Churn Analysis: creazione di un modello di machine learning profittevole in ambito marketing

Ermellino Andrea, Kolyszko Matteo, Serino Antonio, Valente Sofia, Maggi Lucrezia

Dipartimento di Informatica Sistemistica e Comunicazione, DISCO , Università degli Studi di Milano Bicocca

Piazza dell'Ateneo Nuovo, 1, 20126 Milano MI

Abstract: Il customer churn rate riveste un ruolo fondamentale nel monitoraggio dell'andamento del business in termini di percentuale di consumatori che durante un determinato periodo di tempo ha "abbandonato" l'azienda, ponendo fine al rapporto. L'obiettivo di questa analisi è addestrare un modello in grado di predire con precisione il tasso di propensione all'abbandono tanto da poter essere reputato profittevole in ambito marketing. Per far ciò, sono stati addestrati modelli di Neural Network e Random Forest adottando diverse tecniche di campionamento al fine di risolvere il problema della classe sbilanciata, sono state confrontate le performance dei due modelli per individuarne il migliore che è stato anche prodotto in una sua forma più parsimoniosa grazie ad una fase preliminare di feature selection. È stata infine calcolata e confrontata la Propensity to churn con la percentuale media di churners reali, al fine di poter analizzare ancor meglio se le predizioni del miglior modello individuato potessero avere un effettivo riscontro nei dati reali.

Key words: Machine Learning, Churn Analysis, SMOTE, Random Forest, Neural Network.

Indice

1 Preprocessing

1.1 Descrizione del dataset

La versione iniziale del dataset è composta da 102 colonne e 330586 righe contenente informazioni sulle abitudini di navigazione di un gruppo di clienti di una piattaforma online.
In particolare è possibile distinguere macro-categorie di colonne ognuna delle quali descrive un diverso fattore:

- colonne con dati numerici che descrivono quali URL visitati dall'utente funzionavano e quali no;
- colonne con dati binari che indicano il sistema operativo e il browser utilizzato dagli utenti;
- colonne con dati numerici normalizzati nel range [0,1] che indicano il tempo di utilizzo della piattaforma in una determinata fascia oraria e in un determinato periodo della settimana;
- colonne con dati numerici normalizzati nel range [0,1] rappresentanti la lunghezza dei testi letti dai clienti;

- colonne con dati numerici normalizzati nel range [0,1] rappresentanti le categorie di interesse dei clienti.
- colonne con dati binari che indicano i pacchetti acquistati dall'utente e i servizi offerti preferiti.

Di nostro particolare interesse è la colonna *Pdisc*, questa infatti rappresenta la variabile binaria di target che indica se un cliente ha abbandonato l'azienda o meno.

1.2 Analisi esplorativa

L'analisi esplorativa parte dal conteggio di quanti churner sono presenti all'interno del dataset: la classe è particolarmente sbilanciata con il 97% di utenti no churner e il restante 3% churner. Successivamente si è andato a calcolare per i diversi tipi di contratti stipulati e per i diversi sistemi operativi utilizzati dagli utenti, il numero di churners: in particolare è stato possibile notare che la maggior parte dei churner sono quelli che non ususfruiscono del contratto STB_MYSKYHD e che usufruiscono del contratto STB_HD, mentre per i sistemi operativi la maggior parte dei churner sono i clienti che utilizzano Linux. E' stata effettuata inoltre l'analisi della correlazione presente tra la variabile di churn *Pdisc*, le categorie di interesse degli utenti e i pacchetti acquistati: in questa si è verificato che non sono presenti valori di correlazione significativi.

1.3 Preprocessing

Il preprocessing dei dati inizia con l'eliminazione di 8 righe duplicate e di 6 righe nelle quali per le colonne CINEMA, CALCIO, SPORT e SKY_FAMIGLIA, contenenti valori booleani, erano stati ritrovati valori pari a 2. Sono inoltre state eliminate tutte le colonne "admants" in quanto contenenti solo valori pari a 0. Prima di effettuare l'addestramento dei modelli il dataset è stato ordinato sulla base della colonna DATA_RIF che successivamente è stata eliminata insieme alla colonna EXTERNAL_ID in quanto non fornivano informazioni utili per l'analisi.

2 Addestramento Modelli

In questa sezione del documento viene descritto il processo di addestramento dei modelli di Machine Learning selezionati: Random Forest e Artificial Neural Network. La selezione dei due modelli è stata guidata dal fatto che entrambi sono in grado di lavorare con un elevato numero di variabili riuscendo ad ottenere tempi di addestramento computazionalmente sostenibili.
L'obiettivo è addestrare un modello in grado di individuare con precisione il "livello di propensità" di un utente ad abbandonare il servizio analizzando il suo comportamento all'interno della piattaforma online in questione.
Per evitare problemi legati all'overfitting, tutti i modelli sono stati validati mediante K-Fold Cross Validation con 10 split.

2.1 Problema della classe sbilanciata

Prima di procedere con l'addestramento dei modelli è necessario effettuare delle considerazioni sulla distribuzione dei valori della classe da predire (*Pdisc* ha valore 0 se l'utente non ha abbandonato il servizio, 1 se invece ha abbandonato). Come emerso dall'analisi esplorativa, la classe è estremamente sbilanciata: gli utenti che hanno abbandonato il servizio sono appena il 3% delle osservazioni totali, dunque gli utenti che invece non hanno abbandonato il servizio costituiscono il restante 97%.
Siccome l'interesse è per la classe rara ($Pdisc = 1$), nella divisione del dataset in train e test set è impossibile adottare una tecnica di campionamento casuale poiché il set di addestramento potrebbe non contenere osservazioni della classe rara , così come è inutile effettuare un campionamento stratificato perché pur mantenendo le proporzioni originali delle due classi, il modello addestrato su tale set performerebbe come una *ZeroRule*: andando a predire sempre il valore 0 si otterrebbe il 97% di Accuracy ma il modello risulterebbe essere completamente inutile per l'obiettivo (predire la classe rara).
Per risolvere tale problema sono state testate diverse tecniche di campionamento alternative che permettono di bilanciare le due classi nel train e nel test set:

- *Random Undersampling*: tecnica di campionamento che permette di bilanciare le due classi andando ad estrarre casualmente un numero di osservazioni della classe maggioritaria pari al numero di osservazioni della classe minoritaria. In questo modo sia nel train che nel test set c'è un rapporto del 50:50 delle due classi, tuttavia viene utilizzato solamente il 6% delle osservazioni totali del dataset, perdendo informazione;

- *Random Oversampling*: tecnica di campionamento che permette di bilanciare le due classi andando a selezionare casualmente delle osservazioni della classe minoritaria aggiungendole al train e al test set fino ad ottenere un rapporto del 50:50 fra le due classi;

- *Synthetic Minority Over-sampling Technique (SMOTE)*: tecnica di oversampling che permette di bilanciare le due classi andando a creare delle osservazioni sintetiche della classe minoritaria sfruttando i *K nearest neighbors* nello spazio delle variabili delle osservazioni appartenenti alla classe minoritaria.

Entrambi i modelli sono stati addestrati adottando ognuna delle tre tecniche di campionamento descritte per poi essere confrontati.

2.2 Valutazione delle performance

Per valutare la bontà dei modelli sono stati presi in considerazione i principali indicatori di performance: Accuracy, Precision, Recall, F-measure, AUC. Gli indicatori vengono calcolati a partire dal numero di osservazioni correttamente classificate dal modello, e dal numero di osservazioni che invece vengono classificate in modo errato:

- *TP e TN*: numero di istanze appartenenti alla classe positiva (TP) o alla classe negativa (TN) che vengono correttamente classificate;

- *FP e FN*: numero di istanze appartenenti alla classe positiva (FN) o negativa (FP) che vengono classificate in modo errato.

L'Accuracy di un modello di classificazione indica il numero di predizioni corrette effettuate dal modello e assume valore nell'intervallo [0,1]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Nonostante sia un indicatore estremamente diffuso, poiché in questo caso l'obiettivo dell'analisi è predire quando le osservazioni appartengono ad una classe specifica (valore di *Pdisc* = 1), l'Accuracy non è un indicatore che da solo può indicare efficacemente la bontà di un modello. Decisamente più adatti al caso specifico in esame, sono gli indicatori di Precision e Recall di un modello. La Precision indica la frazione di record effettivamente positivi tra tutti quelli predetti come tali:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

La Recall invece, definita anche come True Positive Rate, indica la porzione di record positivi classificati correttamente dal modello:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Entrambi gli indicatori assumono valori nell'intervallo [0,1].
Tuttavia, spesso ci si trova davanti ad un tradeoff tra i due indicatori, un miglioramento della Recall porta ad un peggioramento della Precision e viceversa. Per questo motivo si ricorre all'utilizzo di misure alternative che riassumono i valori di Precision e Recall, come la F-Measure.
La F-Measure è la media armonica tra Recall e Precision e anche questa, come gli indicatori descritti in precedenza, assume valori nell'intervallo [0,1]:

$$F - Measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{4}$$

L'ultimo indicatore numerico di performance preso in considerazione fa riferimento alla curva ROC: una curva che mette in relazione il rate di falsi positivi con il rate dei veri positivi. Tracciata la curva, l'area sottesa al grafico (AUC) risulta essere un indicatore di performance di particolare rilevanza.

2.3 Addestramento con Random Undersampling

Adottando questa tecnica di campionamento vengono utilizzate 19892 osservazioni delle 330mila iniziali ottenendo un rapporto del 50:50 tra le due classi. L'undersampling ha permesso di bilanciare il set di osservazioni, ma allo stesso tempo comporta una notevole perdita di informazione.
Di seguito viene riportato il confronto delle performance dei due modelli:

Classificatore	Accuracy	Precision	Recall	F-Measure	AUC
Neural Network	0.71	0.73	0.67	0.70	0.79
Random Forest	0.69	0.72	0.64	0.68	0.75

Tabella 1: Confronto Random Undersampling

La Neural Network performa leggermente meglio della Random Forest essendo sia più precisa che più

sensibile nella predizione della classe rara. Anche gli indicatori di F-Measure e AUC sono superiori per la Neural Network che dunque risulta essere il modello migliore se si adotta il random undersampling come tecnica di campionamento.
Tuttavia, la perdita di informazione dovuta all'utilizzo di una minima parte (6%) delle osservazioni totali presenti nel dataset influisce notevolmente sulle performance dei modelli, che risultano essere discrete, ma sono facilmente migliorabili adottando tecniche che permettono di sfruttare al massimo le informazioni dell'intero set di dati.

2.4 Addestramento con Random Oversampling

Adottando questa tecnica di campionamento vengono utilizzate tutte le osservazioni presenti nel set di dati, in più per bilanciare le due classi vengono selezionate casualmente delle osservazioni della classe rara e vengono aggiunte al set. In totale vengono utilizzate 641252 osservazioni divise in train e test set. In questo modo non viene persa informazione, ma anzi viene amplificata la presenza di osservazioni appartenenti alla classe rara. Ciò potrebbe portare all'overfitting dei modelli, per questo motivo vengono validati mediante K-Fold Cross Validation.
Di seguito viene riportato il confronto delle performance dei due modelli:

Classificatore	Accuracy	Precision	Recall	F-Measure	AUC
Neural Network	0.71	0.74	0.66	0.69	0.79
Random Forest	0.95	0.96	0.94	0.95	0.99

Tabella 2: Confronto Random Oversampling

Mentre la Neural Network performa in maniera molto simile al modello addestrato con random undersampling, la Random Forest al contrario registra delle performance ottime con 0.96 di Precision e 0.94 di Recall, oltre ad un AUC prossima al valore massimo.
La Random Forest con random oversampling risulta essere fra i migliori modelli addestrati in questa analisi.

2.5 Addestramento con SMOTE

Adottando questa tecnica di campionamento viene sfruttato lo stesso numero di osservazioni del random oversampling (641252), tuttavia le osservazioni appartenenti alla classe rara che vengono aggiunte al set non sono dei duplicati selezionati casualmente ma sono delle osservazioni sintetiche generate sfruttando i K nearest neighbors nello spazio delle variabili delle osservazioni "reali".
Tale tecnica permette di bilanciare le due classi, tuttavia è doveroso chiarire che i modelli vengono addestrati su un set di dati che contiene informazioni sintetiche oltre alle informazioni reali.
Di seguito viene riportato il confronto delle performance dei due modelli:

Classificatore	Accuracy	Precision	Recall	F-Measure	AUC
Neural Network	0.71	0.73	0.68	0.70	0.79
Random Forest	0.95	0.96	0.94	0.95	0.99

Tabella 3: Confronto SMOTE

La Neural Network con tecnica di campionamento SMOTE ottiene le perfomance migliori se confrontate con i modelli addestrati adottando le precedentemente descritte tecniche di campionamento. Si tratta comunque di un miglioramento minimo, poco significativo, se si valutano e confrontano le metriche di performance considerate.
Per quanto riguarda invece la Random Forest, anche con SMOTE le performance risultano essere ottime e praticamente identiche a quelle ottenute adottando la tecnica di campionamento random oversampling.

In conclusione è utile guardare l'evoluzione delle performance dei due modelli adottando le varie tecniche di campionamento:

Classificatore	Campionamento	Accuracy	Precision	Recall	F-Measure	AUC
Neural Network	Random Undersampling	0.71	0.73	0.67	0.70	0.79
Neural Network	Random Oversampling	0.71	0.74	0.66	0.69	0.79
Neural Network	SMOTE	0.71	0.73	0.68	0.70	0.79

Tabella 4: Confronto performance Neural Network

Come si può facilmente notare, le tre tecniche portano a dei risultati estremamente simili nella valutazione della bontà del modello.

Classificatore	Campionamento	Accuracy	Precision	Recall	F-Measure	AUC
Random Forest	Random Undersampling	0.69	0.72	0.64	0.68	0.75
Random Forest	Random Oversampling	0.95	0.96	0.94	0.95	0.99
Random Forest	SMOTE	0.95	0.96	0.94	0.95	0.99

Tabella 5: Confronto performance Random Forest

Nel caso della Random Forest è evidente il miglioramento delle performance nel momento in cui si sfrutta

un numero maggiore di osservazioni per l'addestramento, sia duplicate reali sia sintetiche.
Adottando random oversampling o SMOTE si riesce ad ottenere un modello in grado di prevedere efficacemente le osservazioni appartenenti alla classe rara. La Random Forest addestrata con random oversampling viene utilizzata nell'analisi successiva per la creazione di un modello più parsimonioso mediante una fase preliminare di Feature Selection e, successivamente, per la stima dell'indicatore di *Propensity to churn*

3 Feature Selection

Al fine di ottenere un modello più parsimonioso è stato deciso di implementare la feature selection, ossia la selezione delle caratteristiche maggiormente significative rispetto alle altre. Nel nostro caso si è scelto di utilizzare l'approccio di selezione Univariato, questo approccio funziona selezionando le migliori caratteristiche sulla base di test statistici univariati. Attraverso il test del chi-quadro sono stati attributi dei valori d'importanza ad ogni variabile di input del modello. È stato deciso di prendere in considerazione solo gli attributi il cui valore calcolato fosse maggiore di 100, di conseguenza, si è ottenuto un nuovo dataset con solo 20 variabili di input rispetto alle 75 precedenti. Di seguito vengono mostrate le 20 variabili selezionate.

Specs	Score
how_many_ok_urls	2960.018427
STB_HD	2700.251731
SPORT	1400.260803
categories_shopping	1338.775253
SKY_FAMIGLIA	1007.445180
categories_science	982.586361
how_many_ko_urls	979.685072
FLG_MV	933.097267
FLG_MYSKYHD	915.857791
CINEMA	855.605119
CALCIO	661.140651
categories_technologyandcomputing	505.378480
FLG_SKY_ON_DEMAND	489.173976
STB_MYSKYHD	380.814699
categories_uncategorized	341.148788
categories_sports	281.930905
FLG_MYSKY	228.713395
categories_foodanddrink	190.529817
FLG_HD	176.445615
categories_news	149.398778

Figura 1: Features più significative

Il passo successivo è stato quello di addestrare il modello di Random Forest con oversampling che ha ottenuto le seguenti performance:

Classificatore	Campionamento	Accuracy	Precision	Recall	F-Measure	AUC
Random Forest	Random Oversampling	0.91	0.94	0.87	0.90	0.97

Tabella 6: Random Forest con Feature Selection

Come si può vedere dalla tabella numero 6, nonostante le variabili di input siano diminuite del 74% le performance del modello sono rimaste pressoché invariate, con tutti gli indicatori di performance che si discostano di pochi punti percentuali rispetto al modello addestrato con tutte le variabili di input. Pertanto il modello addestrato con feature selection è da ritenersi valido per l'analisi in questione.

4 Propensity to churn

Una volta addestrata la Random Forest con random oversampling, il modello è stato utilizzato per predire la probabilità che un utente possa essere un churner oppure no. Per ogni osservazione appartenente al test set è stata predetta tale probabilità con lo scopo di andare a verificare statisticamente se effettivamente le predizioni del modello corrispondessero alla realtà.
Il test set è stato ordinato in ordine decrescente in base all'indicatore di *Propensity* al churn, successivamente il set è stato suddiviso in decili ottenendo i seguenti risultati:

```
Decile 100% :

Propensity to churn media: 0.9267377148613825

Percentuale di churners: 99.81285090455397
Percentuale di non churners: 0.18714909544603867

--------------------------------------------------

Decile 90% :

Propensity to churn media: 0.893022181514547

Percentuale di churners: 98.90829694323145
Percentuale di non churners: 1.0917030567685588

--------------------------------------------------

Decile 80% :

Propensity to churn media: 0.8385736450841161

Percentuale di churners: 97.87897691827823
Percentuale di non churners: 2.1210230817217717

--------------------------------------------------

Decile 70% :

Propensity to churn media: 0.7369844359259867

Percentuale di churners: 93.4341859014348
Percentuale di non churners: 6.565814098565189

--------------------------------------------------

Decile 60% :

Propensity to churn media: 0.5906085326647835

Percentuale di churners: 82.7666874610106
Percentuale di non churners: 17.233312538989395

--------------------------------------------------

Decile 50% :

Propensity to churn media: 0.3779230656947566

Percentuale di churners: 24.142233312538988
Percentuale di non churners: 75.85776668746101

--------------------------------------------------
```

Figura 2: Primi 6 decili e Propensity

Si può notare come effettivamente i decili in cui la Propensity media è molto alta (nel range tra 0.73 e 0.93) comprendono una percentuale molto alta di churners. Per i primi tre decili, la Propensity media varia nel range tra 0.83 e 0.93 e la quasi totalità (97.9% - 99.8%) delle osservazioni risulta appartenere alla classe dei churners. Nel quarto e nel quinto decile, la Propensity media è significativamente più bassa rispetto a quella dei primi tre decili (rispettivamente 0.73 e 0.59) ma nonostante ciò la percentuale di churners in questi decili è alta lo stesso (rispettivamente 94.4% e 82.7%). Nel sesto decile la Propensity media è 0.38, la percentuale di churners cala drasticamente al 24.1%. Confrontando le statistiche dei primi cinque decili con quelle del sesto è possibile dedurre che in effetti è molto probabile che il modello riesca ad etichettare un churner con una Propensity superiore a 0.5.

```
Decile 40% :

Propensity to churn media: 0.26569866577250567

Percentuale di churners: 2.885215221459763
Percentuale di non churners: 97.11478477854024

--------------------------------------------------

Decile 30% :

Propensity to churn media: 0.20792404283505045

Percentuale di churners: 0
Percentuale di non churners: 100.0

--------------------------------------------------

Decile 20% :

Propensity to churn media: 0.15625694325424214

Percentuale di churners: 0.140361821584529
Percentuale di non churners: 99.85963817841547

--------------------------------------------------

Decile 10% :

Propensity to churn media: 0.09583670875202005

Percentuale di churners: 0
Percentuale di non churners: 100.0

--------------------------------------------------
```

Figura 3: Ultimi 4 decili e Propensity

Le statistiche degli utlimi quattro decili sono una conferma di quanto emerso dall'analisi dei primi sei. La Propensity media è molto più bassa di quella dei primi sei, così come la percentuale di churners.
In conclusione, il modello riesce con molta probabilità ad etichettare con una Propensity maggiore a 0.5 i churners.

5 Conclusioni e Sviluppi Futuri

In questa analisi sono stati addestrati due modelli che hanno registrato performance ottime nel task di predizione della variabile di churn *Pdisc*: si tratta dei due modelli di Random Forest addestrati con tecniche di campionamento SMOTE e Random Oversampling. La Random Forest con random oversampling è stata successivamente addestrata anche in una versione più parsimoniosa con sole 20 variabili di input rispetto alle 75 iniziali, risultando essere comunque un modello estremamente valido. La stima della Propensity al churn e la successiva verifica sui dati reali ha reso evidente che il modello identificato risulta essere valido e utilizzabile ai fini del business.
Uno sviluppo futuro di questa analisi potrebbe essere la validazione del modello e il test su dati più recenti e raccolti in un range temporale più ampio, in quanto la grande sfida dei modelli anti-churn è proprio quella di mantenere alte le performance nel tempo e con dati raccolti in istanti temporali successivi e distanti dall'addestramento dello stesso.

6 Librerie Utilizzate

- seaborn;
- pandas;
- matplotlib;
- numpy;
- scikit-learn;
- glob;
- os;
- pathlib;
- tensorflow;
- tensorflow_addons;
- keras;
- statistics.

7 Bibliografia

Di seguito vengono elencate le librerie utilizzate per lo svolgimento del progetto:

- `https://seaborn.pydata.org`;
- `https://pandas.pydata.org/docs/`;
- `https://matplotlib.org`;

- https://numpy.org/doc/;
- https://scikit-learn.org/stable/modules/model_evaluation.html;
- https://docs.python.org/3/library/glob.html;
- https://docs.python.org/3/library/os.html;
- https://docs.python.org/3/library/pathlib.html;
- https://www.tensorflow.org/api_docs/python/tf/all_symbols;
- https://www.tensorflow.org/addons;
- https://keras.io/api/;
- https://docs.python.org/3/library/statistics.html;