopravvivenza ha un'importanza maggiore, in quanto riesce a massimizzare il criterio di riferimento per lo split più efficacemente delle altre variabili. Questa misura è meno onerosa a livello computazionale e può essere utilizzata in concomitanza con la VIMP.

1.6.6 Dati mancanti

Per l'imputazione dei dati mancanti nella metodologia CART è stato introdotto l'idea di split surrogato. Se *s* è il miglior split per la variabile *x*, se l'osservazione *i* ha un dato mancante per questa variabile, al momento dello split si utilizza lo split s^* della variabile x^* (scelta tra le variabili tra le quali *i* non registra valore mancante) tale che s^* sia il più possibile simile a *s* in termini di risultati previsionali (Breiman et al., 1984). Tuttavia la tecnica dello split surrogato può non essere adatta alle foreste. Per prima cosa perché è molto onerosa a livello computazionale a causa della moltitudine di alberi stimati e del fatto che gli alberi vengono portati alla saturazione massima, per cui si evita di potarli con dei criteri particolarmente restrittivi. Inoltre può non avere senso dal momento in cui si selezionano in modo casuale le variabili che concorrono allo split, può capitare che non vengano incluse variabili correlate a quella per la quale si registra un dato mancante, di conseguenza può non esistere uno split surrogato adeguato. Infine lo split surrogato rischia di distorcere il valore della VIMP.

Per queste ragioni nelle RF è possibile utilizzare un metodo di imputazione che utilizza un approccio di prossimità. Si inizia con una prima imputazione dei dati: i valori mancanti per le variabili continue sono sostituiti dalla mediana e quelli per le variabili categoriali con la moda. Questi dati vengono poi analizzati utilizzando la RF. Nello specifico i dati mancanti vengono imputati utilizzando la matrice simmetrica di prossimità $n \times n$, in cui le celle registrano la frequenza di co-occorrenza delle osservazioni *i, j* nello stesso nodo terminale normalizzata per il numero di alberi della foresta. Per variabili continue i dati mancanti vengono imputati mediante la media pesata dei valori non-mancanti e i pesi sono i valori della matrice di prossimità. Per le variabili discrete si stimano tramite il valore intero più vicino al valore medio di prossimità maggiore, mentre per le variabili categoriali si usa sempre la moda. I valori aggiornati vengono nuovamente utilizzati nella RF e i valori vengono aggiornati nuovamente.

Di solito si raggiunge una soluzione stabile con poche iterazioni. Per approfondimenti sulle modalità di imputazione dei dati mancanti nelle RSF si veda Ishwaran et al., 2008.

CAPITOLO II

Il CONTESTO DI RIFERIMENTO E LA COSTRUZIONE DELLA BASE DATI

SOMMARIO: 2.1 Contesto aziendale e considerazioni preliminari – 2.2 Costruzione della base dati – 2.2.1 La variabile di risposta – 2.2.2 Le altre variabili – 2.3 Tutela della privacy e rispetto degli accordi di diffusione delle informazioni – 2.4 Risorse computazionali a disposizione – 2.5 Descrizione del flusso di dati e delle modalità di implementazione nei sistemi aziendali.

2.1 Contesto aziendale e considerazioni preliminari

Come accennato l'azienda è una multiutility che offre principalmente servizi integrati B2B e B2C di Energia (gas e luce) e Telecomunicazioni (Voce, Adsl e Mobile). Inoltre permette l'acquisto di tutta una serie di oggetti ai propri abbonati (smartphone, laptop e altri dispositivi elettronici)

e da poco fornisce servizi assicurativi. Per ora il parco clienti che usufruisce di questi di servizi è molto ridotto, specie per quel che riguarda i servizi assicurativi, per cui la parte più approfondita dell'analisi riguarda variabili che hanno a che fare con i servizi Energia e Telecomunicazioni. Il prodotto principale dell'azienda, volendo semplificare, consiste nella formulazione di un contratto unico che di solito ingloba più servizi tra quelli del pacchetto servizi a disposizione e per la cui fruizione viene stabilito un prezzo fisso di base entro certi consumi. Di solito le soglie di spesa vengono calcolate analizzando i consumi riportati sulle bollette precedenti dei clienti, arrivando a formulare un range di consumo entro il quale viene pagato un certo importo fisso in bolletta. Questo contratto è detto a Prodotto Integrato. I vantaggi dichiarati principali consistono nell' ottenere un risparmio in bolletta, nell'avere la relativa sicurezza di pagare un prezzo fisso già stabilito, nella semplificazione derivante da una bolletta unica per una moltitudine di servizi e dal potersi interfacciare con un unico servizio clienti in caso di problemi o domande. I clienti però possono anche sottoscrivere contratti a consumo di tipo classico, non caratterizzati dalle peculiarità del Prodotto Integrato.

Considerate le importanti differenze che caratterizzano il parco clienti sulla base della ragione sociale e dei contratti sottoscritti, si è optato per una partizione preliminare del parco clienti in:

- Clienti Prodotto Integrato Business (PIB)
- Clienti Prodotto Integrato Consumer (PIC)
- Clienti Prodotto Sciolto (PS)

Inoltre è possibile che un cliente contrattualizzi un pacchetto di utenze con la forma del Prodotto Integrato e allo stesso tempo una o più utenze mediante un contratto di tipo Sciolto. In queste situazioni i clienti vengono inseriti comunque nella categoria dei clienti con Prodotto Integrato e la loro condizione contrattuale è tenuta inconsiderazione da una variabile appositamente costruita.

Il progetto è stato portato avanti per tutti e tre i dataset, costruendo tre diverse Random Survival Forest. Date le differenze significative tra i tre tipi di clienti ci sono state alcune variabili disponibili solo per uno o due dei tre dataset frutto di questa partizione del parco clienti. Tuttavia la maggior parte delle caratteristiche sono disponibili e registrabili su tutti e tre i tipi di clienti. Inoltre le RSF sono necessariamente diverse per numero e tipo di variabili indipendenti utilizzate, in quanto sono state portate avanti tre analisi distinte.

In questo lavoro di tesi però ci si limiterà alla descrizione dell'analisi effettuata sul dataset dei Clienti PIB, poiché le modalità di analisi sono le medesime e avrebbe poco senso dilungarsi nel descrivere per tre volte le fasi dell'analisi. La scelta è chiaramente arbitraria e non dipende ne da caratteristiche specifiche del dataset dei clienti PIB né dai risultati ottenuti, ma è stata fatta solo per ragioni di sintesi e per gli scopi specifici di questo lavoro.

2.2 Costruzione della base dati

Nel discutere le modalità di costruzione della base dati, per prima cosa vanno definite le strategie di selezione del campione. Sono stati selezionati tutti i clienti della società che hanno sottoscritto il loro primo contratto non più di 42 mesi fa. Si tratta di un lasso temporale piuttosto ampio e che include tipi di clienti piuttosto eterogenei. La scelta di un taglio a 42 mesi è stata presa in seguito ad una attenta analisi della storia aziendale e dell'evoluzione dei prodotti offerti e del parco clienti stesso. Spingersi più indietro di 42 mesi avrebbe portato alla selezione di clienti con dei pacchetti e delle offerte ormai troppo differenti dall'attuale offerta. Infatti, tra le altre cose, l'azienda negli ultimi anni ha profondamente aggiornato la propria struttura, le strategie di raccolta dei dati, ha aggiunto nuovi servizi e si è aperta maggiormente al mondo dei clienti Consumer. Inoltre dato l'elevato tasso di abbandono dei clienti che caratterizza il settore economico di

riferimento, specie nel campo delle Telecomunicazioni, i clienti che hanno sottoscritto il loro primo contratto più di 42 mesi fa sono solo una parte residuale dell'intero parco clienti.

Tutte queste ragioni hanno portato a pensare che l'inclusione di questi clienti non avrebbe portato significativi vantaggi all'analisi, rischiando di inserire un effetto di distorsione dovuto alla peculiarità dei clienti che hanno contratti troppo vecchi. Sono poi stati esclusi dalla base dati tutti quei clienti che hanno sottoscritto contratti particolari, come quelli dedicati ai dipendenti e agli amici dei dipendenti e i clienti appartenenti alla categoria della Pubblica Amministrazione.

Per la costruzione della base dati è stato necessaria un'attività di *mining* dei dati contenuti nei database aziendali. I dati che vengono prodotti e raccolti in una società di servizi, come è facilmente intuibile, possono essere complessi, estremamente estesi e di varia natura. Per queste ragioni è stato necessario dedicare uno sforzo significativo all'estrazione e alla sintesi di queste informazioni, al fine di costruire una base dati utile ed efficace per gli scopi dell'analisi stessa. Lo strumento aziendale principale per l'immagazzinamento e la gestione dei dati in forma di tabelle relazionali è Microsoft SQL Server.

Questa fase ha portato alla selezione di 256 variabili potenzialmente interessanti. La relativa tabella è quindi di dimensioni significativamente grandi. Ogni riga della tabella rappresenta un cliente distinto. Chiaramente la costruzione della base dati è soggetta a una continua revisione e modifica, in quanto è previsto che l'analisi stessa sia periodicamente rivista e aggiornata, allo scopo di migliorarne le performance. Questo perché al di la del valore che ha la singola analisi, essa è inserita in un contesto aziendale in continua evoluzione e i risultati verranno utilizzati giornalmente per specifiche azioni nei confronti dei clienti. Può accadere infatti che ci si renda conto che alcune variabili cruciali fossero state trascurate, che sia possibile la creazione di nuove variabili di sintesi, che

vengano raccolte nuove informazioni che in precedenza non erano disponibili e così via.

I soggetti presi in considerazione e che non hanno ancora sperimentato l'evento terminale (che cioè sono ancora clienti della società) saranno necessariamente osservazioni censurate a destra.

Per quel che riguarda la numerosità campionaria, il dataset considerato, limitato ai soli clienti PIB, è composto da 36694 osservazioni. Di queste 18033 sono ancora clienti della società e sono quindi censurate a destra, mentre 18661 hanno rescisso il contratto e quindi hanno sperimentato l'evento terminale. Si tratta quindi di un 50% circa per ognuna delle due casistiche.

2.2.1 La variabile di risposta

L'obiettivo è la stima della probabilità di sopravvivenza per ognuno dei singoli tempi *t* considerati. Nello specifico l'unità temporale minima presa in considerazione è un mese, per cui si otterrà per ogni mese una probabilità di sopravvivenza per l'osservazione *i-esima* Di conseguenza per ogni osservazione si otterrà un vettore di risultati. Inoltre, come sottolineato poco sopra, l'evento terminale non si è verificato per tutte le osservazioni, che quindi saranno in alcuni casi censurate a destra.

Ne consegue che la natura stessa del modello di sopravvivenza richiede due variabili di risposta: la prima registra i mesi trascorsi dalla sottoscrizione del contratto da parte del cliente, mentre la seconda è una variabile binaria che registra se per il cliente si è verificato l'evento terminale (*churn*) o meno.

2.2.2 Le altre variabili

Come accennato la tabella SQL che sarà utilizzata come dataset dell'analisi è composta da 256 variabili, di cui 253 candidate (direttamente o in seguito ad un adeguato *preprocessing*) a variabili indipendenti della RSF. Delle

altre 3 infatti una registra l'idcliente e le altre due sono le variabili di risposta poso sopra descritte. Di queste 253 variabili verrà fatta una descrizione aggregata per macro categorie di variabili, poiché si ritiene che la descrizione delle singole variabili sarebbe tediosa e di scarsa utilità ai fini di questo lavoro di tesi. La lista delle sole variabili inserite nel modello che verrà trattato, accompagnate da una breve descrizione, potrà invece essere consultata in appendice.

Una parte significativa delle informazioni inserite riguardano uno specifico modo di registrazione delle interazioni tra l'azienda e i propri clienti. Ogni qualvolta che viene effettuata un azione che riguarda in generale il cliente (ad esempio azioni specifiche che riguardano le utenze, l'anagrafica, le spedizioni dei *device*, la segnalazioni dei reclami, delle offerte e così via) questa viene registrata automaticamente nei database o da un operatore mediante delle apposite procedure presenti sul software CRM (*Customer Relationship Management*) aziendale. Questi eventi verranno detti c*asi*, per cui si parla di apertura di un *caso* (ad esempio un reclamo), della sua risoluzione e della sua conseguente chiusura. Si tratta, fondamentalmente, di un sistema di *ticketing* abbastanza diffuso nella registrazione delle interazioni delle aziende con i clienti tramite ad esempio il servizio clienti. I casi vengono classificati utilizzando una classificazione gerarchica a tre categorie. Un caso di guasto Adsl potrebbe ad esempio avere una etichetta del tipo "Reclamo-Guasto-Adsl". L'operatore una volta classificata l'azione inserisce tutta una serie di informazioni, che possono sia essere previste dal sistema, che essere riportate in forma di nota scritta dall'operatore stesso.

Questo tipo di classificazione da vita a migliaia di casistiche diverse, per cui nella fase preliminare è stato necessario accorpare queste casistiche in poche decine di categorie. Si tratta di un'operazione che ha rappresentato una fase molto delicata della fase di estrazione, elaborazione, sintesi e costruzione della base dati. Una delle strategie adottate è stata quella di

chiedere la collaborazione dei team aziendali che hanno seguito la formulazione e lo sviluppo dei casi stessi. In questo modo si è provato a evitare di escludere casistiche importanti o di includerne altre poco importanti, provando a sintetizzare l'informazione in modo efficace. Infatti, da un lato era impraticabile l'utilizzo di una così grande mole di variabili, spesso contenenti informazioni ridondanti o pochissime occorrenze; dall'altro lato si correva il rischio di ridurre in modo eccessivo la variabilità e quindi l'informazione stessa disponibile o, molto peggio, si correva il rischio di aggregare sotto la stessa categoria casistiche eterogenee. Il confronto costante ha portato all'ottenimento di una base dati ritenuta coerente ed efficace (seppur non definitiva).

Sono state costruite due differenti versioni per ognuna delle variabili che registrano il numero di casi aperti sull'anagrafica del cliente: la prima versione registra il numero di casi aperti nella storia del cliente e la seconda il numero di casi aperti nell'ultimo mese.

Per quel che riguarda le macro categorie di variabili individuate, si possono distinguere:

- Variabili socio-anagrafiche legate al referente del contratto: quali l'età, l'area di residenza, il genere e così via.
- Variabili legate alle caratteristiche contrattuali del cliente: quali numero e tipo di utenze da quando è diventato cliente dell'azienda; numero e tipo di utenze attive al momento dell'analisi; se nella loro storia hanno cambiato tipologia contrattuale (ad esempio convertendo dei contratti PS in contratti PI); per i clienti PIC e PIB si registra il numero di contratti PS se ce ne sono; tutta una serie di variabili sulle specifiche tecniche dell'utenza, ad esempio il tipo di linea, la potenza impiegata e così via.
- Variabili legate a fatturazione e pagamenti: quali fattura media, modalità di spedizione della fattura, modalità di pagamento, numero

di mesi in cui ci sono stati ritardi nei pagamenti, numero di blocchi temporanei delle utenze per morosità e così via.

- Numero di casi aperti sull'anagrafica del cliente che riguardano la richiesta o ricezione di Informazioni: possono essere informazioni di carattere tecnico, sui contratti, a proposito di offerte dedicate e così via.
- Numero di casi aperti sull'anagrafica del cliente che riguardano operazioni di Variazione: possono riguardare variazioni dell'offerta, dell'anagrafica, delle caratteristiche tecniche del servizio e così via.
- Numero di casi aperti sull'anagrafica del cliente che riguardano la ricezione di una Campagna: si tratta delle volte in cui il cliente rientra in una campagna specifica e viene contattato. Le campagne possono avere a che fare con tentativi di *retention*, la necessità di rimodulare la taglia dell'offerta, azioni di *cross-selling* per far sottoscrivere abbonamenti a prodotti aggiuntivi, comunicazione di nuovi servizi e così via.
- Numero di casi aperti sull'anagrafica del cliente che riguardano l'apertura di un Reclamo: si può trattare di reclami per guasti tecnici, per incongruenze di fatturazione, ritardi di attivazione e così via.
- Numero di casi aperti sull'anagrafica del cliente che riguardano un invio di documentazione: si tratta di quei contatti con il cliente quando avviene una richiesta esplicita per l'invio del contratto, della fattura, di modulistica e così via. Può anche riguardare l'invio di documentazione inerente al credito.

2.3 Tutela della privacy e rispetto degli accordi di diffusione delle informazioni

Ci sono due ordini di problemi che sono state affrontati per la stesura di questo lavoro di tesi e che meritano una breve discussione: in primo luogo le modalità per il rispetto della privacy dei soggetti coinvolti; in secondo

luogo il rispetto degli accordi contrattuali sottoscritti da chi scrive sulle modalità di diffusione dei dati e dei risultati del lavoro svolto in azienda.

Per quel che riguarda la privacy dei soggetti coinvolti chi scrive era chiaramente in possesso di tutta una serie di dati personali che l'azienda è autorizzata a rilevare sulla base del contratto che la lega con il cliente. Nel presente lavoro di tesi quindi, nel maggior rispetto possibile della privacy dei soggetti, non verrà fornito alcun tipo di informazione che possa ricondurre ai soggetti coinvolti nello studio e i risultati verranno riportati solo in forma aggregata.

Chi scrive poi è tenuto a rispettare una serie di clausole contrattuali che vietano la diffusione dei risultati e dei progetti aziendali, se non rispettando determinati criteri, per evitare di incorrere in sanzioni legati alla tutela della privacy e per evitare di diffondere informazioni ai competitor. Per questi motivi si eviterà di comunicare informazioni aziendali sensibili. Inoltre il dataset utilizzato non verrà duplicato né reso pubblico, neanche in seguito a un occultamento dei dati sensibili. Il dataset verrà quindi visionato e validato solo dai soggetti coinvolti in questo lavoro (vale a dire chi scrive, il relatore e gli eventuali collaboratori). Invece, in accordo con la controparte aziendale, non c'è ragione di adottare strategie di occultamento del codice utilizzato nell'analisi né del processo di analisi stesso, che d'altronde è inevitabile presentare nei dettagli in un lavoro di tesi.

2.4 Risorse computazionali a disposizione

Come accennato la costruzione della base dati è stata portata a termine servendosi di Microsoft SQL. La parte di analisi è invece stata svolta in R. Le specificità dei pacchetti utilizzati verranno discusse nel capitolo in cui si discute l'analisi deva e propria. Il progetto necessitava di risorse significative in termini di architettura server e di capacità computazionali per tutta una serie di motivi.

In primo luogo per le caratteristiche della metodologia utilizzata. La metodologia RSF, come visto nel capitolo precedente, è piuttosto complessa e computazionalmente onerosa. Si vedrà infatti che un singolo modello può pesare anche più di un gigabyte e che il calcolo di un grosso numero di alberi può necessitare di molte risorse di calcolo. In particolare quando si stima un numero di alberi che arriva già solo all'ordine delle centinaia, il calcolo di misure di performance come la Variable Importance possono necessitare di diverse ore di calcolo anche su macchine ad elevata memoria e capacità di calcolo. Dato che si tratta di misure cruciali nella selezione delle variabili e nella valutazione delle performance del modello, si può facilmente immaginare che queste vadano ricalcolate una moltitudine di volte prima di arrivare a una soluzione soddisfacente.

In secondo luogo i dataset considerati sono di dimensioni piuttosto elevate e richiedono memoria e capacità di calcolo adeguati anche nella fase di *preprocessing*.

In terzo luogo a seguire di parlerà della necessità di un continuo aggiornamento della base dati e dei risultati dell'analisi, e dell'integrazione di questi ultimi sia in ingresso che in uscita con altri sistemi aziendali. Questa necessità richiede chiaramente memoria e capacità di calcolo elevati.

Per tutte queste ragioni la parte di analisi dei dati è stata svolta su un server aziendale ad alte prestazioni. Si tratta di una macchina con sistema operativo Ubuntu sul quale tra i vari programmi è installato R studio. La macchina in questione ha una memoria molto elevata ed ha a disposizione 56 *CPU* ognuno dei quali può eseguire più *thread*. I calcolo più onerosi sono stati quindi parallelizzati su una moltitudine delle *CPU* a disposizione, ottenendo una grossa capacità e velocità di calcolo. A dimostrazione della necessità di grosse risorse in termini computazionale, si fa notare che anche avendo a disposizione tutta questa capacità di calcolo il calcolo della VIMP in alcuni casi ha richiesto anche tre o quattro ore di elaborazione.

2.5 Descrizione del flusso di dati e delle modalità di implementazione nei sistemi aziendali

A questo punto è possibile descrivere in modo più schematico la definizione del flusso di informazioni che è stato necessario all'analisi in primo luogo e al suo utilizzo pratico in secondo luogo. Questa fase ha rappresentato una sfida importante nel progetto che ha in qualche modo dovuto coinvolgere diversi *stakeholders*. Detto in altri termini era necessario definire delle modalità per avere dei risultati sempre aggiornati e un modo per renderli fruibili agli operatori, i quali dovranno poi nella pratica fare delle azioni nei confronti dei clienti.

Come già accennato, si parlerà in maniera approfondita dei vantaggi che l'azienda può trarre dai risultati di quest'analisi nelle conclusioni di questo lavoro di tesi. In questo paragrafo si vuole invece descrivere l'architettura vera e propria del progetto. L'architettura del progetto può essere riassunta nei seguenti punti, messi in atto con cadenza giornaliera:

1. Costruzione della base dati.
2. Stima della probabilità di sopravvivenza di tutto il parco clienti.
3. Calcolo del *lifetime value*.
4. Segmentazione del parco clienti sulla base del *lifetime value*.
5. Scrittura delle tabelle che registrano *lifetime value* e cluster di appartenenza da questo derivato per ogni singolo cliente sul server del Data Warehouse.
6. Visibilità del cluster di appartenenza sull'interfaccia unica del cliente sulla piattaforma di CRM e del menù a tendina che permetta all'operatore di selezionare una serie di azioni sulla base del cluster di appartenenza.
7. Scrittura degli esiti in una tabella sul server del Data Warehouse.

Per quel che riguarda la costruzione della base dati è stato necessario pensare a una strategia per avere i dati costantemente aggiornati. Si è

proceduto alla creazione di una *stored procedure* in Microsoft SQL che permettesse di eseguire in un solo comando lo script di creazione della tabella. Si è fatta una rapida analisi di compatibilità con le procedure che aggiornavano le tabelle da cui vengono presi i dati che servono alla creazione della tabella e si è schedulata la sua creazione ogni mattina. In questo modo si hanno sempre disponibili i dati aggiornati sul Data Warehouse aziendale.

La stima delle probabilità di sopravvivenza nei vari tempi *t* avviene inserendo i clienti obbiettivo come test set nel modello di RSF creato e validato. Anche questa procedura viene effettuata ogni giorno mediante schedulazione automatica dello script R, così da avere le stime della probabilità di sopravvivenza sempre aggiornati a ogni mutamento delle variabili associate ai singoli clienti.

Lo stesso vale per il calcolo del *lifetime value* e la segmentazione del il parco clienti.

Le stime dei tempi di sopravvivenza, il *lifetime value* e il cluster di appartenenza per il giorno in questione insieme ad altre variabili di interesse vengono inseriti in automatico in una tabella sul Data Warehouse. Ogni giorno la nuova tabella viene incollata in coda alla tabella del giorno prima inserendo la data di riferimento. In questo modo si otterrà una grossa tabella storicizzata per valutare come variano i risultati nel tempo.

La piattaforma CRM legge la tabella con i risultati in modo che per ogni cliente venga riportato il cluster di appartenenza. Per ogni cliente l'operatore vedrà un menù a tendina riportante una serie di azioni possibili. Le azioni dipendono dalle caratteristiche del cliente e sono state definite sulla base dei risultati dell'analisi di sopravvivenza. In questo modo si prova ad intervenire su quelle variabili critiche per i modello e si prova a massimizzare il *lifetime* del cliente. Gli esiti dell'azione intrapresa vengono riportati su un apposita tabella per una fase successiva di valutazione delle performance dell'azione stessa.

CAPITOLO III

L'ANALISI DEI DATI

SOMMARIO: 3.1 Librerie R utilizzate – 3.2 Operazioni preliminari sulle variabili – 3.3 La variabile di risposta – 3.4 *Random Survival Forest*: strategie di costruzione e misure di performance – 3.5 *Variable Importance* – 3.6 *Customer lifetime* e *lifetime value* –3.7 Definizione dei risultati più significativi – 3.8 Studio delle relazioni tra le variabili della RSF – 3.9 Segmentazione del parco clienti.

3.1 Librerie R utilizzate

Per l'analisi in R la libreria principale utilizzata per l'implementazione delle RSF è *RandomForestSRC* (Ishwaran, Kogalur, 2019), libreria che include algoritmi di calcolo per le Random Forest nel caso di problemi di classificazione binaria e multipla, di regressione, e per l'analisi della sopravvivenza. Per arricchire la parte grafica è stata utilizzata *ggRandomForest* (Ehrlinger, 2016), che segue una sintassi basata su

ggplot2, anche esso utilizzato per la parte grafica. Per alcune funzioni legate all'analisi della sopravvivenza si è fatto uso della libreria *survival* (Therneau, 2015). Per la parte di *pre-processing*, oltre alle librerie di base, sono state usate le librerie del *Tidyverse* (Hadley, 2017). Le altre librerie utilizzate verranno riportate in appendice insieme al resto del codice R.

3.2 Operazioni preliminari sulle variabili

Una volta costruita la base dati si è proceduto all'importazione del dataset in R studio. Per farlo è stata costruita una funzione in R che crea una connessione al database Microsoft SQL. Una volta definita correttamente la tipologia delle variabili si è proceduto alla creazione di una serie di indici. Si sono sommate tutte quelle variabili che registrato il numero di reclami, informazioni, variazioni e campagne, e che sono state giudicate marginali per numero di osservazioni o per ragioni qualitative. Anche le variabili che registrano il numero di accessi ai vari canali digitali sono state sommate in un'unica variabile. Operazioni simili sono state effettuate per tutta una serie di variabili, che non si ritiene sia il caso di descrivere approfonditamente, perché non sono state cruciali nella formulazione del modello.

Per quel che riguarda il numero di casi aperti dal cliente, poiché viene registrato il numero di casi aperti da quando il cliente ha firmato il suo primo contratto, c'è il rischio che ci sia una distorsione dovuta alla natura stessa dei dati. La distorsione deriva dal fatto che è molto più probabile che i clienti con valori maggiori sulla variabile mesi cliente abbiano aperto necessariamente più casi. Molti dei casi infatti indicano semplici richieste di informazione oppure delle campagne *ricevute* dal cliente, per cui sono chiaramente correlate alla variabile che registra i mesi trascorsi dall'attivazione della prima utenza. Per superare questa peculiarità dovuta alla natura stessa di queste variabili si è proceduto al calcolo della media mensile dei casi aperti dal cliente, queste variabili sostituiranno le corrispondenti registrate in valore assoluto.

Spesso i clienti con Prodotto Sciolto vengono invogliati a passare al prodotto integrato. Inoltre, come accennato, un cliente con un contratto Prodotto Integrato può avere anche un'utenza contrattualizzata come Prodotto Sciolto. Si è costruita allora una variabile che registri se il cliente PIB ha avuto o meno un contratto Prodotto Sciolto ed è stata pensata come flag binario, che si vedrà sarà più utile nell'analisi di quelle che registrato il numero di contratti sottoscritti e quelli ancora attivi per tipologia.

Infine si è proceduto alla standardizzazione delle variabili numeriche, per ragioni di diversità di scala e di range e perché è un'operazione spesso adottata nell'implementazione delle *Random Forests*.

3.3 La variabile di risposta

Come anticipato la percentuale delle osservazioni per cui si è verificato l'evento terminale è poco più della metà, per cui si è in presenza di una grossa percentuale di dati censurati a destra.

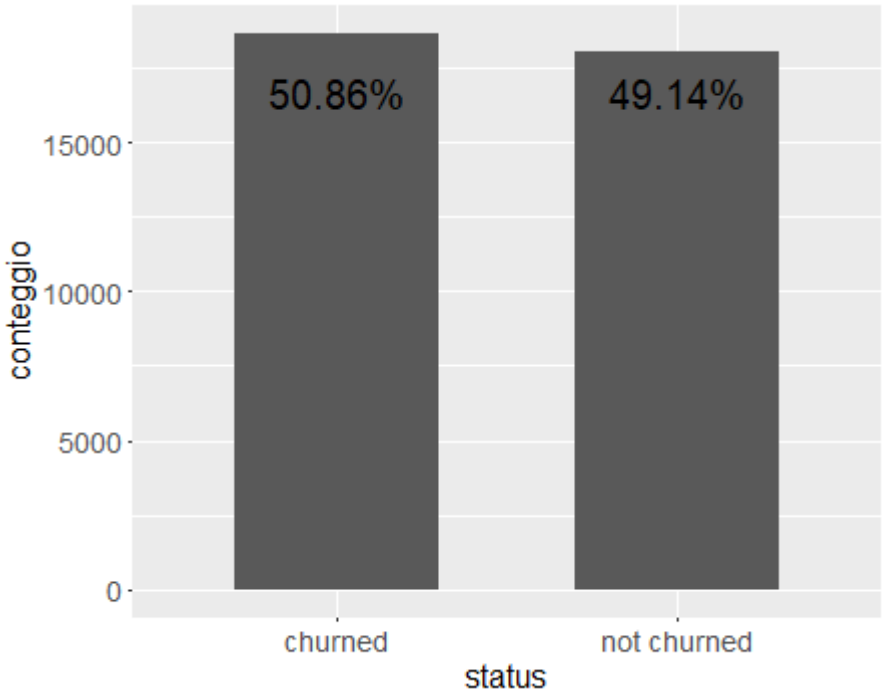

Figura 2 - diagramma a barre della variabile che registra se si è verificato l'evento terminale (churn)

Per quel che riguarda la distribuzione empirica del tempo di sopravvivenza, identificato nella variabile numero di clienti, si ha una distribuzione fortemente asimmetrica, con una concentrazione a sinistra della curva.

Si ricorda che il range di mesi osservato è 42 mesi e l'unità elementare di osservazione è il singolo mese, per cui si tratta di una discretizzazione della variabile continua che registra il tempo prima che si verifichi l'evento terminale. La discretizzazione in mesi è però necessaria per la natura stessa del contratto stipulato dai clienti con l'azienda, che prevede un canone mensile, per cui anche al momento della recessione del contratto quest'ultima sarà effettiva comunque a un mese dall'ultima rata di pagamento delle utenze.

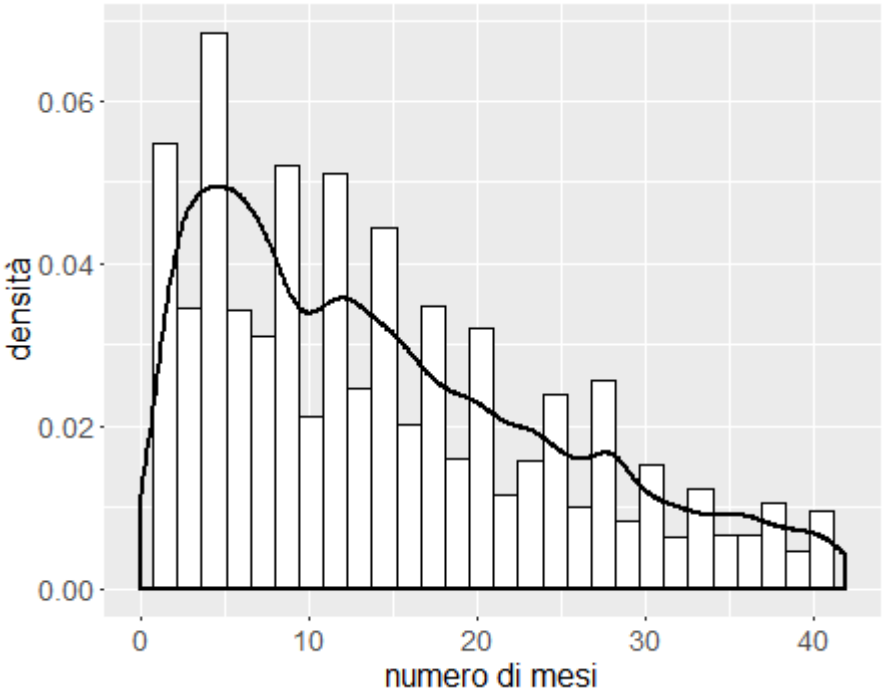

Figura 3 - distribuzione di densità di probabilità della variabile sul numero di mesi di permanenza del cliente

Dal Boxplot in basso si nota con maggiore chiarezza che il 50% delle osservazioni ha un numero di mesi che è compreso tra i 6 e i 21 mesi, con mediana 12 mesi e media 14,6 mesi.

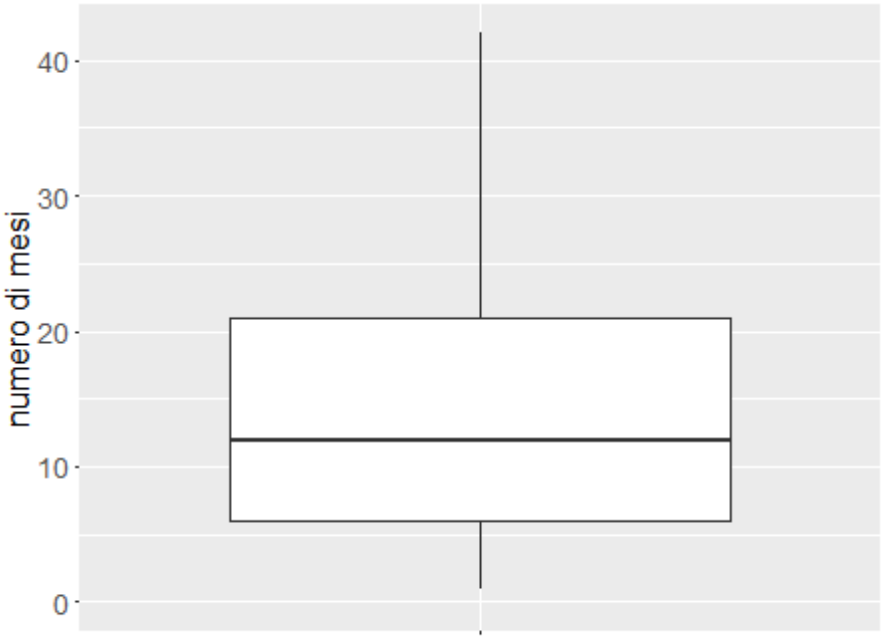

Figura 4 - boxplot della variabile sul numero di mesi di permanenza del cliente

3.4 Random Survival Forest: strategie di costruzione e misure di performance

Come ogni processo di costruzione di un'analisi statistica, quello del presente lavoro di tesi è stato circolare e soggetto a un grosso numero di tentativi, per provare a trovare la miglior soluzione in termini di utilità e performance. Per prima cosa si è proceduto con un'analisi qualitativa del contenuto informativo delle variabili stesse. Alcune variabili infatti, seppur aventi significato differente, nella pratica tendevano a replicare il contenuto informativo di altre variabili e per questa ragione non contribuivano al miglioramento delle performance del modello, piuttosto rischiavano in qualche modo di intaccarne la validità.

Per fare un esempio basta pensare alle variabili che registrano il numero di contratti sottoscritti dal cliente per tipo. Queste variabili sono in

rapporto uno a uno con la variabile che registra il numero di utenze attivate dal cliente nella sua storia di cliente della società, e in relazione quasi perfetta con la variabile che registra il numero di utenze che ha attive al momento dell'analisi (nel caso di clienti ancora attivi). Con quest'ultima variabile, ad esempio, anche se non ci sarà un rapporto uno ad uno, la correlazione sarà di tipo quasi lineare, introducendo di fatto ridondanza nel modello.

Nonostante si tratti di modelli robusti al crescere del numero di covariate, chiaramente, non è sufficiente inserire indistintamente un grosso numero di variabili esplicative per ottenere dei risultati soddisfacenti. Una prima selezione ha mirato quindi ad includere solo quelle variabili che contribuiscono maggiormente all'analisi sia in base al significato dei risultati che ai valori di performance. Si è trattato allora di una selezione che ha alternato criteri di selezione più qualitativi al confronto di indici statistici di correlazione e di misure proprie delle Random Survival Forest, quali le misure di Variable Importance e l'Error Rate.

La strategia principale di selezione delle variabili esplicative è riconducibile in modo più o meno diretto a quello di *backward selection*. Dopo una prima scrematura si è proceduto all'inserimento di un grosso numero di variabili e all'esclusione progressiva delle variabili meno importanti, sulla base di una valutazione delle misure di Variable Importance, Error Rate e di Minimal Depth. Periodicamente si è proceduto a reinserire quelle variabili escluse negli step precedenti ma che si consideravano importanti nell'analisi, per valutare se la loro esclusione era dovuta a un qualche tipo di distorsione della loro reale capacità esplicativa, riconducibile, ad esempio, all'inclusione di un grosso numero di variabili nelle fasi iniziali della costruzione della RSF.

Questo processo di selezione ha portato alla scelta di 54 variabili esplicative. Per la lista completa e una loro sintetica descrizione si veda l'appendice A.

Sono stati stimati 1000 alberi di sopravvivenza. Lo scopo era quello di beneficiare dei vantaggi in termini di stabilità dei risultati e di riduzione dell'Error Rate derivante dalla stima di un grosso numero di alberi. Chiaramente la scelta di stimare un così elevato numero di alberi deriva anche dalle grosse capacità computazionali a disposizione, infatti l'Error Rate si stabilizza ben prima di arrivare alla stima di un così elevato numero di alberi.

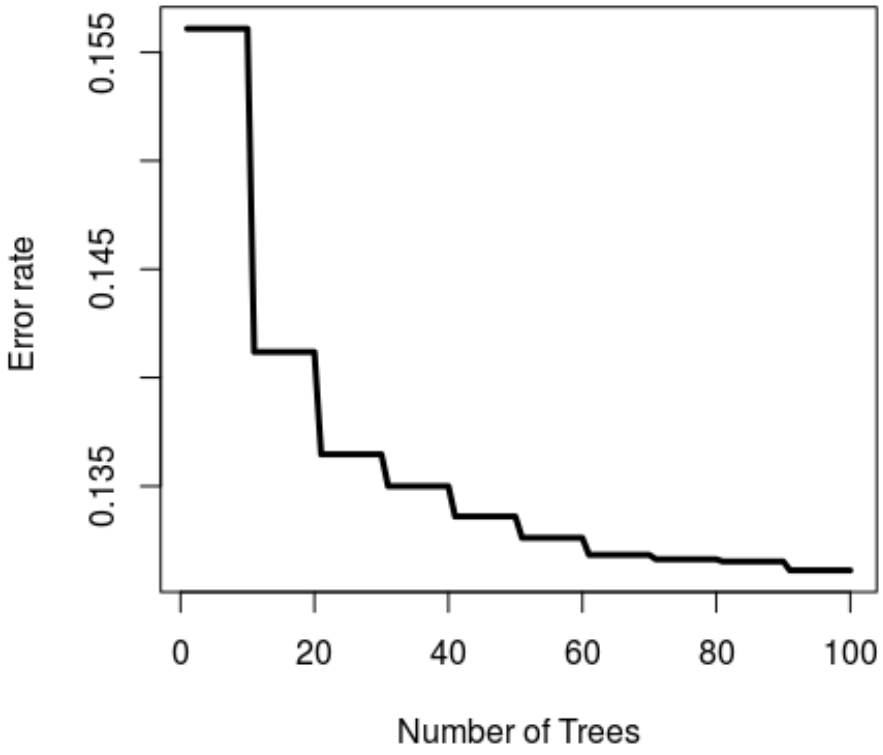

Figura 5 - grafico di discesa del tasso di errore al crescere del numero di alberi stimati nella foresta

Guardando la *Fig.* 5 è chiaro che la curva di riduzione dell'Error Rate inizia a stabilizzarsi già per un numero di alberi che tende a 100. Infatti una RSF a 100 alberi, utilizzando tutte le 54 variabili indipendenti, restituisce un Error Rate dell'13,51%, che è già considerabile un buon risultato.

Spingendosi fino alla stima di 1000 alberi si ottiene tuttavia un miglioramento di oltre due punti percentuali sull'Error Rate, che raggiunge l'11,51%, un risultato molto soddisfacente e ben distante dalla soglia del 50% che indica la situazione in cui il valore della risposta è stimato in modo totalmente casuale. Oltre che sulla base del confronto con il valore teorico del 50% e con le altre applicazioni delle RSF, il risultato è giudicabile molto soddisfacente anche perché si sta analizzando un fenomeno piuttosto complesso, che ha a che fare con tutta una serie di variabili difficilmente controllabili. Infatti, quando si ha a che fare con le interazioni tra le persone (ad esempio quelle registrate nelle variabili che registrano i casi) e con l'esperienza soggettiva dei singoli, ci si aspettano generalmente risultati peggiori di quelli ottenibili in applicazioni mediche e/o di biostatistica, che rappresentano i campi di applicazione per i quali le tecniche di analisi della sopravvivenza sono state sviluppate. Un Error Rate così basso è quindi più che accettabile.

I dati mancanti sono stati imputati automaticamente secondo la metodologia proposta in Ishwaran et al., 2008 ed implementata nel pacchetto R RandomForestSRC (Ishwaran, Kogalur, 2019).

Un altro parametro importante ha a che fare con la numerosità delle osservazioni che andranno a ricadere in ogni nodo terminale. Dopo una serie di tentativi si è optato per lasciare invariate le impostazioni di default, che prevedono che in ogni nodo terminale siano presenti al massimo 15 osservazioni. Così come nella metodologia CART, questo parametro servirà da regola di stop, per cui quando un nodo includerà un numero pari o inferiore a 15 osservazioni non verrà ulteriormente splittato e sarà dichiarato terminale. Come discusso nel *Cap. 1*, i risultati sono dati da una media dei valori per le osservazioni presenti nel nodo terminale. Questo comporta che all'aumentare del numero di osservazioni presenti nei nodi terminali aumenta l'eterogeneità interna al nodo, ma aumenta anche la velocità di calcolo delle RSF. Viceversa al diminuire del numero di

osservazioni presenti in ogni nodo terminale diminuisce l'eterogeneità interna ai nodi, ma si riduce anche la velocità di calcolo delle RSF. Tenendo presente queste considerazioni si è optato per una numerosità massima pari a 15 osservazioni.

Per la scelta di ogni split si considera ogni volta un sottocampione di 8 variabili (si vedano i principi del bagging discussi nel *Cap. 2*).

Uno schema riassuntivo delle caratteristiche della RSF e del setting dei parametri è riportato nella tabella di seguito.

Tabella 1 - tabella di sintesi dei parametri impostati nella stima della Random Survival Forest

	Valore
Numerosità campionaria	36694
Numero di eventi terminali	18661
Imputazione dei dati mancanti	si
Numero di alberi	1000
Numerosità dei nodi terminali	15
Numero medio di nodi terminali	2.949.659
Numero di variabili testate a ogni split	8
Numero totale di variabili	54
Regola di split	logrank
Error Rate	11,51%

Come si è detto dalla RSF si otterrà per ogni cliente la probabilità di sopravvivenza per ognuno dei tempi *t* considerati. Avendo considerato 42 mesi si otterranno quindi 42 valori di probabilità distinti. In aggiunta alle misure e ai grafici disponibili nelle librerie RandomForestSRC e ggRandomForest si è optato per la costruzione di ulteriori misure, basate sulle stime di sopravvivenza nei diversi istanti temporali. In questo modo ci si potrà fare un'idea ancora più approfondita delle performance della RSF.

Una volta calcolata la probabilità di sopravvivenza per ognuno dei mesi su tutti i clienti, una prima misura è stata costruita considerando i soli clienti per i quali si è verificato l'evento terminale. La misura è basata su un confronto tra i mesi realmente trascorsi prima del verificarsi dell'evento

terminale e quelli previsti dalla RSF. La previsione è stata costruita nel modo seguente: il primo mese in ordine crescente nel quale la RSF prevede una probabilità di sopravvivenza inferiore al 50% è considerato il mese in cui la RSF stima l'evento terminale. Detto in altri termini, poiché la curva di sopravvivenza per ogni soggetto è monotona decrescente, l'evento terminale è stimato quando la probabilità che si verifichi l'evento terminale è maggiore della probabilità che quest'ultimo non si verifichi.

A questo punto si è fatto un confronto tra il numero di mesi effettivi prima dell'evento terminale e quelli stimati e si è proceduto alla costruzione di una *Heatmap*.

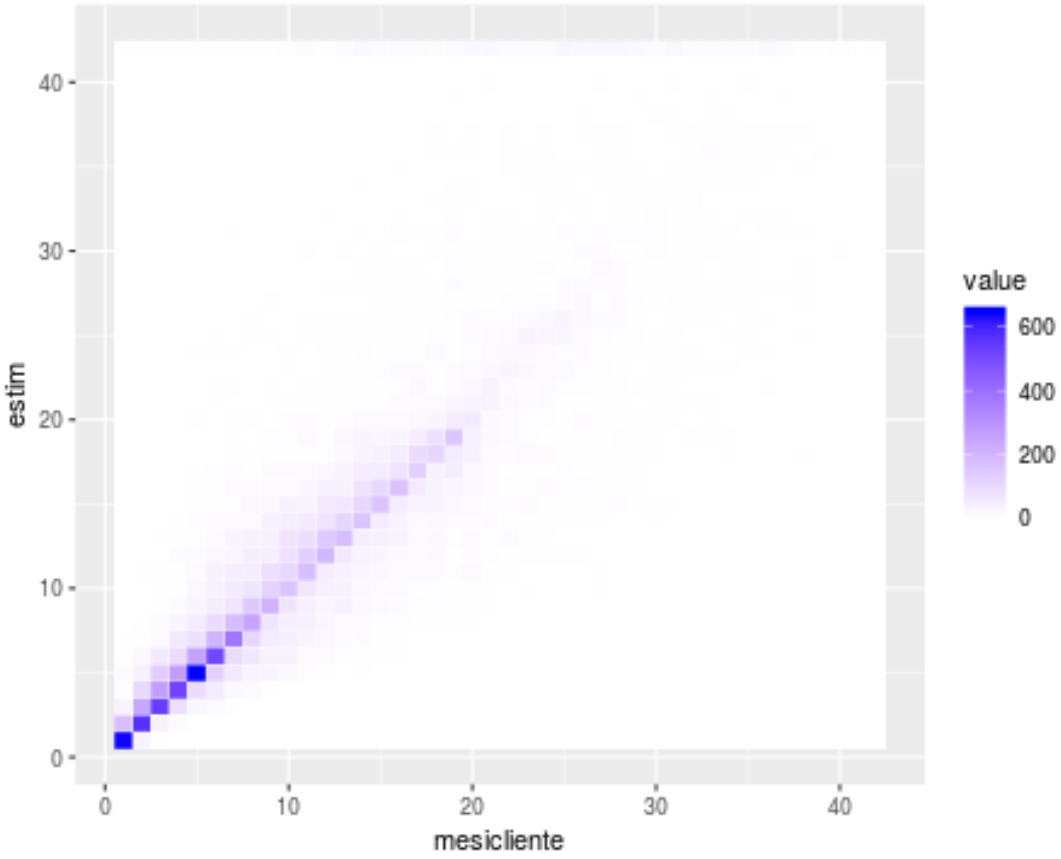

Figura 6 - heatmap dei mesi effettivi del cliente vs i mesi stimati per i soli clienti che hanno sperimentato l'evento terminale (*churn*)

Non si tratta di altro che di una rappresentazione grafica di una matrice di contingenza. L'intensità del colore dei quadrati è proporzionale al numero di soggetti presenti in quella cella. Il fatto che i valori siano concentrati sulla diagonale secondaria, per come è costruito il grafico, indica che per una percentuale molto alta dei casi c'è concordanza tra il valore predetto dalla RSF e quello reale.

Un'altra misura è stata costruita invece solo sui clienti attivi, vale a dire per le osservazioni censurate a destra. Anche in questo caso si tratta di una misura elaborata da chi scrive per arricchire la descrizione delle performance del modello. Si tratta di una misura meno rigorosa di quella precedente, che potrebbe essere definita di *plausibilità*. Si è proceduto al calcolo dei mesi stimati prima che si verifichi l'evento terminale seguendo gli stessi criteri usati per la misura descritta poco sopra. A questo punto si è fatto un confronto con i mesi reali che il cliente ha fin ora trascorso come cliente dell'azienda. Dato che i clienti sono osservazioni censurate a destra, il valore reale del tempo all'evento terminale non è noto. Tutte le volte che però la RSF prevede un tempo di sopravvivenza inferiore ai mesi reali attuali si può essere certi che la stima è errata. Tutte le volte che invece il tempo previsto è superiore a quello reale, è *possibile* che la stima sia corretta. Si tratta quindi più di una stima di quante volte si è certi che la stima sia sbagliata. L'errore di stima, seguendo i criteri appena descritti, è commesso l'8,04% delle volte.

Si è optato poi per la costruzione e la rappresentazione di una curva di sopravvivenza media per tutto il parco PIB. La stima del valore di probabilità di sopravvivenza per ognuno dei tempi t_i, che poi serviranno a tracciare la curva, è ottenuta mediante il calcolo della media delle probabilità di sopravvivenza dei clienti al tempo t_i considerato. Chiaramente si tratta di una misura descrittiva più che di un risultato da cui trarre ulteriori considerazioni nell'analisi dei risultati.

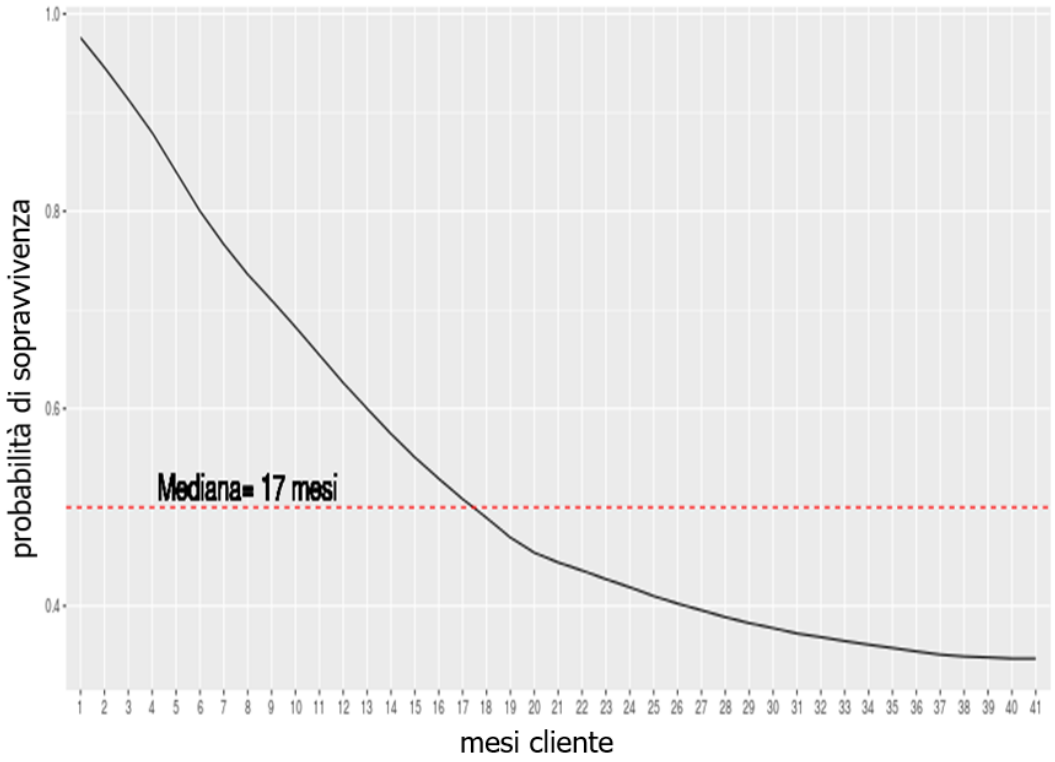

Figura 7 - curva di sopravvivenza media stimata per il parco clienti PIB

Una curva molto simile è utilizzata negli studi di sopravvivenza per descrivere la percentuale di pazienti che sopravvive o che si stima sopravviva dopo un certo periodo temporale. In oncologia umana si parla talvolta di *sopravvivenza mediana*, intendendo il tempo in cui è sopravvissuto il 50% dei pazienti (e quindi il 50% è deceduto). Ancora una volta il concetto è traslabile al campo della *customer churn analysis*, e sulla base delle stime della RSF la sopravvivenza mediana si colloca attorno ai 17 mesi.

3.5 *Variable importance*

Di seguito si discuteranno i risultati di Variable Importance. Ci si limiterà alla presentazione dei risultati, mentre le conclusioni tratte verranno discusse nel prossimo capitolo.

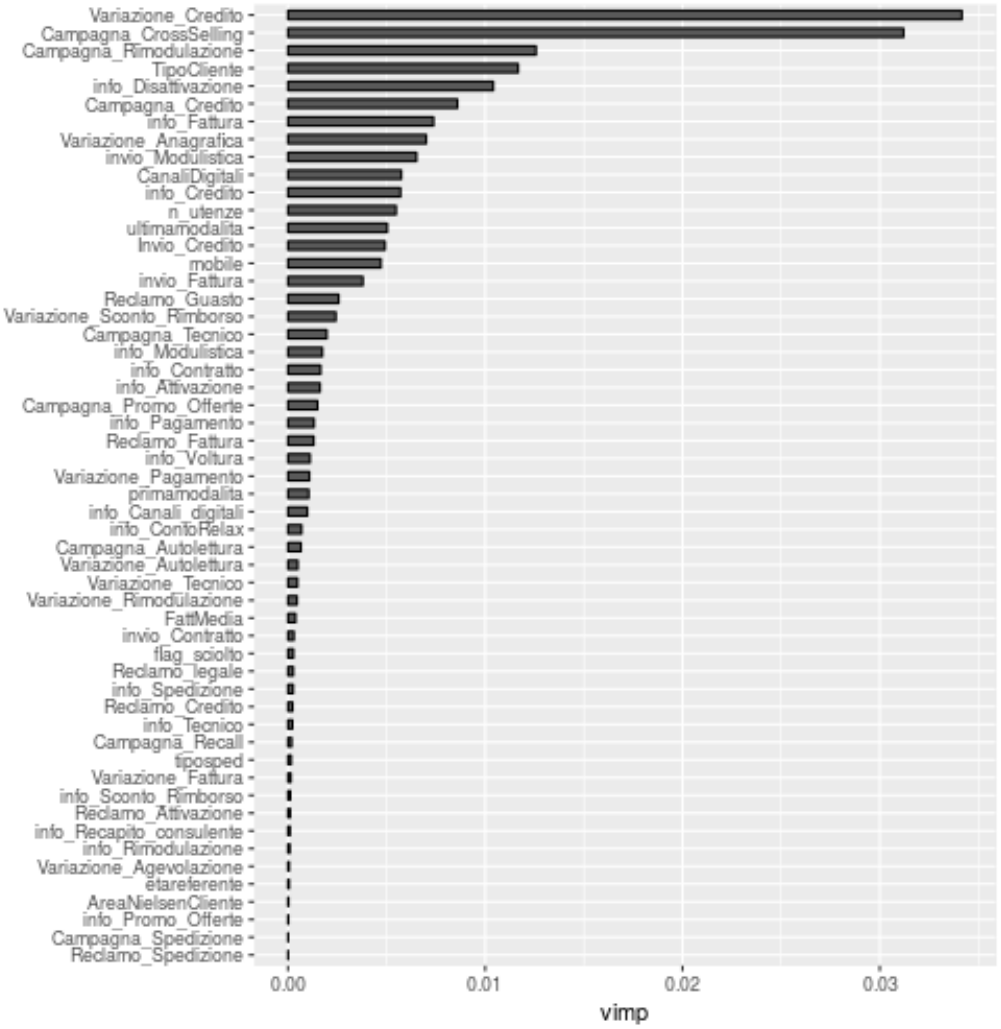

Figura 8 - diagramma a barre dei valori di Variable Importance (VIMP) delle variabili indipendenti inserite nella RSF

Chiaramente non ci sono variabili con VIMP negativa perché sono state escluse durante il processo di selezione delle variabili esplicative. Di seguito una breve descrizione con alcuni esempi delle prime 15 variabili con VIMP maggiore.

La variabile con VIMP maggiore è quella che indica il numero medio mensile di variazioni che hanno a che fare con il credito. Si tratta ad esempio di richieste di dilazione dei pagamenti o di azioni direttamente legate con la situazione creditizia del cliente, per cui clienti che hanno

molte variazioni del credito sono o sono stati morosi, hanno fatture arretrate, e così via.

La seconda variabile indica il numero medio mensile di campagne di cross selling ricevute dal cliente. Queste campagne consistono nel contattare il cliente per provare a convincerlo ad ampliare il suo pacchetto di offerte.

La terza variabile indica il numero medio di campagne di rimodulazione ricevute. Questa campagna è molto importante nella gamma delle varie campagne implementate dall'azienda. Viene fatta per cambiare la taglia (si veda il Cap.2) del cliente perché c'è stato un errore nel calcolo dei consumi medi previsti. Lo scopo è evitare che il cliente sia insoddisfatto perché paga molto rispetto ai suoi bassi consumi o perché paga un sovrapprezzo dovuto a un consumo oltre soglia, che di solito comporta costi più elevati rispetto al consumo previsto dalla taglia di riferimento.

La variabile successiva in termini di VIMP è una variabile categoriale che indica se il cliente ha solo utenze Energia e/o Gas (ENG), solo utenze Voce, Adsl e/o Mobile (TLC), o se ha entrambi i tipi di utenze (ENG/TLC).

Quella successiva registra il numero medio mensile di casi aperti per chiedere informazioni su un eventuale disattivazione di una o più utenze.

A seguire la variabile che registra il numero medio mensile di campagne credito ricevute. Si fa una campagna credito quando ad esempio un cliente ha dei pagamenti arretrati e vengono proposte delle dilazioni nel pagamento.

A seguire la variabile che registra il numero medio mensile di casi aperti per chiedere informazioni sulla fattura. Si può trattare di semplici richieste di chiarimento su alcune voci in fattura oppure di richieste di spiegazione perché, ad esempio, il cliente ritiene che la cifra richiesta non corrisponda ai suoi consumi o al canone concordato.

Segue la variabile che registra il numero medio mensile di richieste di variazione anagrafica. Un cliente può chiedere che le utenze cambino intestatario o può voler correggere i suoi dati anagrafici.

A seguire la variabile che registra il numero medio mensile di accessi ai canali digitali. Sono dati dalla somma del numero di accessi del cliente al bot di Telegram, alla chat live con un operatore in carne ed ossa, al portale web selfcare o ai contatti con il chatbot da sito web o tramite social.

A seguire la variabile che registra il numero medio di casi aperti per chiedere informazioni inerenti al credito. Si può trattare di richieste di informazione in merito al numero di fatture arretrate, alla dilazione dei pagamenti e così via.

La successiva variabile registra il numero di utenze che il cliente ha attivato nella sua storia di cliente dell'azienda.

Quella successiva indica l'ultima modalità di pagamento del canone mensile. Il cliente ha la possibilità di pagare tramite bollettino postale, tramite bonifico bancario o tramite addebito diretto sul conto corrente.

La variabile a seguire indica il numero medio mensile di casi aperti per la richiesta di invio di documentazione inerente al credito.

La quindicesima variabile in termini di VIMP registra se il cliente ha attiva anche un'utenza Mobile.

Il grafico in *Fig.* 8 fa riferimento alla sola VIMP. Un altro grafico interessante che determina il rank delle variabili indipendenti nella RSF è dato dall'intreccio dei risultati della VIMP con quelli del rank dato dalla misura di Minimal Depth.

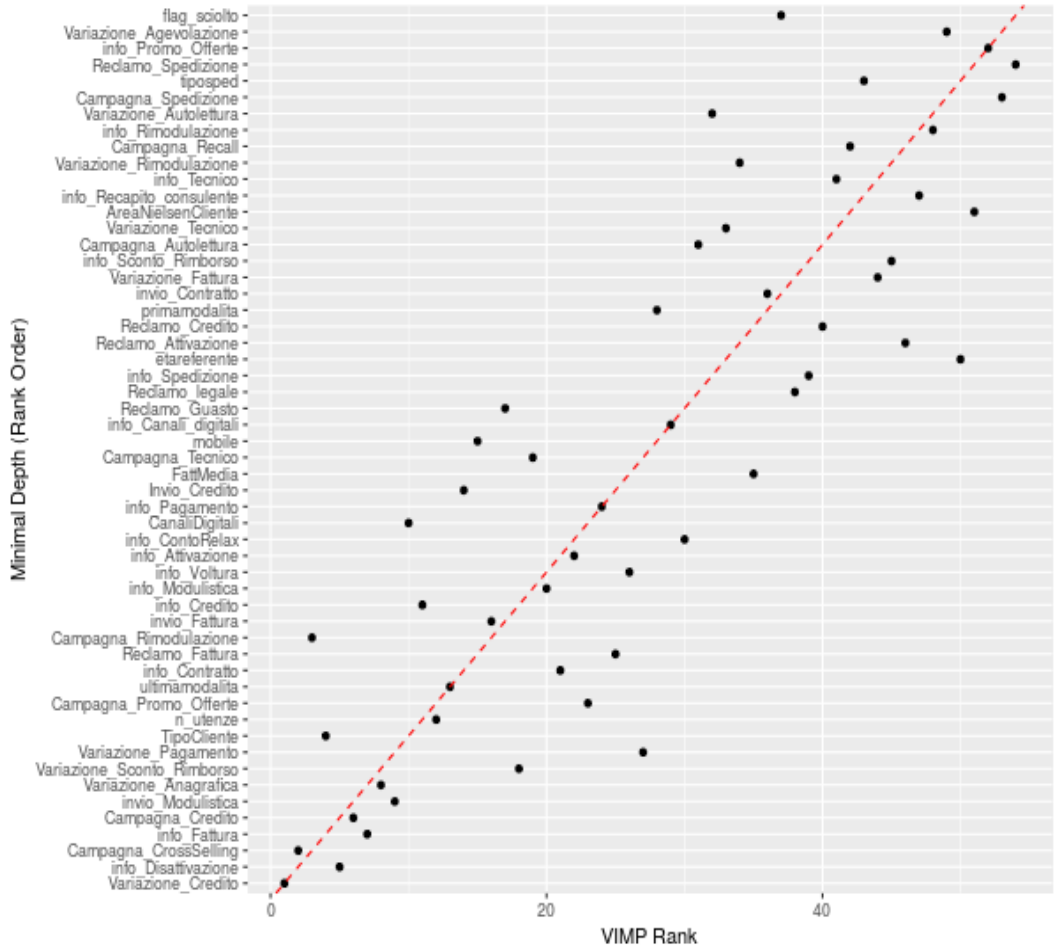

Figura 9 - Rankplot VIMP vs Minimal Depth delle variabili indipendenti inserite nella RSF

Ogni punto indica una variabile. Sull'asse delle ordinate le variabili sono ordinate dal basso verso l'alto in base al ranking di importanza definito dalla misura di Minimal Depth. Se un punto si trova a sinistra della diagonale tratteggiata si posiziona meglio nel ranking delle variabili definito dalla VIMP rispetto a quello definito dalla misura di Minimal Depth. Viceversa se un punto si trova a destra della diagonale si posiziona più in basso nel ranking delle variabili definito dalla VIMP rispetto a quello definito dalla misura di Minimal Depth. Il fatto che i punti siano concentrati intorno alla diagonale tratteggiata indica che i due indici sono concordi nel definire un ranking di importanza delle variabili. Questa concordanza

rassicura sulla credibilità del ranking di importanza delle variabili, che non è molto sensibile al cambiamento della misura utilizzata.

3.6 *Customer lifetime* e *lifetime value*

Uno dei principali scopi di questo lavoro di tesi è quello di stimare il valore prospettico del cliente per l'azienda. Questa misura viene definita l*ifetime value*. Il suo calcolo necessità di una stima dei mesi di sopravvivenza e di una misura del valore economico del cliente, espresso in termini di margine sul prodotto o servizio venduto.

Una volta ottenute le probabilità di sopravvivenza stimate per ogni tempo *t* considerato per ognuno dei clienti, si vuole quindi definire una stima del tempo in mesi prima che si verifichi l'evento terminale. Una misura che stima il tempo di sopravvivenza in mesi, tenendo presente che il range necessariamente varierà da 0 a 42 mesi, è dato dalla somma delle probabilità di sopravvivenza nei singoli mesi. In questo modo si può quindi definire una misura del *lifetime* stimato per ogni cliente. Chiaramente si tratta solo di una possibile misura. Si sarebbe potuto, ad esmpio, stimare il *lifetime* dichiarando l'evento terminale quando per l'osservazione *i*-esima si verifica che $t_i < 0.5$, così come per le misure di performance del paragrafo 3.3.

Per quel che riguarda la misura del valore economico del cliente, si farà riferimento al margine in euro ottenuto dalla vendita dei servizi ai clienti. Si tratta cioè della differenza tra il prezzo pagato dal cliente e le spese che l'azienda sostiene per erogare il servizio. Si è optato per l'utilizzo di un margine primario, in quanto non sono inclusi tutti i costi indiretti legati alla gestione del cliente, ai servizi post vendita, ai costi di acquisizione e così via, che se ritenuto necessario verranno sottratti o tenuti in considerazione in un'altra sede. Data la natura dei servizi offerti, il margine è ottenuto come media del ricavo marginale ottenuto in tredici mesi. Nello specifico i primi dodici mesi sono quelli che precedono

l’analisi, mentre il tredicesimo valore è ottenuto dal margine previsto per il mese successivo all’analisi. Lo scopo è quello di tenere in qualche modo dentro al calcolo del ricavo marginale medio, oltre che la storia del cliente, anche il ricavo previsto nel breve termine.

Per farsi un’idea dei valori assunti da questa variabile, si può fare riferimento alla sua distribuzione. Inoltre vengono presentate alcune statistiche descrittive nella *Tab. 2*.

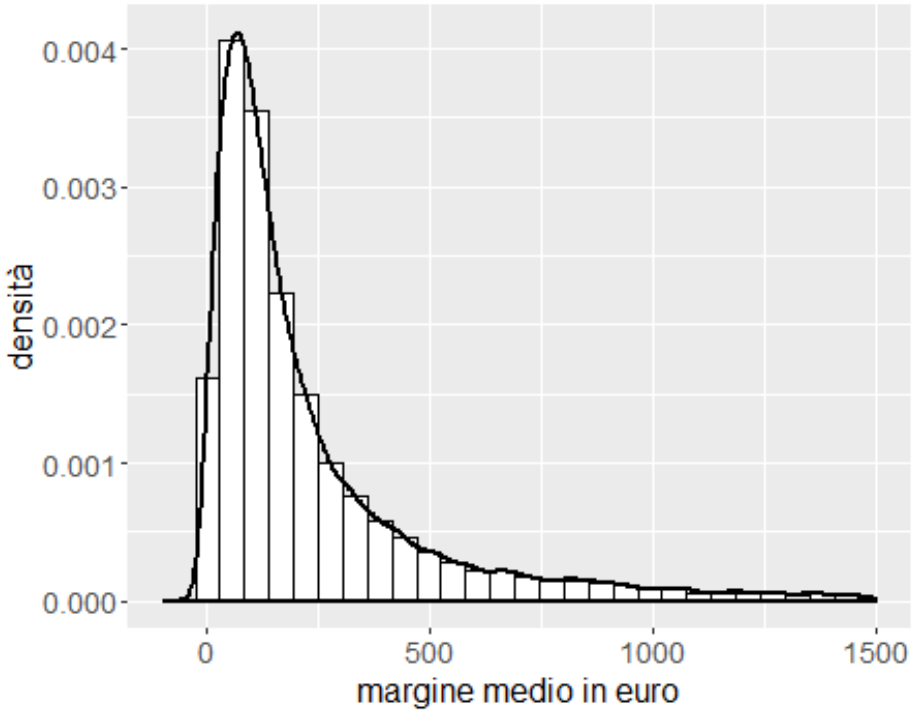

Figura 10 - distribuzione di densità di probabilità della variabile sul margine mensile medio in euro

La distribuzione è fortemente asimmetrica. Ha una forte concentrazione al di sotto dei 300 euro e la densità diventa molto schiacciata superati i 1000 euro. Il grafico è stato tagliato poiché il range di variazione molto elevato avrebbe reso il grafico illeggibile. Ci sono infatti un gran numero di valori estremi, che riguardano quei clienti Business sui quali si margina un valore molto più elevato della media.

Tabella 2 - statistiche di sintesi della variabile sul margine mensile medio in euro

Min	1Q	Mediana	Media	3Q	Max
-€184	€68.90	€138.90	€285.50	€300.70	€20 872

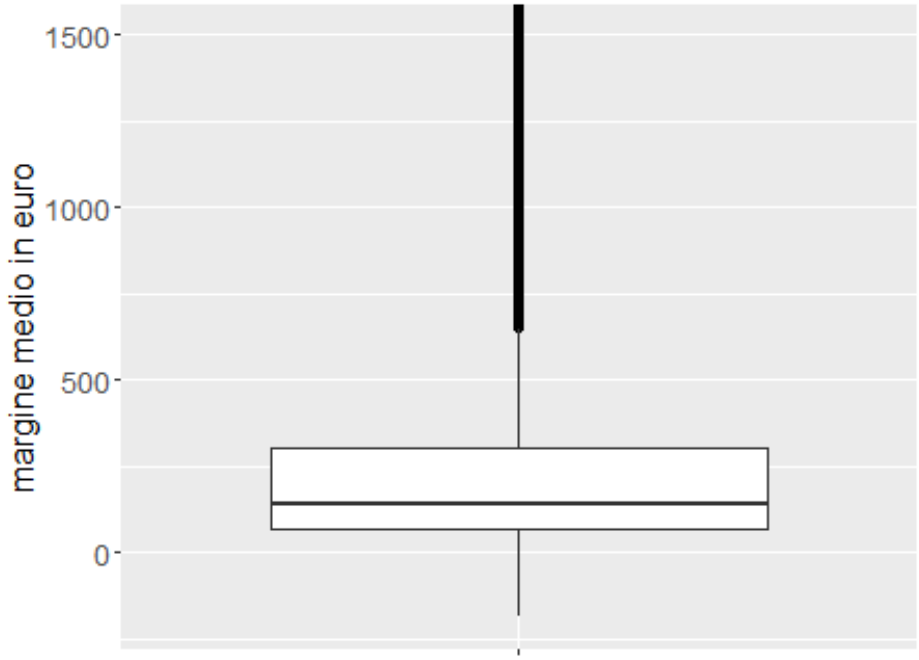

Figura 11 - boxplot variabile sul margine mensile medio in euro

Il range oscilla da un minimo di -184 euro (perdita) a un massimo di 20872 euro. Il margine può essere anche negativo a causa di quei clienti morosi o che hanno avuto offerte particolarmente vantaggiose di cui godono per i primi mesi.

Il ricavo medio dell'intero parco clienti PIB è di circa 140 euro mensili. Tra il primo e il terzo quartile si concentrano quei clienti sui quali il margine medio mensile è all'incirca tra i 70 e i 300 euro. I pallini più scuri, che nel boxplot vanno a disegnare una linea a causa della loro elevata numerosità, sono outliers.

A partire da questi dati il *Lifetime Value* è calcolato come somma dei valori delle probabilità di sopravvivenza stimate nei singoli mesi per il ricavo marginale medio:

$$\text{LTV} = MM \times \sum_{i=1}^{T} p_i \qquad (3.1)$$

Dati gli obiettivi dell'analisi, ha più senso discutere il valore del Lifetime Value per i soli clienti ancora attivi.

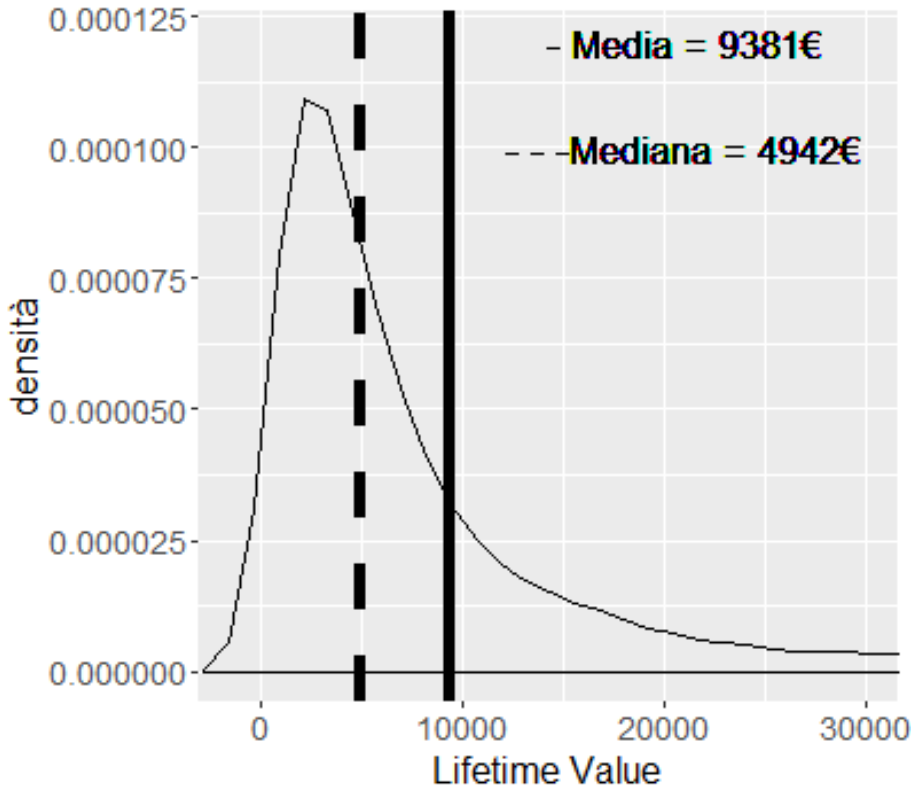

Figura 12 - distribuzione di densità di probabilità della variabile sul *lifetime value* stimato

La distribuzione, così come per il margine medio è asimmetrica. La distribuzione è molto concentrata al di sotto dei 10000 euro e diventa molto schiacciata superati i 20000 euro. Il range dei valori va da un minimo negativo di cirva -3000 euro a un massimo di oltre 600000 euro. Nel caso di valori che vanno molto al di sotto dello zero si tratta di situazioni in cui il cliente non paga o ha un offerta temporanea, situazioni che chiaramente non potrebbero protrarsi nel tempo, per cui questo valore non è verosimile. Nel

caso di valori sopra le centinaia di migliaia di euro si tratta di clienti molto particolari, che rappresentano una minima parte del parco. La media è pari a 9381 euro ma, a causa della forte influenza dei valori estremi, è bene considerare la mediana, che è pari a 4942 euro. Il 50% dei clienti attivi hanno valori che variano tra 2500 euro circa e 10000 euro circa.

Tabella 3 - statistiche di sintesi della variabile sul *lifetime value* stimato

Min	1Q	Mediana	Media	3Q	Max
-€2752	€2511	€4942	€9381	€10054	€620981

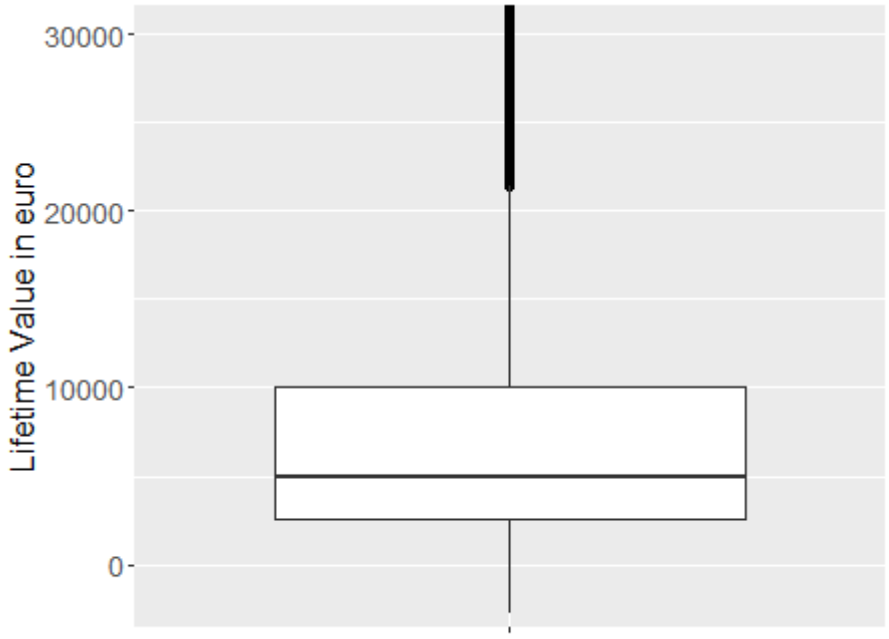

Figura 13 - boxplot della variabile sul *lifetime value* stimato

3.7 Definizione dei risultati più significativi

Si è parlato dell'analisi e dei risultati in termini di capacità predittiva della RSF. A questo punto si sono ottenute le stime della probabilità di sopravvivenza per ogni singolo cliente per ognuno dei tempi t_i considerati, si è fatta una stima del tempo di sopravvivenza (*lifetime value*) e si è ottenuta una misura del valore economico potenziale del cliente (*lifetime*

value). L'utilizzo di queste misure può essere utile in diverse fasi dei processi decisionali e può portare alla formulazione di un gran numero di strategie utili all'azienda.

Uno degli obiettivi, una volta ottenute queste stime, è quello di provare a influenzare o controllare le variabili esplicative, così da poter condizionare il valore della risposta. Appare naturale infatti che l'obiettivo primario dell'analisi della sopravvivenza, oltre che quello di una stima di probabilità, sia massimizzare il tempo di sopravvivenza[3] dei pazienti (o in questo caso il tempo prima che il cliente rescinda il contratto) e comprendere in che modo le variabili indipendenti influenzano il tempo di sopravvivenza.

Un altro obiettivo è quello di utilizzare queste stime per segmentare il parco clienti ed agire in modo specifico su ogni segmento. Si potrà decidere ad esempio di investire più o meno risorse su quei soggetti che hanno un valore di *lifetime value* basso in senso assoluto, o rispetto agli altri clienti.

In primo luogo, quindi, vanno tratte delle conclusioni dai risultati della RSF che permettano di ipotizzare un qualche tipo di relazione tra le variabili indipendenti e la risposta. Per riuscirci efficacemente è necessaria anche da una valutazione preliminare della natura delle variabili indipendenti. In secondo luogo vanno utilizzati i risultati della RSF per definire una segmentazione del parco clienti.

3.8 Studio delle relazioni tra le variabili della RSF

Il poter controllare o meno le variabili indipendenti, con lo scopo di influenzare il valore della risposta, dipende direttamente dalla natura delle variabili indipendenti stesse. Tra le variabili inserite in quest'analisi (si veda

[3] In alcuni casi, come è facile intuire, lo scopo potrebbe essere quello invece di minimizzare il tempo di sopravvivenza. È il caso ad esempio di uno studio sul tempo di sopravvivenza di un certo tipo di batterio pericoloso per la salute, oppure del tempo prima che un ticket aziendale venga gestito e risolto.

l'appendice A) alcune sono di tipo socio-anagrafico e chiaramente non possono essere influenzate da azioni specifiche da parte dell'azienda. La maggior parte delle altre variabili, invece, possono essere influenzate in modo più o meno diretto mediante specifiche azioni nei confronti dei clienti. Ad esempio, le variabili sul tipo di spedizione della fattura o sulla modalità di pagamento possono essere influenzate promuovendo l'utilizzo specifico di un tipo di modalità di spedizione o di pagamento a discapito di un altra, quelle sugli accessi ai canali digitali promuovendo o meno l'utilizzo dei canali digitali, e così via. Inoltre va ricordato che buona parte delle variabili riguardano il sistema di ticketing, quelli che fin ora sono stati chiamati casi, e che registrano l'interazione tra cliente e azienda. L'analisi di queste variabili può portare a risultati molto interessanti, e in molti casi si tratta di variabili controllabili o quantomeno influenzabili direttamente o indirettamente.

Come anticipato, bisogna trovare una qualche chiave di interpretazione della relazione tra queste variabili e la risposta. Le RSF, così come altri algoritmi sofisticati di *machine learning*, hanno molti vantaggi, di cui già si è parlato nel corso di questo lavoro di tesi, ma hanno anche un grosso difetto che ha a che fare con la difficoltà di interpretazione dei risultati ottenuti. Questi modelli raggiungono performance molto elevate, permettono di individuare e trattare relazioni non lineari, di tenere conto di un grosso numero di variabili e delle interazioni tra esse. Queste caratteristiche però implicano l'impossibilità di usare gli strumenti interpretativi dei classici modelli statistici. Non si può, tra le altre cose, parlare di effetto moltiplicativo delle covariate sulla risposta o di una qualche forma relazionale standard o ricorrente tra le variabili oggetto di studio. Questa caratteristica ha fatto spesso paragonare queste metodologie a una *black box*. In realtà restano comunque degli strumenti interpretativi a cui fare riferimento per aprire questa *black box* e comprenderne i meccanismi.

Un approccio che si sta sviluppano di recente è quello che si basa sull'assunto che ogni relazione a livello locale può essere considerata in qualche modo lineare. Su quest'idea è basata la libreria di Python *lime* (Thomas, Michael, 2019), disponibile anche in R. In questo lavoro però verrà utilizzato un approccio più descrittivo, mediante una strategia differente di analisi.

Lo strumento principale per capire i risultati del modello è stato di tipo grafico. Sarebbe meglio dire che è di tipo descrittivo e si serve di strumenti grafici. Per prima cosa sono stati presi in considerazione i risultati ottenuti dalla stima del *lifetime*. Questo, come spiegato, può assumere valori compresi tra un minimo di 0 a un massimo di 42 mesi ed è stato suddiviso in intervalli temporali di 5 mesi. In questo modo i soggetti per i quali la stima del *lifetime* è, ad esempio, compresa tra 10 e 15 mesi, ricadranno nella stessa classe di risposta. Si sono poi considerate le variabili indipendenti, ordinate in base ad un ranking definito dal valore di Variable Importance. Si è proceduto allora alla costruzione di una *heatmap*, con il *lifetime* ricodificato sull'asse delle ascisse e le covariate sull'asse delle ordinate. La ricodifica della variabile sul *lifetime* stimato non era strettamente necessaria, ma come si vedrà è utile a rendere il grafico più leggibile e l'interpretazione più agevole.

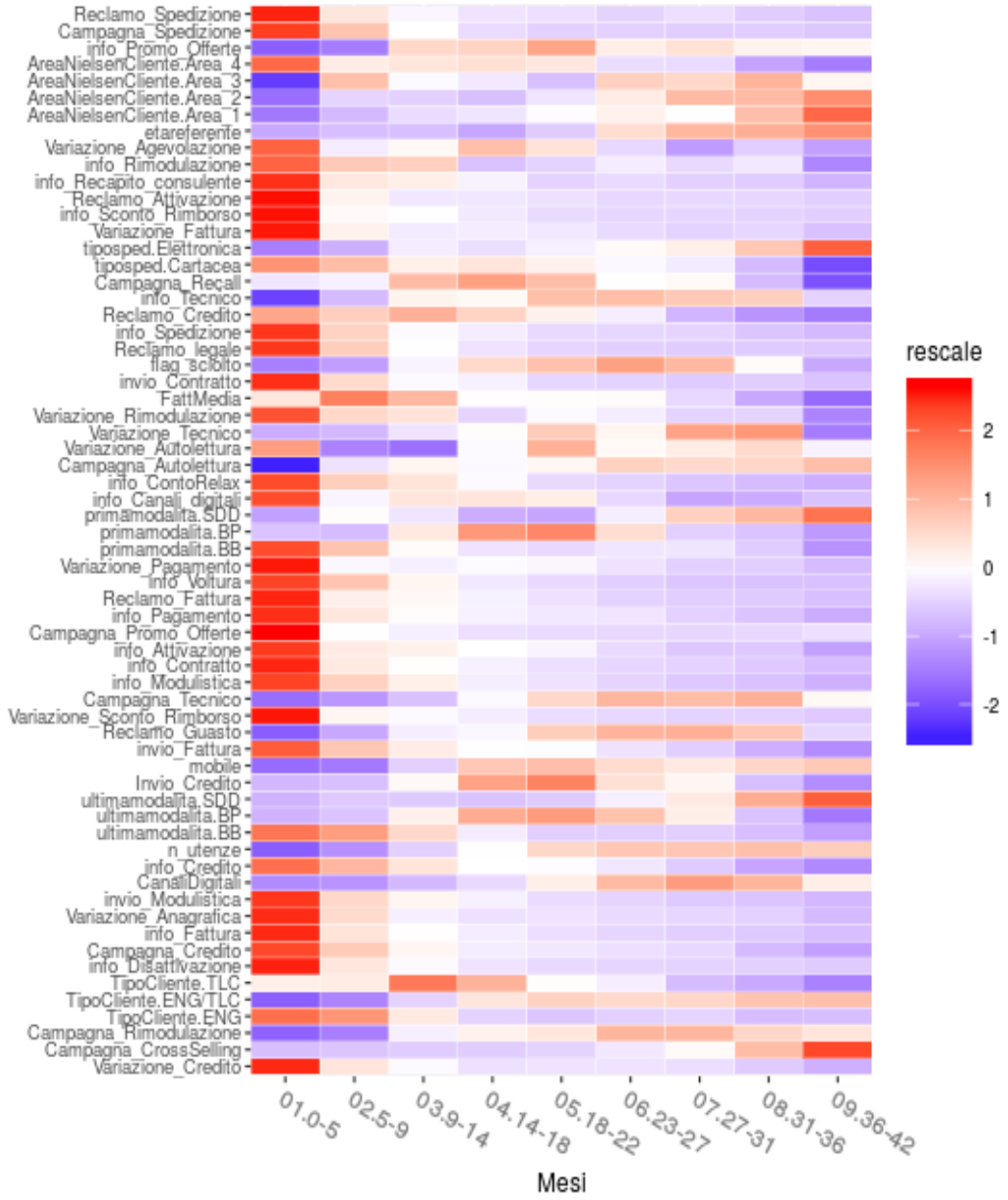

Figura 14 - Heatmap del lifetime stimato in classi di mesi vs variabili indipendenti inserite nella RSF

Le variabili sono ordinate in ordine di VIMP dal basso verso l'alto. Un colore rosso di intensità maggiore all'interno dei rettangoli indica che i clienti con quella stima del *lifetime* hanno un valore molto superiore rispetto alla media per quella variabile a cui il rettangolo corrisponde. Un colore vicino al bianco indica che questi valori sono in media. Infine un colore blu di intensità maggiore indica che i valori della variabile sono sotto media per

la variabile in questione data una certa stima del *lifetime*. Ad esempio, guardando al primo rettangolo in basso a destra, si può dire che per quei clienti per i quali la RSF stima un valore di *lifetime* compreso tra 0 e 5 mesi, è stato effettuato un numero medio di variazioni del credito molto maggiore rispetto alla media dei clienti.

Questo chiaramente vale per le variabili numeriche. Per le variabili categoriali viene riportata una riga per ogni categoria della variabile stessa. L'interpretazione tuttavia è simile. Nel caso della variabile che registra il tipo di cliente le categorie possibili sono tre: cliente Energia e/o Gas (ENG), cliente Voce, Adsl e/o Mobile (TLC) o cliente misto (ENG/TLC). In questo caso, ad esempio, il colore tendente al blu per la riga all'intreccio tra la categoria ENG/TLC e la stima del *lifetime* tra 0 e 5 mesi, indica che nel cluster di clienti per i quali la RSF stima tempo di permanenza basso ci sono meno clienti con servizio misto ENG/TLC rispetto alla media del parco clienti.

Questo tipo di strategia interpretativa a primo acchito potrebbe sembrare un po' complessa o macchinosa. Tuttavia è parsa a chi scrive essere piuttosto efficace, sia per gli scopi dell'analisi dei risultati che per i vantaggi derivanti dalla visualizzazione diretta. La lettura dei risultati risulta molto più immediata una volta compreso il meccanismo. Infatti l'interpretazione risulterebbe molto complessa anche se si avessero a disposizione i classici coefficienti di regressione, specie a causa dell'elevato numero di variabili indipendenti, tra le quali sono presenti anche alcune variabili categoriali, che farebbero aumentare ulteriormente il numero di coefficienti da stimare e da interpretare.

Prima di procedere però va fatta qualche ulteriore considerazione. Si potrebbe pensare che nonostante la premessa ci si sia ricondotti in qualche modo a un'interpretazione della relazione tra le variabili e la risposta di tipo lineare. Qui non si vuole stabilire un rapporto causa effetto. L'interpretazione non può essere del tipo «all'aumentare del numero medio

di casi mensilmente aperti di variazioni del credito diminuisce il tempo di permanenza stimato del cliente», piuttosto si potrà formulare una frase del tipo «tra le *caratteristiche* dei soggetti per i quali il tempo di permanenza stimato è più basso c'è quella di aver aperto un numero di variazioni del credito superiori alla media del parco clienti».

La differenza è sottile ma sostanziale, perché non viene ipotizzata una relazione diretta. Si potrebbe piuttosto dire che la variabile che registra i casi di variazione credito è molto importante nel determinare il tempo di permanenza perché prima nel ranking definito dalla VIMP e che, inoltre, è verosimile che il verificarsi di casi di variazione del credito facciano registrare valori bassi sul tempo di permanenza stimato. Insomma, si invita alla cautela in quanto non c'è un modo sistematico e comprovato per definire la relazione tra la variabile e la risposta a causa della natura stessa della metodologia impiegata.

Detto ciò si può passare al commento della *heatmap*. Il numero di casi di variazione del credito aperti, si è detto, è molto maggiore della media per quei clienti con *lifetime* stimato fino a 5 mesi. Per gli altri clienti è in media, ma per i clienti per i quali si stima un tempo di permanenza più elevato è inferiore alla media. Probabilmente dipende dal fatto che i clienti che rescindono prima il contratto possono essere cattivi pagatori e quindi viene aperta una qualche azione di credito, che può variare dalla messa in mora all'approvazione di una dilazione nei pagamenti.

Nel caso dei casi aperti per una campagna di cross selling si verifica l'opposto. I clienti con tempo di permanenza stimato più basso hanno ricevuto meno campagne rispetto alla media del parco, mentre se il tempo stimato è maggiore di 30 mesi si collocano sopra la media. Si potrebbe pensare quindi che la campagna cross selling in qualche modo allunghi il tempo di permanenza. Tuttavia è molto più verosimile che quest'ultima venga fatta maggiormente per quei clienti che già sono ritenuti buoni clienti dal marketing per tutta una serie di criteri specifici.

I clienti che hanno un *lifetime* stimato inferiore a 9 mesi hanno subito meno campagne di rimodulazione rispetto alla media. Da 10 a 22 mesi questa variabile resta in media, mentre per i clienti con una stima superiore ai 22 mesi si attesta a un valore superiore alla media, con una tendenza al ritorno in media per quelli con una stima superiore ai 36 mesi. Questo può dare un sentore del fatto che la campagna di rimodulazione sembra avere un effetto positivo, magari perché quando si intercettano eventuali problemi nella definizione della taglia (e quindi dei consumi e del canone) si mette in atto un'azione di *customer care* efficace.

Per quel che riguarda la tipologia di cliente, tra quelli con una stima inferiore del *lifetime* ci sono più clienti con sole utenze Energia e/o Gas della media e meno con utenze miste. Questa tendenza si inverte all'aumentare del valore del *lifetime* stimato. I clienti solo TLC in media sono presenti meno nel gruppo di quelli con *lifetime* stimato inferiore a 9 mesi mentre sono più della media per il gruppo per cui si stimano dai 9 ai 18 mesi. Si scende sotto la media per quelli la cui stima di permanenza è maggiore di 23 mesi. Sembrerebbe quindi che la RSF stimi tempi di permanenza maggiori per i clienti con utenze miste ENG/TLC. Per i clienti solo TLC, il risultato potrebbe essere compreso tenendo presente che nel mercato dei soli TLC c'è un maggiore *turnover* di clientela.

I clienti con permanenza stimata inferiore a 5 mesi hanno aperto casi di informazione sulla disattivazione di una o più utenze molto più della media del resto del parco. Si potrebbe pensare che questo genere di ticket siano un sentore di disdetta specie all'inizio della storia del cliente, infatti di solito chi disattiva un'utenza all'inizio le disattiva tutte, mentre può capitare che chi è cliente da più tempo cambi ad esempio solo il gestore telefonico senza cambiare le utenze Energia e/o Gas.

Per il commento dei casi di campagna sul credito e di informazione sul credito possono essere fatte considerazioni molto simili a quelle fatte per quelli di variazione del credito.

Il numero di casi di informazione sulla fattura aperti è molto maggiore della media per quei clienti con *lifetime* stimato fino a 5 mesi. Per gli altri clienti è in media, mentre per i clienti per i quali si stima un tempo di permanenza maggiore è inferiore alla media. Sembrerebbe che i clienti per i quali si stima un tempo di permanenza inferiore chiedano più spesso delucidazioni sulla fattura, ad esempio perché non comprendono bene le voci in fattura o ricevono fatture con importi diversi da quello che si aspettavano.

Il trend è quasi uguale per le variabili che registrano i casi aperti di variazione dell'anagrafica e di invio della modulistica. Nel primo caso sarebbe necessaria un'analisi più approfondita poiché non è facile capire il motivo di un'alta concentrazione di variazioni sui dati di anagrafica per quei clienti la cui stima di permanenza è bassa. Potrebbe accadere, ad esempio, che in fase di immissione dei dati nei sistemi aziendali si sia sbagliato a inserire l'indirizzo e quindi il cliente non riceve la fattura. Nel secondo caso la richiesta esplicita di modulistica potrebbe essere legata a un qualche problema, per la cui soluzione è necessario presentare un qualche tipo di modulo. La richiesta esplicita è anche indice del fatto che il cliente non ha trovato autonomamente il modulo on-line, per cui potrebbe ad esempio aver avuto problemi nell'interazione con i canali digitali.

I clienti che si stima restino meno hanno effettuato meno accessi ai canali digitali rispetto alla media. Man mano che si considerano gruppi di clienti con stime maggiori gli accessi tendono ad essere più alti della media, tranne che per quelli con stime di permanenza superiori ai 36 mesi, per i quali gli accessi sono in media con il resto del parco clienti. Pare che i clienti che si stima rescindano prima il contratto facciano meno uso dei canali digitali rispetto agli altri clienti. Viceversa tra quelli con stime maggiori si trovano utilizzatori più assidui dei canali digitali.

I clienti con stime di permanenza basse sono quelli che hanno un numero di utenze inferiore alla media del parco. Viceversa, più i tempi di

permanenza stimati sono elevati, più si tendono a trovare clienti con un numero maggiore di utenze rispetto alla media del parco. Questo risultato è in linea con gli altri risultati, come quello sul tipo di cliente a utenze miste. D'altronde è plausibile che clienti con più utenze cambino fornitore di servizi più difficilmente.

La variabile che registra l'ultima modalità di pagamento si colloca anch'essa tra le prime dieci variabili per Variable Importance. Le categorie della variabile sono: bonifico postale, bollettino bancario e addebito diretto su conto corrente. Per stime di permanenza basse si registra presenza di clienti che pagano con bollettino postale superiore alla media del parco, mentre la presenza di clienti che utilizzano le altre due modalità di pagamento è sotto media. Per stime di permanenza intermedie è la presenza di clienti che pagano con bonifico bancario a essere sopra la media. Infine per stime molto elevate è maggiormente presente la forma di pagamento dell'addebito diretto su conto corrente. Questo tipo di modalità di pagamento, ad esempio, evita ritardi nei pagamenti, è più comoda e, probabilmente, potrebbe dare meno l'impressione che un mese si paghi una bolletta un po' più alta del solito.

Per quel che riguarda la presenza di un'utenza mobile, i clienti con tempi stimati di permanenza più bassi hanno in meno occasioni della media anche un'utenza mobile. Il risultato sembra essere in linea con quello sul numero delle utenze. Inoltre spesso quando si ha un contratto integrato il bundle mobile viene dato a costo zero o quasi, per cui potrebbe essere determinate nel prolungare il tempo di permanenza.

Fin qui si sono commentati i risultati delle prime quindici variabili in ordine di VIMP. A questo punto dovrebbe essere abbastanza il tipo di approccio adottato nell'interpretazione dei risultati. Si procederà a commentare il resto delle variabili in modo più sintetico, soffermandosi sulle variabili e i risultati più interessanti.

Un risultato inaspettato riguarda il fatto che i clienti con stime di permanenza inferiori a 9 mesi hanno aperto meno casi di reclamo per il guasto di un'utenza rispetto alla media del parco. Quelli con stima da 9 a 18 mesi sono in media, mentre quelli con stima compresa tra i 18 e i 36 mesi hanno aperto più casi di reclamo per guasto rispetto alla media. Sembrerebbe che l'apertura del reclamo per guasto incida positivamente sul tempo di permanenza. In realtà un'analisi più attenta riconduce questo risultato al fatto che è poco probabile che si verifichi un guasto dopo poco tempo dall'attivazione dell'utenza. Di conseguenza, quei clienti da cui la RSF ha "appreso" quali sono le caratteristiche che possono portare a rescindere prima il contrato, avranno sperimentato raramente l'esperienza del guasto dell'utenza. Bisognerebbe indagare meglio l'effetto del guasto sulla propensione al *churn*, tuttavia è abbastanza banale immaginare che sia piuttosto rilevante, specie se si verifica più di un guasto in tempi brevi. Attraverso un'analisi più approfondita si potrebbe scoprire, ad esempio, che molte persone se sono clienti da poco e hanno subito un guasto all'utenza decidono di cambiare gestore senza neanche sporgere reclamo, per cui si tratterebbe di una variabile molto importante ma che non viene registrata dal sistema di ticketing.

Un altro risultato interessante è quello del caso di richiesta di informazioni sull'attivazione. Il gruppo di soggetti con tempo di permanenza stimato inferiore a 5 mesi ha fatto questo tipo di richieste con frequenza molto maggiore rispetto alla media del parco. Per il resto dei gruppi con tempo stimato maggiore questa variabile è in media o leggermente inferiore alla media del parco. I ritardi di attivazione delle utenze sembrano avere un effetto importante sul tempo di permanenza, infatti risultati molto simili si riscontrano anche per la variabile che registra il numero medio di reclami per l'attivazione di un'utenza. A partire da questo risultato si potrebbe fare un'analisi più approfondita sui tempi e i

problemi tecnici di attivazione delle varie utenze nelle diverse zone di erogazione dei servizi.

Come prevedibile, è rilevante anche la variabile che registra la fattura media. I clienti con tempo di permanenza stimato inferiore a 14 mesi pagano fatture più elevate rispetto alla media del parco. Quelli con tempo stimato tra i 14 e i 27 medi pagano fatture di importo in media rispetto a quelle dell'intero parco, mentre quelli con tempi stimati maggiori, vale a dire più di 27 mesi, pagano fatture inferiori rispetto alla media. Sembrerebbe quindi che i clienti con fatture più elevate e sui quali probabilmente si margina di più, sono anche quelli che restano meno. Questo risultato comporta tutta una serie di possibili considerazioni sulle politiche di *pricing* aziendali. Di solito sarebbe conveniente mantenere ricavi più bassi ma fidelizzare il cliente, poiché si solito i costi di acquisizione sono elevati, e nel medio e lungo periodo si riescono a ottenere margini più elevati. Inoltre l'immagine aziendale non può che trarre beneficio da un parco maggiormente fidelizzato, specie in un mercato in cui il posizionamento di marketing è fatto soprattutto sul risparmio in bolletta e risulta difficile trattenere i propri clienti nel caso in cui i competitor adottano strategie di riduzione dei prezzi.

I soggetti con stima di permanenza inferiore a 9 mesi hanno o hanno avuto in meno occasioni rispetto alla media del resto dei gruppi anche un'utenza contrattualizzata come Prodotto Sciolto. Una presenza di clienti di questo tipo superiore alla media si riscontra per quelli con un tempo di permanenza stimato tra i 14 e i 31 mesi. Sembrerebbe che avere o aver avuto un'utenza Prodotto Sciolto caratterizzi quei clienti con *lifetime* medio-alto, tuttavia quelli con *lifetime* superiore a 36 mesi hanno sperimentato il solo Prodotto Integrato in più casi rispetto alla media del parco.

Un altro risultato rilevante riguarda la modalità di spedizione della fattura. I soggetti con tempo stimato nella parte inferiore del range

considerato, vale a dire sotto i 22 mesi, ricevono la fattura tramite spedizione cartacea in più casi rispetto alla media del parco. Viceversa i soggetti con tempo di permanenza stimato maggiore di 27 mesi ricevono la fattura in formato elettronico in più casi rispetto alla media del parco. Uno dei motivi potrebbe essere che la fattura cartacea può non arrivare a destinazione o arrivare in ritardo, determinando ritardi nel pagamento e comunque la percezione di un servizio inefficiente. Infatti i soggetti con mesi di permanenza stimata molto bassi hanno anche fatto molte più richieste di invio della fattura rispetto alla media del parco clienti. Nel caso di fatturazione elettronica è assai difficile che il cliente non riesca a reperire la fattura.

Inoltre, per il gruppo di clienti con tempi di permanenza stimati inferiori a 5 mesi, si riscontra una frequenza di richieste di informazioni sul recapito del consulente della rete vendita molto maggiore rispetto alla media del parco. Questo risultato è indicativo del fatto che, quando il consulente che fa sottoscrivere il contratto al cliente non si rende reperibile, si ha un effetto molto negativo sui tempi di permanenza. La questione è legata ad una più ampia che riguarda la gestione della rete vendita. Chiaramente quando i consulenti vendono in modo poco appropriato, o non forniscono supporto post vendita all'acquirente, si possono generare tutta una serie di effetti negativi, specie per quel target di clienti che fa particolarmente affidamento sul contatto personale con il singolo venditore piuttosto che con l'azienda in generale.

Per quel che riguarda l'età del cliente, i clienti che hanno tempo di permanenza stimato che ricade nella parte inferiore del range dei 42 mesi sono maggiormente clienti giovani. Viceversa quelli che hanno stime più elevate sono mediamente più grandi di età rispetto alla media del parco clienti. Questo risultato potrebbe dipendere dal fatto che i più giovani sono più inclini al cambiamento di gestore o tendono maggiormente a ricercare offerte più convenienti sul mercato. Poiché non è possibile agire su una

variabile del genere, questo risultato è utile più che altro per capire, ad esempio, su quale target puntare. Si potrebbe ad esempio decidere che si vuole maggiormente fidelizzare i giovani proponendo servizi o sconti particolari o, viceversa, si potrebbe decidere che gli sforzi maggiori devono essere indirizzati ai clienti più in avanti con l'età per consolidare la loro tendenza a restare più a lungo clienti dell'azienda.

Un'ultima variabile su cui si vuole puntare l'attenzione è quella che registra l'area di residenza dei clienti. Tra quelli con tempo di permanenza stimato inferiore a 5 mesi sono presenti residenti al sud e Sicilia (Area Nielsen 4) in quantità molto maggiore rispetto alla media del parco. Viceversa i residenti nel resto d'Italia sono presenti in misura maggiore in quei gruppi con stima di sopravvivenza elevata. Non è facile trovare una spiegazione univoca per questo risultato. Potrebbe, ad esempio, dipendere dalle caratteristiche della rete vendita che opera nelle diverse aree, oppure dalle caratteristiche delle aziende che compongono il sottoinsieme di clienti (si ricorda l'analisi riguarda i soli clienti business).

3.9 Segmentazione del parco clienti

Una volta discussi i principali risultati dell'analisi, si può passare alla descrizione delle strategie di segmentazione del parco clienti adottate. Come anticipato la segmentazione è basata sulla misura di *lifetime value*, le cui modalità di calcolo sono state descritte nel paragrafo 3.6. Sulla base di un confronto con le altre funzioni aziendali, sono stati stabiliti dei valori soglia (*threshold*) per definire i segmenti. Si è optato per una divisione in cinque segmenti, che hanno preso rispettivamente i nomi di segmento *Iron, Bronze, Silver, Gold* e *Platinum*. La segmentazione può essere rappresentata mediante una piramide con alla base il segmento *Iron* e al vertice quello *Platinum*.

Figura 154 - Descrizione di tipo piramidale della segmentazione del parco clienti sulla base della misura di lifetime value

La rappresentazione a piramide è efficace perché i valori soglia sono stati scelti in modo tale da formare una larga fascia di clienti a basso *lifetime value*, per poi definire segmenti sempre meno numerosi ma a *lifetime value* maggiore. A seguire si farà riferimento di nuovo al solo parco clienti ancora attivo, poiché, chiaramente, la parte di analisi del valore economico potenziale è utile a fini pratici solo su questo sottoinsieme dei dati analizzati.

Tabella 4 - Percentuale di clienti e lifetime value per segmento

PIB	Iron	Bronze	Silver	Gold	Platinum
% clienti	51,31%	26,79%	11,82%	5,78%	4,70%
% lifetime_value	14,34%	21,57%	18,94%	14,83%	30,32%

Dalla tabella 4 si legge che, dati i valori soglia scelti, il segmento *Iron* rappresenta più del 50% dell'intero parco clienti business con contratto Prodotto Integrato e, tuttavia, vale solo poco più del 14% del totale del valore economico potenziale stimato sull'intero parco PIB. Il segmento *Gold* vale la stessa percentuale del *lifetime value* stimato totale ma è composto da meno del 6% del parco. Il segmento *Platinum* invece vale da solo più del 30% del valore economico potenziale stimato pur essendo composto da meno del 5% del parco.

Queste considerazioni sono di cruciale importanza per le decisioni strategiche aziendali, che vanno abbinate ai risultati descritti dalla *Heatmap* per definire delle strategie di *customer care* congruenti con gli obiettivi aziendali prefigurati.

La *Heatmap* in *Fig. 13* può essere replicata inserendo sull'asse delle ascisse il cluster di appartenenza al posto del *lifetime* stimato, e può essere utilizzata in modo simile a quanto descritto poco sopra.

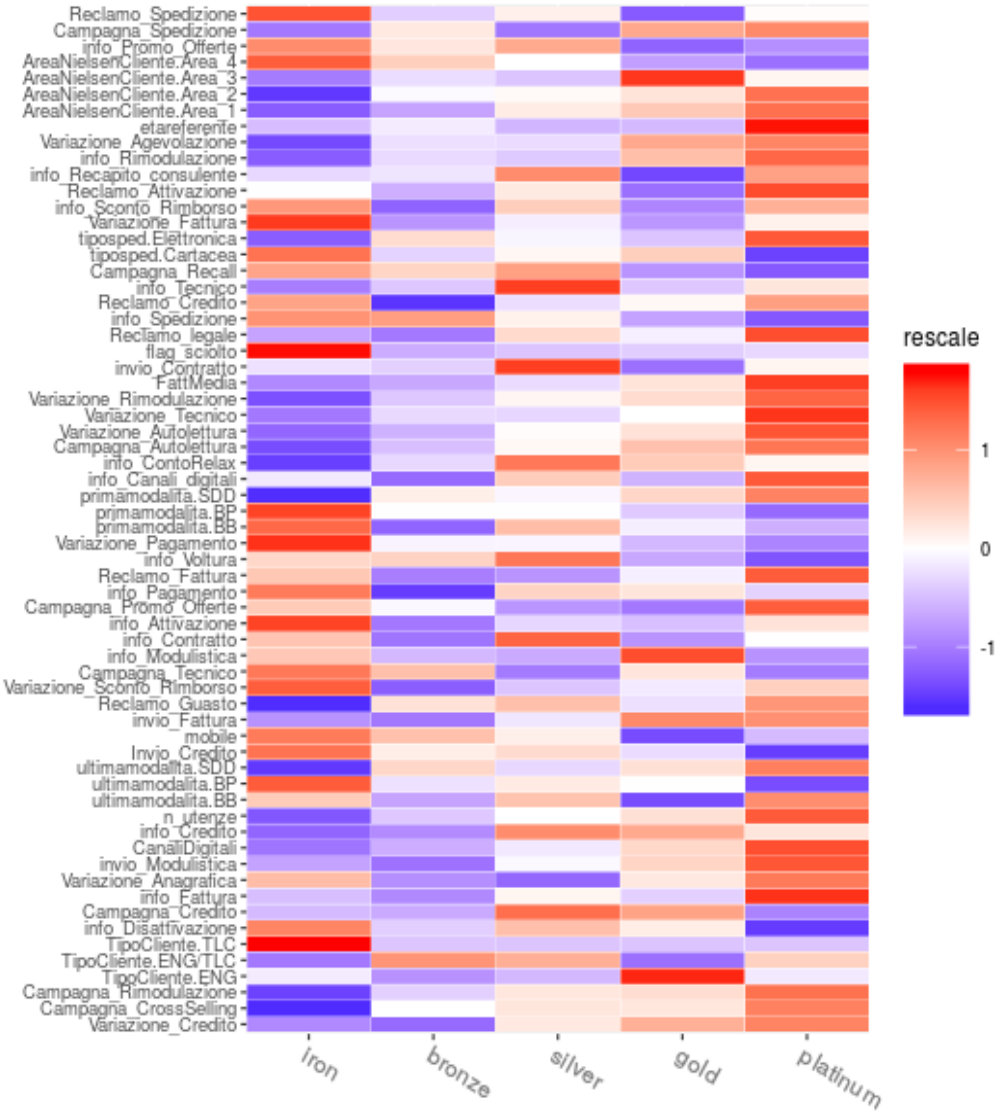

Figura 1516 - Heatmap dei segmenti basati sul valore di *Lifetime Value* vs variabili indipendenti inserite nella RSF

In questo modo si potranno definire strategie sui singoli cluster sulla base delle variabili che caratterizzano maggiormente i clienti che li compongono. Tuttavia bisogna utilizzare con attenzione questo strumento perché è pur sempre il risultato di due misure diverse, vale a dire il *lifetime* e il margine mensile medio. Va quindi tenuto presente che nel cluster dei *Platinum* potrebbero ricadere clienti sui quali si fa un margine estremamente elevato ma che hanno un'altrettanta elevata probabilità di abbandono nel breve

periodo. Viceversa potrebbe accadere che nel cluster degli *Iron* ci siano clienti altamente fidelizzati ma sui quali si fa un margine minimo.

Sulla base di questa considerazione si possono pensare non solo semplici strategie che puntano allo "spostamento" dei clienti verso i segmenti più alti della piramide. Si potrebbe ad esempio decidere di applicare degli sconti anche a dei clienti *Platinum*, per aumentare la fidelizzazione a discapito di una parte del margine nell'immediato, per poi marginare di più nel medio e nel lungo periodo. Allo stesso modo si può decidere di non applicare alcuno sconto ai clienti *Iron*, ma di continuare a proporre dei servizi aggiuntivi per cercare di aumentare il margine e, allo stesso tempo, applicare delle strategie di caring che mirino sull'offerta di servizi aggiuntivi piuttosto che su sconti e promozioni, dato che non è conveniente continuare ad abbassare il margine che si ottiene da questo tipo di clienti.

Conclusioni

In questo lavoro di tesi è stata svolta un'analisi approfondita di alcuni aspetti del parco clienti di una società multiutility che opera principalmente nel settore energetico e delle telecomunicazioni. La parte più importante dell'analisi consiste in un'analisi statistica della sopravvivenza e, nello specifico, in un'applicazione della metodologia Random Survival Forest (Ishwaran et al., 2008).

Attraverso questo strumento si è giunti a una stima della probabilità di abbandono dei clienti (*customer churn*) in un arco temporale di 42 mesi. Sulla base di questi risultati si è ottenuta una stima del tempo prima che il cliente rescinda il proprio contratto (*lifetime*). Si è proceduto poi al calcolo di una misura del valore economico potenziale dei singoli clienti (*lifetime value*). Infine, sulla base di questa misura si è operata una segmentazione del parco clienti in 5 *clusters*.

Questi risultati forniscono dei fondamentali strumenti per la costruzione di tutta una serie di strategie aziendali di azione nei confronti dei clienti. Per la valutazione degli aspetti cruciali su cui agire è stata fatta un'analisi approfondita delle caratteristiche dei clienti condizionata alle stime del tempo di permanenza. Sulla base di questo tipo di analisi, dei valori di *lifetime* e *lifetime value* e, non per ultimo, del *cluster* di appartenenza dei clienti, si è proceduto alla costruzione di una serie di azioni possibili nei confronti dei clienti. Queste azioni vengono proposte ai clienti dagli operatori del servizio clienti, con il supporto di un'apposita funzione sviluppata sulla piattaforma di *Customer Relationship Management* aziendale, la cui discussione esula dagli scopi di questo lavoro di tesi.

Lo scopo primario è quello di riuscire ad applicare delle azioni di *customer care*, che puntino alla fidelizzazione del cliente e quindi alla massimizzazione del tempo di permanenza e, di conseguenza, del suo

valore economico potenziale, innescando un circolo virtuoso che possa coinvolgere diversi ambiti di business.

In conclusione, va sottolineato che l'analisi è pur sempre provvisoria e sarà posta a una revisione periodica, con lo scopo di ottenere miglioramenti non solo delle performance statistiche, ma anche di carattere più qualitativo. Si potrà procedere, ad esempio, trasformando, rimuovendo o aggiungendo degli aspetti non considerati in prima battuta nell'analisi qui presentata. Un contesto aziendale è in continuo mutamento ed è necessario applicare analisi statistiche complesse che riescano a tener conto della complessità crescente del mercato. Appare poi cruciale un'intensa attività di monitoraggio delle performance del modello, specie in una prima fase, ma anche nel medio e lungo periodo. Gli algoritmi statistici come quello sviluppato in questo lavoro, grazie allo sviluppo di capacità di calcolo e strutture IT avanzate degli ultimi anni, sono "incastonati" in una serie di processi aziendali in produzione. Per gli statistici si aprono tutta una serie di sfide inedite che hanno a che fare con l'integrazione delle proprie analisi in un sistema complesso e in continuo mutamento, alle quali è necessario rispondere con strumenti ibridi, che spesso si collocano tra la statistica e l'informatica.

Riferimenti Bibliografici e Sitografici

Ansell J., Harrison T, Archibald T. (2007). *Identifying cross-selling opportunities, using lifestyle segmentation and survival analysis.* Marketing Intelligence & Planning. 25 pp. 394-410.

Bogaerts K., Komarek A., Lesaffre E. (2018). *Survival Analysis with Interval-Censored Data: A Practical Approach with Examples in R, SAS, and BUGS*. Chapman and Hall/CRC, Boca Raton, Florida.

Breiman L (2001a). *Random Forests*. Machine Learning, 45(1), 5–32.

Breiman L. (1996). *Bagging predictors*. Machine Learning 26 123–140.

Breiman L. (2001). Random forests. Machine Learning 45 5–32.

Breiman L., Friedman J. H., Olshen R. A., Stone C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, California.

Collett D., (2013). *Modelling Survival Data in Medical Research. Third edition*, Chapman & Hall, London.

Cox D. R., Oakes D. (2018). *Analysis of Survival Data*. Chapman and HALL/CRC, New York.

Cox D.R. (1972). *Regression models and life-tables (with discussion)*. Journal of the Royal Statistical Society B, 34, 187-220.

Ehrlinger J., (2016). *ggRandomForests: Exploring Random Forest Survival*. URL arXiv:1612.08974

Hadley W. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016. URL CRAN.R-project.org/package=gg plot2

Hadley W. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1.* URL CRAN.R-project.org/package=tidyverse

Hadley W., Romain F., Lionel H. and Kirill M. (2018). *dplyr: A Grammar of Data Manipulation. R package version 0.7.6.* URL CRAN.R-project.org/package=dplyr

Harrell F., Califf R., Pryor D., Lee K., Rosati R. (1982). *Evaluating the yield of medical tests.* J. Amer. Med. Assoc. 247 2543–2546.

Ishwaran H. (2007). *Variable Importance in Binary Regression Trees and Forests*. Electronic Journal of Statistics, 1, 519–537.

Ishwaran H., Kogalur U. B. (2007). *Random survival forests for R*. Rnews 7 25–31.

Ishwaran H., Kogalur U. B. (2008). *RandomSurvivalForest* 3.2.2. R package.

Ishwaran H., Kogalur U. B., (2019). *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*. R package version 2.9.1. URL cran.rproject.org/web/packages/randomForestSRC/randomForestSRC.pdf

Ishwaran H., Kogalur U. B., Blackstone E. H., Lauer M. S. (2008). *Random Survival Forest.* The Annals of Applied Statistics Vol. 2, No. 3, 841–860.

Jerald F. Lawless (2002). *Statistical Models and Methods for Lifetime Data, Second Edition (Wiley Series in Probability and Statistics)*. Wiley Series in Probability and Statistics.

Junxiang L. *Modeling Customer Lifetime Value Using Survival Analysis − An Application in the Telecommunications Industry*. Paper 120-28 URL support.sas.com/resources/papers/proceedings/proceedings/sugi28/120-28.pdf

Junxiang L. *Predicting Customer Churn in the Telecommunications Industry — An Application of Survival Analysis Modeling Using SAS*. Paper 114-27. URL support.sas.com/resources/papers/proceedings/

Kaplan E. L., Meier P. (1958). *Nonparametric estimation from incomplete observations*. Journal of the American Statistical Association, 53, 457-481.

Machin D, Cheung Y. B., Parmar M. (2006) .*Survival Analysis: A Practical Approach, Second Edition*. John Wiley & Sons, Chichester, England.

Mantel N., Haenszel W. (1959). *Statistical aspects of the analysis of data from retrospective studies of disease*. Journal of the National Cancer Institute, 22, 719- 748.
proceedings/sugi27/p114-27.pdf

Ribeiro M. T., Singh S., Guestrin C.(2016). *Why Should I Trust You?: Explaining the Predictions of Any Classifier*. URL arxiv.org/pdf/ 1602.04938.pdf

Schmid M., Wright M. N., Ziegler A. (2016). *On the use of Harrell's C for clinical risk prediction via random survival forests.* Expert Systems with Applications: An International Journal archive, Vol. 63 Issue C, pp. 450-459.

Therneau T. (2015). *A Package for Survival Analysis in R. version 2.38.* URL: CRAN.R-project.org/package=survival

Thomas L. P. and Michaël B. (2019*). lime: Local Interpretable Model-Agnostic Explanations. R package version 0.5.0.* URL CRAN.R-project.org/package=lime

Tripepi G., Catalano F. (2004). *L'analisi di sopravvivenza con il metodo di Kaplan-Meier*. Giornale Italiano di Nefrologia n. 6, pp. 540-546.

Appendice A: descrizione completa delle variabili inserite nella Random Survival Forest

Cliente attivo: il cliente è considerato attivo se ha un'utenza energia, gas, voce o adsl attiva. I clienti con solo mobile attivo sono considerati disattivi se hanno disattivato tutte le utenze ad eccezione di quella mobile. Si tratta infatti di clienti che sono di fatto disattivi ma che non disattivano la utenza mobile. I clienti con solo mobile attivo ma con le altre utenze in attivazione sono considerati invece attivi. Si tratta infatti di clienti di fatto attivi, ma che al momento dell'analisi hanno la data di attivazione popolata solo per l'utenza mobile.

Mesi cliente: mesi trascorsi dall'attivazione della prima utenza alla disattivazione di tutte le utenze intestate al cliente. Se il cliente è attivo al momento dell'analisi il valore è dato dalla differenza in mesi dalla prima attivazione alla data di svolgimento dell'analisi.

numero di utenze: numero di utenze di cui il cliente è intestatario.

Flag sciolto: registra se il cliente ha attivato almeno una volta un'utenza con contratto Prodotto Sciolto.

Tipo Cliente: ha a che fare con il tipo di utenze attivate dal cliente. Si possono distinguere tre categorie:

- ENG: cliente che ha avuto solo utenze Energia e/o Gas
- TLC: cliente che ha avuto solo utenze Adsl, Voce e/o Mobile
- TLC/ENG: cliente che ha avuto sia utenze Energia e/o Gas che utenze Adsl, Voce e/o Mobile.

Fattura Media: fattura media del cliente in euro.

Area Nielsen: area di residenza del cliente. La distinzione è fatta in quattro aree:

- Area N. 1: Nord Ovest (Piemonte – Valle d'Aosta – Liguria – Lombardia)

- Area N. 2: Nord Est (Veneto – Friuli V. Giulia – Trentino A.A-Emilia R.)
- Area N. 3: Centro (Toscana – Lazio – Marche – Abruzzo– Molise-Sardegna)
- Area N. 4: Sud e Sicilia (Campania – Calabria – Basilicata –Puglia – Sicilia).

Età: età del cliente in anni.

Prima modalità di pagamento: modalità di pagamento del cliente al momento della sottoscrizione del suo primo contratto, le opzioni sono:

- BB: bonifico bancario
- BP: bollettino postale
- SDD: servizio di incasso tramite addebito diretto

ultima modalità di pagamento: modalità di pagamento del cliente al momento dell'analisi.

Canali Digitali: numero medio mensile di accessi del cliente ai canali digitali considerando gli accessi a chat, selfcare, chatbot e telegram.

mobile: registra se il cliente ha attivato almeno una volta un'utenza mobile.

tipo spedizione: modalità di spedizione della fattura. Il cliente può scegliere tra spedizione cartacea o elettronica.

Campagna autolettura: numero medio mensile di campagne per l'autolettura ricevute dal cliente.

Campagna credito: numero medio mensile di campagne che riguardano il credito ricevute dal cliente.

Campagna cross selling: numero medio mensile di campagne per la vendita di ulteriori prodotti e/o servizi dell'azienda ricevute dal cliente.

Campagna promozioni e/o offerte: numero medio mensile di campagne per proporre promozioni e/o offerte dedicate ricevute dal cliente.

Campagna recall: numero medio mensile di campagne per ricontattare il cliente in seguito ad esempio a un tentativo di contatto fallito, una lamentela e così via.

Campagna rimodulazione: numero medio mensile di campagne di rimodulazione della taglia di prezzo e consumi ricevute dal cliente. Ad esempio può accadere ci si renda conto che il cliente consuma sistematicamente di meno o di più di quanto stabilito nella sua taglia di consumo, per cui si voglia evitare che incorra in sovrapprezzi e risulti scontento o presenti reclami.

Campagna spedizione: numero medio mensile di campagne di che hanno a che fare con la spedizione di un *device* ricevute dal cliente.

Campagna tecnico: numero medio mensile di campagne informative legate ad aspetti tecnici ricevute dal cliente.

Informazione attivazione: numero medio mensile di informazioni su questioni legate all'attivazione delle utenze richieste dal cliente.

Informazione canali digitali: numero medio mensile di informazioni su questioni legate ai canali digitali e ai relativi servizi richieste dal cliente.

Informazione conto relax: numero medio mensile di informazioni su questioni legate al conto relax richieste dal cliente. Ha a che fare con il prodotto integrato ed è la modalità di fruizione dei servizi energia e gas: sulla base della taglia di consumo assegnata al cliente, se quest'ultimo consuma meno del previsto potrà accumulare l'eccedenza e usufruirne nei mesi successivi.

Informazione contratto: numero medio mensile di informazioni su questioni legate al contratto stipulato richieste dal cliente.

Informazione credito: numero medio mensile di informazioni su questioni legate alle modalità di fruizione del credito o alla propria situazione creditizia richieste dal cliente.

Informazione disattivazione: numero medio mensile di informazioni su questioni legate alle modalità di disattivazione di una o più utenze richieste dal cliente.

Informazione fattura: numero medio mensile di informazioni su questioni legate alla fattura richieste dal cliente. Ad esempio si può trattare di semplici richieste di informazioni sulle voci in fattura, che però non sfociano, almeno al momento di questa prima richiesta, in un reclamo.

Informazione modulistica: numero medio mensile di informazioni su questioni legate alla modulistica richieste dal cliente. Ad esempio si può trattare di domande sul suo reperimento, il suo utilizzo, e così via.

Informazione pagamento: numero medio mensile di informazioni su questioni legate al pagamento richieste dal cliente. Ad esempio si può trattare di informazioni sulle modalità di pagamento, sullo stato di un pagamento effettuato, e così via.

Informazione promo e offerte: : numero medio mensile di informazioni su questioni legate a promozioni e/o offerte dedicate richieste dal cliente.

Informazione recapito consulente: numero medio mensile di informazioni sul recapito del consulente di area richieste dal cliente. Molti contratti vengono stipulati tramite la rete di consulenti e, ad esempio, può accadere che il cliente non riesca a contattare il consulente della rete con cui ha firmato il contratto, che il consulente sia cambiato, e così via.

Informazione rimodulazione: numero medio mensile di informazioni su questioni legate alla rimodulazione della propria taglia di consumo richieste dal cliente. Ad esempio può accadere che il cliente si renda conto di consumare di meno o di più di quanto stabilito nella sua taglia di consumo e voglia cambiare taglia per evitare di pagare più di quanto consuma o di pagare un consumo a prezzo rincarato per i kilowatt consumati non previsti dal suo contratto.

Informazione sconti e rimborsi: numero medio mensile di informazioni su questioni legate a sconti dedicati o rimborsi da ricevere richieste dal cliente.

Informazione spedizione: numero medio mensile di informazioni su questioni legate alle spedizioni dei *device* (modem, tablet, etc.) richieste dal cliente.

Informazione tecnico: numero medio mensile di informazioni su questioni legate a questioni di carattere tecnico richieste dal cliente. Ad esempio il cliente può chiedere informazioni sulla configurazione della rete, sulla potenza energetica erogabile, e così via

Informazione voltura: numero medio mensile di informazioni su questioni legate alla voltura di un utenza richieste dal cliente.

Invio credito: numero medio mensile di casi aperti per l'invio di comunicazioni legate al credito. Ad esempio si può comunicare l'accettazione del pagamento dilazionato, la necessità di saldare una fattura scoperta, e così via.

Invio contratto: numero medio mensile di casi aperti per l'invio del contratto stipulato.

Invio fattura: numero medio mensile di casi aperti per l'invio della fattura in seguito a richiesta del cliente. Ad esempio può capitare che al cliente non venga recapitata la fattura o che richieda l'invio di una vecchia fattura.

Invio modulistica: numero medio mensile di casi aperti per l'invio di modulistica di vario genere.

Reclamo attivazione: numero medio mensile di casi di reclamo aperti dal cliente legati all'attivazione di una o più utenze. Ad esempio il cliente può riscontrare ritardi o problemi tecnici in attivazione e sporgere un reclamo.

Reclamo credito: numero medio mensile di casi di reclamo aperti dal cliente legati al credito. Ad esempio il cliente può ritenere errata l'attribuzione di una certa condizione creditizia e sporgere un reclamo.

Reclamo fattura: numero medio mensile di casi di reclamo aperti dal cliente legati alla fattura. Ad esempio il cliente può ritenere errate delle voci in fattura o ritenere di pagare un prezzo diverso dal previsto e sporgere un reclamo.

Reclamo guasto: numero medio mensile di casi di reclamo aperti dal cliente legati al guasto di un'utenza.

Reclamo legale: numero medio mensile di casi di reclamo aperti dal cliente legati.

Reclamo spedizione: numero medio mensile di casi di reclamo aperti dal cliente legati alla spedizione di un *device*. Ad esempio il cliente può non ricevere il *device* ordinato o ricevere un oggetto errato e sporgere un reclamo.

Variazione agevolazione: numero medio mensile di casi aperti che registrano una variazione legata a una agevolazione prevista a norma di legge. Ad esempio può capitare che si debba aggiornare il prezzo dei servizi per dei clienti che vivono in una zona in cui è avvenuta una catastrofe naturale.

Variazione anagrafica: numero medio mensile di casi aperti che registrano una variazione legata all'anagrafica del cliente. Ad esempio può riguardare il numero di telefono, la modifica del nome a causa di un errore, e così via.

Variazione autolettura: numero medio mensile di casi aperti che registrano una variazione legata all'autolettura del contatore fornita dal cliente.

Variazione credito: numero medio mensile di casi aperti che registrano una variazione legata alla situazione creditizia del cliente.

Variazione fattura: numero medio mensile di casi aperti che registrano una variazione legata ai valori delle voci in fattura del cliente.

Variazione pagamento: numero medio mensile di casi aperti che registrano una variazione delle modalità di pagamento del cliente.

Variazione rimodulazione: numero medio mensile di casi aperti che registrano una variazione della taglia del cliente. Il momento della variazione si distingue dalla campagna in quanto registra l'atto vero e proprio di variazione dell'offerta.

Variazione sconto e/o rimborso: numero medio mensile di casi aperti che registrano l'applicazione o la variazione di uno sconto o di un rimborso al cliente.

Variazione tecnico: numero medio mensile di casi aperti che registrano una variazione di carattere tecnico sulle utenze del cliente. Ad esempio possono essere variate le caratteristiche del contatore, la potenza erogata, il tipo di connessione ad internet, e così via.

Appendice B: codice R

Di seguito verrà riportato il codice R utilizzato per l'analisi. Si fa presente che verrà riportata una versione semplificata, in cui vengono inclusi solo gli elementi utili all'analisi e solo per il dataset PIB, che è il dataset su cui è stata svolta l'analisi oggetto di questa tesi.

```
###librerie
library(survival)
library(randomForestSRC)
library(ggRandomForests)
library(ggplot2)
library(RJDBC)
library(DBI)
library(varhandle)
library(reshape)
library(plyr)
library(reshape2)
library(dplyr)

# viene richiamata la funzione custom per l'import dei dati
# da SQL server
# (non inserita per intero perché contenente i dati di
# accesso al database)
source("C:/Users/Flavio/Desktop/functions/importDatiFromSQL
.R")

###Data Pre-processing
gc()
options(java.parameters ="-Xmx12192m")
PIB <- importDatiFromSql(sqlText = "select * from PIB where
mesi_dalla_prima_attivazione <= 42")

# si definiscono le variabili categoriali come factor
PIB = PIB %>%
  mutate(TipoCliente = as.factor(TipoCliente)
                ,AreaNielsenCliente =
                 as.factor(AreaNielsenCliente)
```

```
                ,primamodalita = as.factor(primamodalita)
                ,ultimamodalita = as.factor(ultimamodalita)
                ,tiposped = as.factor(tiposped))

# si definisce la variabile status, che sarà 1 se si
# verifica l'evento terminale (churn)
PIB <- PIB %>% mutate(status = ifelse(PIB$clienteattivo==0,
1, 0))

# per il ricavo marginale, viene fatta la media mensile sia
# per il ricavo sui contratti prodotto integrato che
# eventuali prodotto sciolto. Dato che si ha il ricavo solo
# per 13 mesi:
vett = PIB %>%
  mutate(vett = case_when(mesicliente >= 13 ~ 13,
                          mesicliente == 0 ~ 1,
                          TRUE ~ mesicliente)) %>%
  mutate(ricavomarginale_PI_mean =
          ricavomarginale_tot13mesi/vett,
          ricavomarginale_sciolto =
          ricavomarginale_sciolto/mesicliente)

# si sommano i ricavimarginali fatti sulle utenze Prodotto
# Integrato e Prodotto Sciolto
PIB <- PIB %>%
  mutate(ricavomarginale_mean =
      ifelse(is.na(ricavomarginale_PI_mean),0,
      ricavomarginale_PI_mean) +
         ifelse(is.na(ricavomarginale_sciolto_mean),0,
         ricavomarginale_sciolto_mean)) %>%

  # poiché il ricavo per quelli che hanno 0 non é realmente
  # 0 ma é NA, si riassegna il valore NA
  mutate(ricavomarginale_mean = ifelse(ricavomarginale_mean
           == 0, NA, ricavomarginale_mean)) %>%

  # Si costruisce la variabile CanaliDigitali, data dalla
  # somma degli accessi ai canali digitali
  mutate(CanaliDigitali = chat + selfcare + chatbot +
           telegram) %>%
  # Si costruisce la variabile che indica se un clt PI ha o
```

```
  # ha avuto anche un contratto PS
  mutate(flag_sciolto = ifelse(PIB$Sciolto > 0,1,0))

# Si fa la media mesile di casi aperti per ogni tipo di
# caso
names_casi_mean <- c("Campagna_Autolettura"
                     ,"Campagna_Credito"
                     ,"Campagna_CrossSelling"
                     ,"Campagna_Promo_Offerte"
                     ,"Campagna_Recall"
                     ,"Campagna_Rimodulazione"
                     ,"Campagna_Spedizione"
                     ,"Campagna_Tecnico"
                     ,"info_Attivazione"
                     ,"info_Canali_digitali"
                     ,"info_ContoRelax"
                     ,"info_Contratto"
                     ,"info_Credito"
                     ,"info_Disattivazione"
                     ,"info_Fattura"
                     ,"info_Modulistica"
                     ,"info_Pagamento"
                     ,"info_Promo_Offerte"
                     ,"info_Recapito_consulente"
                     ,"info_Rimodulazione"
                     ,"info_Sconto_Rimborso"
                     ,"info_Spedizione"
                     ,"info_Tecnico"
                     ,"info_Voltura"
                     ,"Invio_Credito"
                     ,"invio_Contratto"
                     ,"invio_Fattura"
                     ,"invio_Modulistica"
                     ,"Reclamo_Attivazione"
                     ,"Reclamo_Credito"
                     ,"Reclamo_Fattura"
                     ,"Reclamo_Guasto"
                     ,"Reclamo_legale"
                     ,"Reclamo_Spedizione"
                     ,"Variazione_Agevolazione"
                     ,"Variazione_Anagrafica"
```

```
                      ,"Variazione_Autolettura"
                      ,"Variazione_Credito"
                      ,"Variazione_Fattura"
                      ,"Variazione_Pagamento"
                      ,"Variazione_Rimodulazione"
                      ,"Variazione_Sconto_Rimborso"
                      ,"Variazione_Tecnico")

PIB[,names_casi_mean] <- (PIB[,names_casi_mean] /
PIB[,"mesicliente"])

# Prima di standardizzare viene fatta una copia del dataset
# non standardizzato
pib_unscaled <- PIB

names_da_std <- c(names_casi_mean
                  ,"n_utenze"
                  ,"FattMedia"
                  ,"etareferente"
                  ,"CanaliDigitali"
                  ,"mobile")

PIB[,names_da_std] <- scale(PIB[,names_da_std])

# apply(PIB, 2, function(x) any(is.na(x))) # per capire
# quali colonne contengono NAslibrary(randomForestSRC)

### RANDOM SURVIVAL FOREST
rsf_pib <- rfsrc(formula = Surv(mesicliente,status)~
                       n_utenze
                     +flag_sciolto
                     +TipoCliente
                     +FattMedia
                     +AreaNielsenCliente
                     +etareferente
                     +primamodalita
                     +ultimamodalita
                     +CanaliDigitali
                     +mobile
                     +Campagna_Autolettura
                     +Campagna_Credito
```

+Campagna_CrossSelling

+Campagna_Promo_Offerte
+Campagna_Recall
+Campagna_Rimodulazione
+Campagna_Spedizione
+Campagna_Tecnico
+info_Attivazione
+info_Canali_digitali
+info_ContoRelax
+info_Contratto
+info_Credito
+info_Disattivazione
+info_Fattura
+info_Modulistica
+info_Pagamento
+info_Promo_Offerte
+info_Recapito_consulente
+info_Rimodulazione
+info_Sconto_Rimborso
+info_Spedizione
+info_Tecnico
+info_Voltura
+Invio_Credito
+invio_Contratto
+invio_Fattura
+invio_Modulistica
+Reclamo_Attivazione
+Reclamo_Credito
+Reclamo_Fattura
+Reclamo_Guasto
+Reclamo_legale
+Reclamo_Spedizione
+Variazione_Agevolazione
+Variazione_Anagrafica
+Variazione_Autolettura
+Variazione_Credito
+Variazione_Fattura
+Variazione_Pagamento
+Variazione_Rimodulazione
+Variazione_Sconto_Rimborso

```
                    +Variazione_Tecnico
                    +tiposped

           ,data=PIB, ntree=1000,forest=T, rf.cores=15,
           na.action = "na.impute", seed = 123
           ,importance=TRUE)

### ANALISI ESPLORATIVA VAR Y
# barplot status
png(file =
"C:/Users/Flavio/Desktop/plots/status_barplot.png", height
= 350, width = 450);

PIB %>% mutate(status=factor(ifelse(status==0,"not
                churned", "churned"))) %>%
  ggplot(aes(x=factor(status))) +
  geom_bar(width=0.6) +
  geom_text(aes(y = ((..count..)/sum(..count..)),
                label =
scales::percent((..count..)/sum(..count..))),
                stat = "count", vjust = -15, size=7) +
  xlab("status") +
  ylab("conteggio") +
  theme(axis.title.x = element_text(size = 16),
        axis.text.x = element_text(size = 14),
        axis.text.y = element_text(size = 14),
        axis.title.y = element_text(size = 16))
dev.off()

# density mesicliente
png(file =
"C:/Users/Flavio/Desktop/plots/mesicliente_density.png",
height = 350, width = 450);

PIB %>% ggplot(aes(x=mesicliente)) +
  geom_histogram(aes(y=..density..),colour="black",
                fill="white") +
  geom_density(alpha=.2, size=1,linetype="solid") +
  xlim(0,42) +
  xlab("numero di mesi") +
  ylab("densità") +
```

```
  theme(axis.title.x = element_text(size = 16),
        axis.text.x = element_text(size = 14),
        axis.text.y = element_text(size = 14),
        axis.title.y = element_text(size = 16))
dev.off()

# boxplot
png(file =
"C:/Users/Flavio/Desktop/plots/mesicliente_boxplot.png",
height = 350, width = 450);

PIB %>% ggplot(aes(x="",y=mesicliente)) +
  geom_boxplot(outlier.colour="black", outlier.shape=16,
               outlier.size=2, notch=FALSE) +
  coord_cartesian(ylim = c(0, 42)) +
  ylab("numero di mesi") +
  xlab("") +
  theme(axis.title.y = element_text(size = 16),
        axis.text.y = element_text(size = 14))
dev.off()

summary(PIB$mesicliente)

### GRAFICO RIDUZIONE ERROR RATE
plot(rsf_pib)

### MISURE AGGIUNTIVE ACCURACY
# heatmap sui soli clt churned (threshold 0.5): mesi
# cliente vs mesi stimati
esempio = as.matrix(rsf_pib$survival.oob)

sur1  = array(NA,dim = nrow(esempio))
for(i in 1:nrow(esempio)){

  sa=esempio[i,]
  sur1[i] =
ifelse(is.na(which(sa<=0.5)[1]),42,which(sa<=0.5)[1])

}
```

```
tavola <- table(estim = sur1[which(PIB$clienteattivo ==
                0)],
              mesicliente = PIB[which(PIB$clienteattivo ==
                0),]$mesicliente)

tavola = melt(tavola)

p <- ggplot(tavola, aes(mesicliente, estim)) +
     geom_tile(aes(fill = value),
                colour = "white") + scale_fill_gradient(low
                = "white", high = "blue", na.value =
                "white")

png(file =
"C:/Users/Flavio/Desktop/plots/heatmap_pib_churned0.5.png",
height = 350, width = 450); p; dev.off()

# per i clienti ancora non churned: misura di plausibilità
length(which(sur1[which(PIB$clienteattivo == 1)] <
     PIB[which(PIB$clienteattivo == 1),]$mesicliente)) /
length(sur1[which(PIB$clienteattivo == 1)]) * 100


###  VIMP
modello_ridotto_pib$importance
select_pib <- var.select(modello_ridotto_pib)
select_pib$varselect$vimp

# vimp plot
VimpPlot_pib <- plot(gg_vimp(modello_ridotto_pib))
png(file = "C:/Users/Flavio/Desktop/plots/vimp_pib.png",
height = 500, width = 500);
VimpPlot_pib;
dev.off()

# Rank plot
rankplot_pib<-plot(gg_minimal_vimp(modello_ridotto_pib))
png(file =
"C:/Users/Flavio/Desktop/plots/rankplot_pib.png", height =
500, width = 600);
rankplot_pib;
```

```
dev.off()

vimp_pib <- rankplot_pib[[1]] %>%
           arrange(vimp) %>%
           select(-col)

# grafico sola minimal depth
gg_md2 <- gg_minimal_depth(modello_pic, lbls = st.labs)
plot(gg_md2)

### LIFETIME
# somma delle probabilità di survival
estim_lft <- apply(rsf_pib$survival.oob, 1, sum)
PIB$estim_lft <- estim_lft

### RICAVO MARGINALE
# Density
png(file =
"C:/Users/Flavio/Desktop/plots/ricavomarginale_density.png"
, height = 350, width = 450);

PIB  %>%
  ggplot(aes(x=ricavomarginale_mean)) +
  geom_histogram(aes(y=..density..),colour="black",
                 fill="white") +
  geom_density(size=1,linetype="solid") +
  xlim(-100,1500) +
  xlab("margine medio in euro") +
  ylab("densità") +
  theme(axis.title.x = element_text(size = 16),
        axis.text.x = element_text(size = 14),
        axis.text.y = element_text(size = 14),
        axis.title.y = element_text(size = 16))
dev.off()

# boxplot
png(file =
"C:/Users/Flavio/Desktop/plots/ricavomarginale_boxplot.png"
, height = 350, width = 450);

PIB %>% ggplot(aes(x="",y=ricavomarginale_mean)) +
```

```
  geom_boxplot(outlier.colour="black", outlier.shape=16,
               outlier.size=2, notch=FALSE) +
  coord_cartesian(ylim = c(-200, 1500)) +
  ylab("margine medio in euro") +
  xlab("") +
  theme(axis.title.y = element_text(size = 16),
        axis.text.y = element_text(size = 14))
dev.off()

summary(PIB$ricavomarginale_mean)

### LIFETIME VALUE
# calcolo
PIB <- PIB %>% mutate(lft_value = estim_lft *
                      ricavomarginale_mean)
summary(PIB$lft_value)

# Dataset con solo i clienti ancora attivi
attivi_pib <- PIB %>% filter(clienteattivo == 1)
summary(attivi_pib$lft_value)

# density
mediana <- median(attivi_pib$lft_value, na.rm = T)
media <- mean(attivi_pib$lft_value, na.rm = T)

png(file
="C:/Users/Flavio/Desktop/plots/lft_value_pib_density.png",
height = 350, width = 400 )

attivi_pib %>% ggplot(aes(x=lft_value)) +
  geom_density() + coord_cartesian(xlim = c(-1500,30000)) +
  geom_vline(xintercept= c(mediana, media),
             linetype=c("dashed","solid"), size=2) +
  theme(plot.title = element_text(lineheight=.8,
face="bold")) +
  geom_text(size = 6, aes(x = 21000,y=1e-04,label="- - -
Mediana = 4942€")) +
  geom_text(size = 6, aes(x = 21000,y=1.2e-04,label="-
Media = 9381€")) +
  theme(legend.position="none" )  +
  xlab("Lifetime Value") +
```

```
  ylab("densità") +
  theme(axis.title.x = element_text(size = 16),
        axis.text.x = element_text(size = 14),
        axis.text.y = element_text(size = 14),
        axis.title.y = element_text(size = 16))
dev.off()

# boxplot
png(file =
"C:/Users/Flavio/Desktop/plots/lft_value_boxplot.png",
height = 350, width = 450);

attivi_pib %>% ggplot(aes(x="",y=lft_value)) +
  geom_boxplot(outlier.colour="black", outlier.shape=16,
outlier.size=2, notch=FALSE) +
  coord_cartesian(ylim = c(-2000,3e+04)) +
  ylab("Lifetime Value in euro") +
  xlab("") +
  theme(axis.title.y = element_text(size = 16),
        axis.text.y = element_text(size = 14))
dev.off()


### HEATMAP LIFETIME VS VARIABILI INDIPENDENTI
# Probabilità di sopravvivenza per cliente
d <- apply(rsf_pib$survival.oob,1,sum,na.rm=T)
bs <- cbind.data.frame(rsf_pib$xvar,d)
nomi <- names(bs)
ne <- nomi[-which(nomi %in%
c("TipoCliente","AreaNielsenCliente","primamodalita","tipos
ped","ultimamodalita"))]

# Codifica disgiuntiva completa dei factor
TipoCliente <- to.dummy(bs$TipoCliente, "TipoCliente")
AreaNielsenCliente <- to.dummy(bs$AreaNielsenCliente,
"AreaNielsenCliente")
primamodalita <- to.dummy(bs$primamodalita,
"primamodalita")
ultimamodalita <- to.dummy(bs$ultimamodalita,
"ultimamodalita")
tiposped <- to.dummy(bs$tiposped, "tiposped")
```

```
b3 <-
cbind.data.frame(bs[,ne],TipoCliente,AreaNielsenCliente,pri
mamodalita,tiposped,ultimamodalita)

# Si ordinano le variabili per VIMP Rank
b3 <- b3[, c("cluster","Variazione_Credito"
,"Campagna_CrossSelling","Campagna_Rimodulazione"
,colnames(TipoCliente),"info_Disattivazione"
,"Campagna_Credito","info_Fattura"
,"Variazione_Anagrafica","invio_Modulistica"
,"CanaliDigitali","info_Credito"
,"n_utenze",colnames(ultimamodalita)
,"Invio_Credito","mobile","invio_Fattura"
,"Reclamo_Guasto","Variazione_Sconto_Rimborso"
,"Campagna_Tecnico","info_Modulistica"
,"info_Contratto","info_Attivazione"
,"Campagna_Promo_Offerte","info_Pagamento"
,"Reclamo_Fattura",info_Voltura"
,"Variazione_Pagamento",colnames(primamodalita)
,"info_Canali_digitali","info_ContoRelax"
,Campagna_Autolettura","Variazione_Autolettura"
,"Variazione_Tecnico","Variazione_Rimodulazione"
,"FattMedia","invio_Contratto"
,"flag_sciolto","Reclamo_legale"
,"info_Spedizione","Reclamo_Credito"
,"info_Tecnico", “Campagna_Recall"
,colnames(tiposped),"Variazione_Fattura"
,"info_Sconto_Rimborso","Reclamo_Attivazione"
,"info_Recapito_consulente","info_Rimodulazione"
,"Variazione_Agevolazione","etareferente"
,colnames(AreaNielsenCliente),"info_Promo_Offerte"
,"Campagna_Spedizione","Reclamo_Spedizione")]

b4 <- melt(b3,id.vars="d")
b4$d  <- round(b4$d,0)

# Costruzione classi di mesi
b4$d <- cut(b4$d,9,labels = c("01.0-5","02.5-9","03.9-
          14","04.14-18","05.18-22","06.23-27","07.27-
          31","08.31-36","09.36-42"))
```

```
b5 <- b4 %>%
  group_by(d,variable) %>%
  dplyr::summarise(value=mean(value,na.rm=T))

b5 <- ddply(b5, .(variable), transform, rescale =
           scale(value))
base_size <- 9

b5$Mesi <- as.factor(b5$d)

b5 <- ddply(b5, .(variable), transform, rescale =
           scale(value))

p <- ggplot(b5, aes(Mesi,variable)) + geom_tile(aes(fill =
     rescale), colour = "white") +
     scale_fill_gradient2(low="blue",mid="white",high="red"
     ,midpoint =0)
q <- p + theme_grey(base_size = base_size) +
      labs(y = "") + scale_x_discrete(expand = c(0, 0)) +
     theme(axis.text.x = element_text(size = base_size ,
           angle =    330, hjust = 0, colour = "grey50"))

png(file =
"C:/Users/Flavio/Desktop/plots/heatmap_classimesivsVar_pib.
png", height = 570, width = 550, pointsize = 18, bg =
"transparent", res = 95); q; dev.off()

### SEGMENTAZIONE DEL PARCO E ANALISI DEL LIFETIME VALUE
PIB %>%
     mutate(cluster = case_when(lft_value <= 5000 ~ "iron",
        lft_value > 5000 & lft_value <= 10000 ~ "bronze",
        lft_value > 10000 & lft_value <= 20000  ~ "silver",
        lft_value > 20000 & lft_value <= 30000  ~ "gold",
        lft_value > 30000 ~ "platinum",
         TRUE ~ NA))

# Clienti oggetto di interesse (ancora attivi),
# sovrascritto per aggiunta della variabile cluster
attivi_pib <- PIB %>% filter(clienteattivo == 1)
```

```
# Percentuale clt in ogni cluster
table(attivi_pib$cluster)/nrow(attivi_pib)

# Percentuale lft_value per gruppo sul totale
sum(attivi_pib$lft_value[which(attivi_pib$cluster ==
     "iron")])/ sum(attivi_pib$lft_value)
sum(attivi_pib$lft_value[which(attivi_pib$cluster ==
     "bronze")])/ sum(attivi_pib$lft_value)
sum(attivi_pib$lft_value[which(attivi_pib$cluster ==
     "silver")])/ sum(attivi_pib$lft_value)
sum(attivi_pib$lft_value[which(attivi_pib$cluster ==
     "gold")])/ sum(attivi_pib$lft_value)
sum(attivi_pib$lft_value[which(attivi_pib$cluster ==
     "platinum")])/ sum(attivi_pib$lft_value)


# Heatmap cluster vs variabili indipendenti

# riutilizzo dei dati non standardizzati, poiché si sta
# usando solo il sottoinsieme attivi mentre la
# standardizzazione era stata fatta su sutto il dataset,
# per cui è necessario standardizzare solo per il
# sottoinsieme di clienti attivi
a <- pib_unscaled %>% select(cliente, names_da_std)
b <- attivi_pib3 %>% select(-names_da_std)
attivi_pib_unscaled <- b %>% inner_join(a, by="cliente")

iii <- names(modello_pib$xvar)
bs <- attivi_pib_unscaled

ne <- iii[-which(names(modello_pib$xvar) %in%
     c("TipoCliente","AreaNielsenCliente","primamodalita","
     tiposped","ultimamodalita"))]

ne <- c(ne,"cluster")
TipoCliente <- to.dummy(bs$TipoCliente, "TipoCliente")
AreaNielsenCliente <- to.dummy(bs$AreaNielsenCliente,
"AreaNielsenCliente")
primamodalita <- to.dummy(bs$primamodalita,
"primamodalita")
```

```
ultimamodalita <- to.dummy(bs$ultimamodalita,
"ultimamodalita")
tiposped <- to.dummy(bs$tiposped, "tiposped")


b3 <- cbind.data.frame(bs[,ne],
      TipoCliente,AreaNielsenCliente,primamodalita,tiposped,
      ultimamodalita)

# Si ordinano le variabili per VIMP Rank
b3 <- b3[, c("cluster","Variazione_Credito"
,"Campagna_CrossSelling","Campagna_Rimodulazione"
,colnames(TipoCliente),"info_Disattivazione"
,"Campagna_Credito","info_Fattura"
,"Variazione_Anagrafica","invio_Modulistica"
,"CanaliDigitali","info_Credito"
,"n_utenze",colnames(ultimamodalita)
,"Invio_Credito","mobile","invio_Fattura"
,"Reclamo_Guasto","Variazione_Sconto_Rimborso"
,"Campagna_Tecnico","info_Modulistica"
,"info_Contratto","info_Attivazione"
,"Campagna_Promo_Offerte","info_Pagamento"
,"Reclamo_Fattura",info_Voltura"
,"Variazione_Pagamento",colnames(primamodalita)
,"info_Canali_digitali","info_ContoRelax"
,Campagna_Autolettura","Variazione_Autolettura"
,"Variazione_Tecnico","Variazione_Rimodulazione"
,"FattMedia","invio_Contratto"
,"flag_sciolto","Reclamo_legale"
,"info_Spedizione","Reclamo_Credito"
,"info_Tecnico", “Campagna_Recall"
,colnames(tiposped),"Variazione_Fattura"
,"info_Sconto_Rimborso","Reclamo_Attivazione"
,"info_Recapito_consulente","info_Rimodulazione"
,"Variazione_Agevolazione","etareferente"
,colnames(AreaNielsenCliente),"info_Promo_Offerte"
,"Campagna_Spedizione","Reclamo_Spedizione")]

b4 <- melt(b3,id.vars="cluster")

b5 <- b4 %>%
```

```
  group_by(cluster,variable) %>%
  dplyr::summarise(value=mean(value,na.rm=T))

b5 <- ddply(b5, .(variable), transform, rescale =
     scale(value))
base_size <- 9


b5$cluster <- factor(b5$cluster,levels
           =c("iron","bronze","silver","gold","platinum"))
p <- ggplot(b5, aes(cluster,variable)) +
     geom_tile(aes(fill = rescale), colour = "white") +
     scale_fill_gradient2(low="blue",mid="white",high="red"
                          ,midpoint =0)
q <- p + theme_grey(base_size = base_size) +
     labs(y = "") +
     scale_x_discrete(expand = c(0, 0)) +
     theme(axis.text.x = element_text(size = base_size ,
          angle = 330, hjust = 0, colour = "grey50"))

png(file =
"C:/Users/Flavio/Desktop/plots/heatmap_clustVsVar_pib.png",
height = 570, width = 550, pointsize = 18, bg =
"transparent", res = 95 ); q; dev.off()
```