Statistica in azienda
Statistici in azienda
SIS 2010
Riunione Satellite

Churn Predictive Analytics

Debora Dominissini e Luca Massaron
AGSM Energia s.p.a. Sede legale Lung. Galtarossa 8 – 37133 Verona www.agsmperte.it – e-mail serv.comm@agsm.it
MKTG Operativo

Convegno degli statistici italiani che lavorano nelle aziende, 15 Giugno 2010

agsm Energia

Introduzione

Dal 1 luglio 2007 in Italia, come nel resto d'Europa, in attuazione di specifiche direttive EU, vige la completa liberalizzazione della domanda di energia elettrica e di gas.

La liberalizzazione ha imposto ai player del settore una maggiore centralità delle esigenze e dei bisogni del cliente finale, al fine di prevenirne la migrazione verso altri competitor (processi di churn management, azioni mirate a livello commerciale e marketing).

Infatti, in quei paesi europei in cui già da alcuni anni è stato completato il libero accesso al mercato dell'energia, il tasso di churn stimato per gli operatori del settore è di circa il 20-25% della propria base clienti.

Poiché riconquistare un cliente costa molto più che mantenerlo, l'anticipazione del churn della propria clientela permette azioni mirate volte alla retention con conseguente riduzione delle perdite di fatturato e redditività. Tali azioni sono possibili disponendo di efficaci modelli predittivi in grado di quantificare la probabilità che i propri clienti abbandonino l'organizzazione per passare a fruire dei servizi offerti dai concorrenti.

Soluzione metodologica e tecnologica

Attraverso l'analisi delle caratteristiche e del comportamento dei clienti, è possibile quantificare la probabilità di ciascun cliente di divenire un churner, rendendo possibile alle funzioni di marketing e customer care di:

- progettare azioni di fidelizzazione mirate (campagne promozionali, azioni pubblicitarie);
- supportare il processo di definizione di nuovi prodotti/servizi rivolti alla miglior retention dei clienti.

AGSM Energia, fortemente radicata sul territorio locale e rivolta da sempre alle esigenze della propria clientela, ha pianificato e realizzato una apposita infrastruttura informativa (database di marketing - DBM) per seguire, servire e fidelizzare al meglio i propri clienti.

Il DBM aziendale raccoglie, organizza e conserva tutte le informazioni disponibili sui clienti come dati demografici, servizi di fornitura, modalità e canali di contatto utilizzati dal cliente verso l'azienda (call center, customer care e ufficio relazioni con il pubblico), indici di valore fatturato/consumo del cliente, modalità di pagamento, variabili di segmentazione del cliente per caratteristiche demografiche e di utenza.

Partendo da questa base dati opportunamente strutturata, il marketing operativo di AGSM Energia ha curato l'addestramento di appositi modelli predittivi per il churn management, grazie al software di Data Mining Clementine di IBM SPSS ed al linguaggio statistico open-source R.

Fasi di realizzazione dell'analisi predittiva

Le fasi di realizzazione dell'addestramento dei modelli predittivi sono state:

- Definizione e calcolo delle variabili target
- Data Preparation delle variabili predittive
  EDA (exploratory data analysis), verifica incidenza dei livelli delle variabili qualitative nominali/ordinali con accorpamento dei livelli non significativi, verifica delle medie e delle distribuzioni delle variabili quantitative con linearizzazione delle più correlate con le variabili target (le funzioni di linearizzazione sono state individuate attraverso una serie di smoothed scatterplots sviluppate in linguaggio R)
- Individuazione dei casi anomali
  calcolo di un anomaly index previa cluster analysis k-means in grado di rappresentare le distribuzioni presenti nei dati
- Ranking dei predittori
  ranking da stima della relazione bivariata target/predittore con calcolo dell'indice CRAMER V e scoring di importanza prodotto da un modello Random Forest sviluppato grazie ad una libreria del linguaggio R.
- Cluster analysis dei dati
  segmentazione con algoritmi two-steps e self-organizing maps di Kohonen.
- Bilanciamento e partizione dei dati
  il 25% dei dati per la validazione dei modelli
- Costruzione di un modello baseline
  modelli statistici di regressione logistica ed analisi discriminante
- Costruzione dei modelli di data mining
  CHAID, C&RT, QUEST, C5.0 con boosting, reti neurali Multiplayer Perceptron
- Valutazione dei modelli attraverso Curve ROC
- Simulazioni di costo/opportunità di possibili azioni di marketing retention

Algoritmo C5.0 con boosting

Fra gli algoritmi di data mining utilizzati per la previsione dei churner, uno dei più efficaci si è rivelato essere il C5.0. Si tratta di un algoritmo usato per generare alberi di segmentazione (decision tree), sviluppato dal ricercatore e specialista in intelligenza artificiale John Ross Quinlan (University of Technology of Sidney, RAND Corporation).

L'algoritmo C5.0 è la versione commerciale dell'algoritmo C4.5, il quale a sua volta è la versione migliorata dell'algoritmo ID3. Tutti questi algoritmi sono stati ideati da Quinlan a partire dagli anni '80 e si basano tutti sul principio del rasoio di Occam: le teorie e quindi gli alberi di segmentazione più semplici prevalgono su quelli più complessi.

La formalizzazione del principi di Occam in algoritmo matematico è operata attraverso il concetto di entropia informativa, massimizzata attraverso una opportuna segmentazione:

$$E(S) = -\sum_{j=1}^{n} f_S(j) \log_2 f_S(j)$$

E(S) è l'entropia informativa del sotto-campione **S** ;

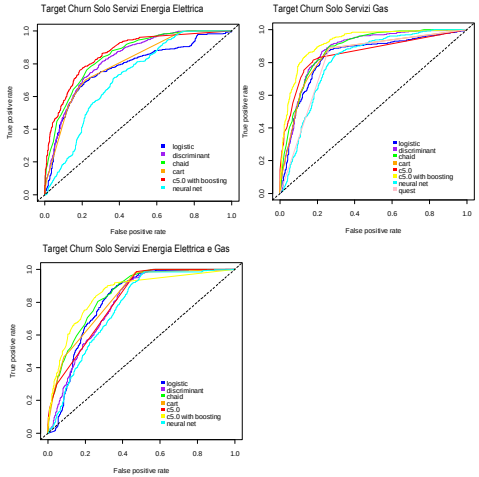
n è il numero di differenti livelli manifestati dalla variabile in oggetto nel sotto-campione S (L'entropia è quindi calcolata per singole variabili)

fs(j) è la frequenza (proporzione) del livello j della variabile entro il sotto-campione S

L'algoritmo C5.0 è potenziato dal boosting, procedura nella quale, durante l'addestramento, vengono creati più alberi di segmentazione allo scopo di classificare via via i casi che gli alberi precedenti non sono riusciti a prevedere correttamente. Il risultato è un sistema (ensemble) di previsioni che vengono sintetizzate in un'unica previsione finale attraverso una somma pesata.

Questo procedimento permette di ottenere la maggior precisione possibile nella previsione, senza rischiare overfitting e quindi scarsa generabilità dei risultati di previsione ottenuti.

Curve ROC



L'analisi di AGSM Energia, richiede di prestare particolare attenzione ai "costi di previsione", ovvero al costo derivante dal prevedere come churner un cliente che potrebbe invece non esserlo.

La letteratura scientifica, per questo tipo di esigenze, suggerisce il ricorso alle curve ROC (Receiver Operating Characteristic). Le curve ROC sono costruite raffrontando, sulla base della probabilità prevista decrescente (nel nostro caso che un cliente diventi un churner), la quota totale di veri positivi contro la quota totale di falsi positivi ottenuti con il proprio modello.

Le curve ROC inizialmente più ripide indicano i modelli più interessanti in quanto sono in grado di individuare subito molti churner senza errori, limitando quindi i costi di un'azione di marketing (minori sprechi nelle attività rivolte a clienti che non diventeranno churner in ogni caso).

Conclusioni

AGSM Energia ha infine scelto di adottare i tre modelli derivanti dall'applicazione dell'algoritmo C5 Boosting.

La scelta è stata effettuata sulla base del criterio AUC (l'area delimitata dalla curva ROC del modello), indice del miglior risultato sia a livello di TP rate (tasso percentuale dei veri positivi, indice della capacità del modello di invidiuare i churner) e sia del FP rate (tasso percentuale dei falsi positivi, indice della capacità del modello di non generare falsi allarmi e quindi di non attivare azioni di marketing non necessarie).

Target Churn Solo Servizio Energia Elettrica

	AUC	TP.rate	FP.rate
logistic	0,774	0,614	0,275
discriminant	0,828	0,756	0,260
chaid	0,847	0,215	0,017
cart	0,798	0,090	0,012
c5			
c5 with boosting	**0,862**	**0,206**	**0,010**
best neural net	0,711	0,881	0,397
quest			

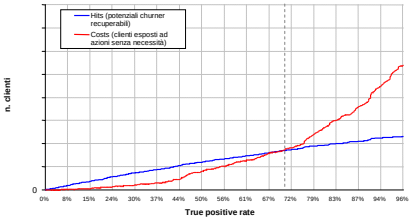
Target Churn Solo Servizio Gas

	AUC	TP.rate	FP.rate
logistic	0,835	0,821	0,286
discriminant	0,875	0,879	0,242
chaid	0,881	0,290	0,032
cart	0,850	0,073	0,001
c5	0,854	0,430	0,043
c5 with boosting	**0,922**	**0,565**	**0,048**
best neural net	0,807	0,865	0,234
quest	0,812	0,887	0,263

Target Churn Solo Servizio Energia Elettrica e Gas

	AUC	TP.rate	FP.rate
logistic	0,809	0,846	0,426
discriminant	0,793	0,921	0,427
chaid	0,848	0,250	0,041
cart	0,828	0,256	0,032
c5	0,809	0,150	0,008
c5 with boosting	**0,853**	**0,147**	**0,009**
best neural net	0,766	0,985	0,516
quest			

Applicazione operativa del modello



A completamento del lavoro analitico di individuazione e costruzione dei migliori modelli di data mining per la previsione del churn nella clientela di AGSM Energia, è stato simulato, attraverso fogli di calcolo opportunamente programmati, il risultato economico di eventuali azioni mirate di marketing. Basandosi sui risultati dei modelli applicati al validation set, si è potuto prefigurare l'ipotetico risultato ottenibile in termini di guadagni economici (individuazione di veri churner) rispetto agli eventuali costi sia fissi (costi della campagna) che variabili (inclusione nell'azione di marketing di falsi churner). L'individuazione da simulazione del punto di equilibrio fra guadagni e costi, permetterebbe l'ottimizzazione della campagna usando solo lo scoring del modello come leva di controllo.