Analisi del dataset “Telco Dataset Churn”

Autori:

- Giuliano Bisinella - 2139735 - bisinella.2139735@studenti.uniroma1.it
- Bernardo Casarosa - 2152655 - casarosa.2152655@studenti.uniroma1.it
- Stella Sofia Cillis - 2141304 - cillis.2141304@studenti.uniroma1.it

Abstract

Il progetto analizza il fenomeno del *Customer Churn* utilizzando il dataset *"Telco Customer Churn"*. L'obiettivo è svolgere un’analisi classificatoria al fine di identificare i clienti a rischio abbandono.
Dopo un'analisi esplorativa dei dati e un preprocessing mirato, che verranno più approfonditamente esplicati nei prossimi paragrafi, sono stati addestrati diversi modelli di regressione logistica, volti ad una progressiva analisi e comparazione degli stessi:
dapprima è stato testato un modello “grezzo” di regressione il quale dà eguale peso agli errori compiuti sulla classe 0 e sulla classe 1; consapevoli dello sbilanciamento del dataset e della conseguente incorrettezza di un tale trattamento degli errori, abbiamo proceduto alla creazione di un secondo modello, volto a correggere tale iniquità.
È stata successivamente posta enfasi sull'ottimizzazione della *threshold* della funzione sigmoide mediante la metrica F2, al fine di ottenere un modello più performante e coerente con il nostro obiettivo.
Infine, sono stati testati modelli "tematici" (Finanziario, Tecnico, Demografico), tentando di circoscrivere l’analisi a gruppi di features più o meno correlate con la variabile target.

Tra tutti, il modello che si rivelerà essere più performante sarà il **modello tematico finanziario**, composto da cinque features (selezionate delle ventinove presenti dopo la dummificazione) inerenti agli elementi, appunto, meramente finanziari del rapporto impresa-cliente, e con una *threshold* inferiore a quella standard.

Introduzione e definizione del problema

Contesto

Il dataset si inserisce nel dominio delle telecomunicazioni, dove la *customer retention* è cruciale. L'analisi dello stesso mira a comprendere la presenza di fattori che possono portare un cliente ad abbandonare il servizio.

Quesito dell'analisi

L'intento analitico, decisamente pragmatico, si concretizza nella seguente domanda:
"è possibile prevedere il rischio di abbandono di un cliente, al fine di permettere azioni commerciali preventive?"

Descrizione del dataset

Il dataset `WA_Fn-UseC_-Telco-Customer-Churn.csv` contiene 7043 records e 21 campi (Churn inclusa), ovviamente ripartite in:

- **Target:** `Churn`
- **Feature:** inerenti a informazioni demografiche, servizi sottoscritti e informazioni contabili.

Analisi Esplorativa dei Dati (EDA)

Pulizia iniziale

Appena caricato il dataset e visualizzate le prime informazioni essenziali riguardo samples e features, ci accorgiamo di due dettagli:

- Il primo campo, `CustomerID`, è assolutamente irrilevante e non apporta alcun contributo utile al nostro scopo; decidiamo di rimuoverlo.
- Il campo `TotalCharges` risulta essere l'ovvio prodotto di `MonthlyCharges` e `tenure`. Dato il rischio di multicollinearità insito nella permanenza dei tre campi nel dataset, decidiamo di rimuovere anche questo.

Visualizzazioni chiave

L'EDA è stata condotta tramite:

- **Matrice di Correlazione (Heatmap):** Evidenzia una forte ridondanza tra variabili (es. `No internet service` ripetuto su più colonne) e ci permette di intendere la correlazione di molte variabili, sia tra loro, sia con Churn
- **Scatterplot:** Relazione tra `tenure` e `MonthlyCharges`. Emerge con evidenza un pattern definito all'interno del grafico: si nota che i clienti con alta spesa mensile e bassa anzianità tendono ad abbandonare più frequentemente. Per accertarcene e meglio comprendere l'entità di tale fenomeno, decidiamo di generare dei subplots, ripartiti proprio in base alla durata del contratto;
- **Boxplot:** Osservando I boxplot notiamo che non abbiamo particolare rumore nei dati, segno molto positivo. Notiamo, come ci aspettavamo, che:
    - In media chi ha speso maggiormente mensilmente ha abbandonato il servizio;

    - In media ha abbandonato il servizio chi è rimasto per meno tempo, tranne qualche eccezione rappresentata come rumore sul grafico.

Bilanciamento classi

La visualizzazione di alcune statistiche sulla variabile target `Churn` ci permette di introdurre l'analisi del bilanciamento delle classi: una media di ~ 0.265 mette in risalto la presenza di un numero contenuto di clienti che disdicono rispetto al totale.

Una più approfondita analisi conferma quest'intuizione: circa il **73%** dei clienti rimane (Churn=0) mentre il **27%** abbandona (Churn=1).
Tale sbilanciamento introduce un problema: un modello che dovesse predire sempre "No Churn" avrebbe una Accuracy fuorviante del 73% ma una Recall nulla.
Per ovviare a questo problema, abbiamo proceduto seguendo due modalità:

- abbiamo utilizzato (a partire dal secondo modello) un peculiare argomento nella funzione di costruzione della regressione logistica (class_weight="balanced") in modo tale da penalizzare differentemente gli errori compiuti sulle due classi;

- abbiamo rivolto maggiore attenzione alla Recall, particolarmente sensibile ai falsi negativi.

Preprocessing e Feature Engineering

Selezione delle Feature

In seguito all'analisi della heatmap sono state rimosse le colonne ridondanti generate dal One-Hot Encoding che portavano la stessa informazione e che avrebbero generato multicollinearità; le elenchiamo:

- `OnlineSecurity_No internet service`
- `OnlineBackup_No internet service`
- `DeviceProtection_No internet service`
- `TechSupport_No internet service`
- `StreamingTV_No internet service`
- `StreamingMovies_No internet service`
- `MultipleLines_No phone service`.

Encoding

Le variabili categoriche sono state trasformate, subito dopo la ricerca dei valori nulli, tramite **One-Hot Encoding** (`pd.get_dummies`). Utilizziamo inizialmente l'argomento `drop_first=False` per preservare tutte le features prodotte dalla dummificazione, ai fini di una visualizzazione completa e dettagliata all'interno dei grafici.
In un secondo momento, al fine di evitare multicollinearità, abbiamo riprodotto l'encoding,

questa volta utilizzando l'argomento `drop_first=True`. Utilizzeremo questo secondo encoding all'interno dei modelli prodotti.

Normalizzazione

Come tecnica di normalizzazione usiamo MinMaxScaler, la quale comprime le features numeriche entro un intervallo [0, 1], senza considerare le features dummificate le quali già risiedono all'interno dello stesso intervallo. Evitiamo la tecnica di standardizzazione perché non gode della stessa proprietà e diversi dati rimarrebbero al di fuori dell'intervallo [0, 1]. Questo potrebbe portare ad attribuire maggior peso a variabili che esulano da tale intervallo.

Metodi e Modelli

È stata scelta la **Regressione Logistica** come modello per risolvere questo problema di classificazione.
A differenza della regressione lineare, che prevede valori continui, la regressione logistica stima la probabilità che un dato sample appartenga ad una tra due classi, risultando ottimale per il problema da noi affrontato.

Strategia di Modellazione

Abbiamo proceduto mediante la creazione di tre modelli principali, comprendenti tutte le features (al netto dunque di quelle già eliminate), e tre modelli "tematici", incentrati su un numero ristretto di features differenti per ciascuno, al fine di valutare la maggiore o minore incidenza delle stesse sulla probabilità di disdetta da parte di un cliente.

Esponiamo i primi tre modelli:

1. **Modello Base (Non bilanciato):** Regressione logistica "standard", senza alcun accorgimento tecnico particolare.
   Memori di quanto detto prima circa lo sbilanciamento delle classi e l'Accuracy calcoliamo quest'ultima, ottenendo un esito di poco sufficiente: ~80%.
   La Recall si attesta invece intorno al 55%. Il modello sta evidentemente prediligendo la classe maggioritaria, con un conseguente innalzamento, per altro non molto soddisfacente, dell'Accuracy. Decidiamo di tralasciare, da qui in poi, la considerazione di quest'ultima metrica.
   Il valore AUC si attesta a 0,72, riflettendo la possibilità limitata di distinguere l'appartenenza all'una o all'altra classe, dovuta dallo sbilanciamento del dataset a favore dei clienti fedeli
2. **Modello Bilanciato:** Utilizziamo in questo modello l'argomento `class_weight="balanced"` per penalizzare maggiormente gli errori sulla classe minoritaria (ossia Churn=1).
   Questo aiuterà ad ottenere un modello più equilibrato, bilanciato appunto.
   Gli esiti sono evidenti: la Recall sale al 78%, mentre la Precision si attesta sul 51%.
   Al fine di ottenere esiti più soddisfacenti, decidiamo di costruire un ulteriore modello

nel quale abbassare la *threshold* non in modo arbitrario, bensì usando una metrica di ottimizzazione.
Il valore AUC stavolta si attesta a 0.75, notiamo come l'introduzione del parametro `"balanced"` migliori la capacità di distinguere le classi.

3. **Modello con Soglia Ottimizzata (F2 score):** Qui si introduce una modifica sostanziale, utile anche al nuovo calcolo della soglia ottimizzata: l'utilizzo della cross-validation. Questa tecnica permette al modello di calcolare probabilità basandosi, per ciascuna delle 5 iterazioni, su una differente porzione del training set non utilizzata per l'addestramento, evitando una "fuga di dati" e una conseguente alterazione ottimistica del risultato.
   La cross-validation ha dunque calcolato, per ogni x appartenente al training set, la probabilità che Churn(x)=1.
   Si colloca qui il calcolo della soglia ottima mediante la metrica F2, intrinsecamente volta alla penalizzazione dei falsi negativi e ad un conseguente innalzamento della Recall.
   Riportiamo la formula della F2[1]:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

   L'abbassamento della soglia è sostanziale: passa dallo standard 0.5 a 0.35, causando un innalzamento della Recall (90%) dovuto ad un abbassamento dei falsi negativi, con un conseguente aumento dei falsi positivi: la Precision, in effetti, si attesta sul 45%. Giustificheremo più avanti la bontà di tali risultati.

Modelli Tematici

Per indagare quali aspetti pesano di più, sono stati addestrati tre modelli separati su specifici sottoinsiemi di feature. Li riportiamo di seguito:

- **Gruppo Finanziario:** `MonthlyCharges`, `Contract`, `PaymentMethod`, `Contract_One year, Contract_Two year, PaperlessBilling`.
- **Gruppo Tecnico:** `InternetService`, `TechSupport`, `OnlineSecurity`.
- **Gruppo Demografico:** `SeniorCitizen`, `Partner`, `Dependents`, `Gender`.

I risultati ottenuti sono significativi, seppur in sensi diametralmente opposti: li commenteremo nella prossima sezione.

Valutazione e Analisi dei Risultati

Premessa e necessarie assunzioni

Data la natura del dataset, eminentemente commerciale, il pragmatico obiettivo dell'analisi è quello di ridurre il più possibile l'**erronea previsione** di **permanenza** del cliente (falso negativo), la quale, assumiamo, apporta un danno economico ben superiore rispetto all'**erronea previsione**, opposta e contraria, di **abbandono** di un cliente: se quest'ultima ipotesi potrà risultare in un "aumento della soglia d'allarme", richiedendo un aumento di costi legati al tentativo di mantenere il legame commerciale attivo, la prima ipotesi risulterebbe in una perdita economica nettamente superiore, data la disdetta del contratto.
Tale assunzione è necessaria per l'elezione di un modello "vincitore".

Metriche

Dato quanto detto sopra, la metrica guida sarà la **Recall**, la quale esprime la capacità di trovare tutti i clienti che abbandonano, bilanciata, seppur con minor peso, dalla Precision.
La ragione alla base di questa scelta consiste proprio nella volontà di valorizzare il più possibile la riduzione di falsi negativi, tentando però di evitare il problema diametralmente opposto a quello presente in un modello "naive": un modello che dovesse prevedere sempre Churn = 1 massimizzerebbe inevitabilmente la Recall, ma sarebbe altrettanto sbagliato.
Per questa ragione consideriamo, accanto alla Recall, anche la Precision, in quanto è in grado di fornire un riferimento solido in casi, come questo, di dataset sbilanciati, mantenendo l'attenzione anche sui falsi positivi, il cui **eccessivo** innalzamento avrebbe certamente e comunque un impatto economico negativo.
L'Accuracy è considerata secondaria a causa dello sbilanciamento: la ragione è esposta nel paragrafo **§Bilanciamento classi**.

Quanto detto sopra giustifica la nostra volontà di andare oltre un "semplice bilanciamento" tra le due metriche appena citate: la nostra predilezione, che potrà apparire lievemente aggressiva, per un aumento della Recall, vuol essere un mezzo di ottimizzazione economico-commerciale.

Confronto Risultati

(I valori numerici esatti dipendono dall'esecuzione finale del codice, qui riportiamo l'analisi qualitativa basata sul codice)

1. **Modello Base:** Alta Accuracy (~80%) ma **bassa Recall** (~55%). Il modello tende a ignorare i clienti che abbandonano, prevedendo troppo spesso la permanenza del cliente (1'091 volte su 1'409).
   Questo avviene a causa della negativa influenza del forte sbilanciamento delle classi, non compensata in alcun modo da una differente valutazione degli errori compiuti sulle due classi. Rimandiamo a quanto esposto nei paragrafi **§Bilanciamento classi**

e **§Metodi e modelli**.

2. **Modello Bilanciato:** L'introduzione di `class_weight="balanced"` abbassa leggermente l'Accuracy totale ma alza significativamente la Recall, intercettando più clienti a rischio.
   La ragione segue da quanto detto al punto superiore.
   Interessante la valutazione del costo che, con l'aggiunta dell'argomento sopra citato, viene ripartito secondo tale calcolo:
   (Totale campioni / (numero classi * frequenza classe)).
   Nel caso della classe 0, il peso assegnato a ciascun errore sarà
   (100 / (2 * 73)) = 0,68.
   Nel caso della classe 1, il peso assegnato a ciascun errore sarà
   (100 / (2 * 27)) = 1,85.
   Il contributo di quel singolo argomento è enorme: la Recall cresce di 24 punti percentuali raggiungendo il 78%, mentre Precision si attesta su un 51%.
   Il dimezzamento dei falsi negativi è (apparentemente) controbilanciato dal raddoppio dei falsi positivi: in questa specifica analisi, dato lo scopo perseguito, questo risulta essere un costo sopportabile.

3. **Modello con Soglia Ottima:** con l'abbassamento della soglia sopra menzionato (che, ripetiamo, scende a 0.35 con il calcolo della F2), i risultati cambiano drasticamente:
   La Recall su Churn = 1 sale sino all'90% mentre la Precision scende al 45%: prevediamo correttamente la disdetta di 9 clienti su 10, ma più della metà delle previste disdette sono, in realtà, falsi allarmi.
   In attesa di valutare i modelli successivi, e fermi sul punto e sull'obiettivo prima esposto, giudichiamo comunque positivamente il risultato ottenuto.

Analisi "Modelli Tematici"

Dal confronto dei sotto-modelli (Finanziario vs Tecnico vs Demografico) emerge che le variabili **Finanziarie** (tipo di contratto, metodo di pagamento, costi) sono i predittori più forti. Le variabili Tecniche hanno un impatto decisamente minore, mentre le sole variabili Demografiche hanno potere predittivo praticamente nullo.
Abbiamo deciso, in tutti e tre i modelli tematici, di considerare l'ipotesi migliore, implementando le migliorie prima sperimentate sul "modello base": bilanciamento dell'errore calcolato sulle diverse classi e calcolo della soglia ottimale mediante F2.
Abbiamo deciso di creare una funzione per racchiudere in un'unica cella di codice i calcoli prima, per altre ragioni, frazionati e ripetuti.

Vediamo un'analisi più approfondita:

1. **Modello Finanziario**: alcune delle variabili qui presenti, più specificamente MonthlyCharges, Contract_One year e Contract_Two year, sono le stesse utilizzate nell'analisi soprastante, visualizzata negli scatterplot: avevamo dunque già avuto evidenza della forte correlazione che sussisteva tra queste features e Churn.
   In effetti tale modello restituisce ottimi esiti: 92% di Recall, con una gestione altamente efficace dei falsi negativi, ma 41% di Precision, 4 punti percentuali al di

sotto del precedente. Il valore AUC è di 0.79, dimostrando come le variabili di costo siano i predittori singoli più forti per il Churn.

2. **Modello Tecnico**: la sua costituzione deriva da un'evidenza di forte correlazione, emersa già con l'osservazione delle heatmap, tra le feature che lo compongono e la variabile target Churn.
   Il risultato però non appaga le aspettative: la Recall è nuovamente al 90%, ma la Precision scende al 35%: notiamo, rispetto al modello precedente:
   a. un lieve aumento dei falsi negativi;
   b. un importante aumento dei falsi positivi (superano, per la prima volta, i veri negativi);

   La soglia ci restituisce una chiara motivazione di questo esito: essendo stata abbassata dalla F2 a 0.20, i falsi positivi hanno subito un'impennata che giudichiamo, ora sì, decisamente eccessiva.
   L'AUC scende nuovamente a 0.75, denotando il minor impatto delle variabili tecniche sulla decisione di abbandono rispetto a quelle economiche.

3. **Modello Demografico**: la costruzione di questo gruppo parte dalla volontà opposta rispetto alla strutturazione dei due precedenti: dimostrare il basso impatto di queste features sulla creazione di un modello efficiente.
   I risultati trascendono questa nostra intuizione: la rilevanza delle features demografiche sulla variabile target è talmente bassa (come già vedevamo nella heatmap) che la F2, al fine di innalzare la Recall, spinge la soglia a 0.
   Questo restituisce effettivamente una Recall del 100% e un annullamento dei falsi negativi, il che potrebbe essere considerata una vittoria se non si dovessero fronteggiare ora 1305 falsi positivi, il cui impatto economico sarebbe rovinoso.
   Includiamo comunque il risultato nel confronto finale, per completezza.
   L'AUC si riduce a 0.65, notiamo dunque come le caratteristiche anagrafiche abbiano scarso potere discriminante se prese singolarmente

Conclusioni

Come già detto, a fronte del tipo di dataset che abbiamo deciso di analizzare, ossia dei clienti di un'azienda di telecomunicazioni, l'approccio che abbiamo usato è in ultima analisi "business-oriented", ossia abbiamo dato priorità alla riduzione dei falsi negativi (in quanto economicamente più impattanti dei falsi positivi) utilizzando la Threshold consigliata dalla metrica F2, che considera la Recall addirittura quattro volte più importante della Precision.

Rispetto ai modelli tematici, facendo una controprova con `drop_first=False` abbiamo notato come l'assenza in essi delle features "eliminate" da `drop_first=True` non cambi in alcun modo i risultati.

Come emerge dall'analisi dei dati, l'azienda dovrebbe concentrarsi sui clienti con contratti mensili (Contract_Month-to-month), pagamenti elettronici (Electronic check) e servizio internet con annessa fibra ottica (InternetService_Fibre optic), che risultano essere i più

volatili.
Qualche operazione di marketing mirato, un potenziamento del servizio clienti e qualche eventuale leggera riduzione del canone attuale sono azioni che potrebbero ridurre il rischio di abbandono, limitando le perdite economiche derivanti.
Il nostro modello tematico si spinge proprio in questa direzione: minimizza i falsi negativi, dunque i clienti considerati “non a rischio”, a fronte di un aumento dei clienti considerati, di contro, a rischio disdetta: è a loro che saranno rivolte le azioni mirate ora citate, al fine di fortificare e sviluppare un importante processo di fidelizzazione.
Quest’obiettivo è supportato da dati già presenti nel dataset: tra i clienti fedeli (con alti valori di tenure), e soprattutto tra coloro che hanno un contratto biennale, il rischio di Churn si abbatte considerevolmente, anche in fasce alte di prezzo del servizio.

Di contro, limitare l'analisi alle sole variabili demografiche è inefficace, come platealmente dimostrato, appunto, dal modello tematico demografico.

Pur mantenendo ferma la nostra iniziale assunzione (§Premessa e necessarie assunzioni), variamente ripresa lungo l’intero documento, siamo consapevoli che il limite maggiore del nostro approccio è diretta conseguenza proprio della massimizzazione della Recall, che porta inevitabilmente ad una minor Precision, dunque ad un numero potenzialmente elevato di falsi positivi: si tratta del fenomeno degli “Sleeping dogs”, ossia dei clienti insoddisfatti che continuano ad usufruire di un servizio per inerzia o disattenzione. Andare a “svegliare il can che dorme”, ossia proporre offerte a persone che sarebbero rimaste pur a prezzo pieno, può produrre in molti casi l’effetto contrario.

[1] Abbiamo trovato la formula nella pagina Wikipedia (https://en.wikipedia.org/wiki/Precision_and_recall) linkata alla diapositiva 8 del pacco slides “08_metrics_model_selection”