Generali Data Challenge: Churn Prediction

Massimiliano Rubino

January 19, 2021

1 Introduzione

Il mancato rinnovo di una polizza da parte di un cliente (Customer Churn) è un problema classico per una compagnia assicurativa.Un problema a cui vanno incontro tutti i modelli di business basati su contratti a scadenza. La previsione del Customer Churn è il tema di questa Challenge. Generali ha messo a disposizione dati estratti dai propri sistemi, anonimizzando il significato di esse, e fornendo due Dataframe uno di train e l'altro di test sul quale verrà valutato il modello predittivo.
Il dataframe di train è cosi formato:

- index: identificativo osservazione;
- feature_0, . . . , feature_294: features anonimizzate. Sono presenti features quantitative e categoriali. Sono presenti dati mancanti identificati da uno spazio vuoto;
- target: vettore di 0 e 1. L'uno identifica il Churn del cliente.

La metrica scelta per valutare il modello è l'F1 Score, cioè la media armonica di precisione e recupero (precision e recall) per la categoria target=1 (Churn):

$$F_1 = \frac{(Precision_1 * Recall_1)}{(Precision_1 + Recall_1)}$$

Il modello selezionato come previsione migliore è stato un' Ensambe di due modelli (GredientBusting e Regressione Logistica) con oversampling della classe minore , modificando il threshold da 0.5 a 0.45. In questo modo il modello è riuscito a migliorare la sua capacità predittiva,diminuendo la percentuale di falsi churn.

2 Summary of the modelling process

1. Preprocesing

   Prima di iniziare con l'analisi, sono stati aggreagati i due dataframe in modo tale da riuscire a gestire missing che si trovavano nel test set e non nel train.
   Seconda operazione effettuata è stata quella di andare a verificare la presenza di variabili a varianza nulla, queste non danno un contributo per la previsione e per questo

vanno tolte. Ne sono state trovate 55 che sono state tolte. Il dataframe è passato cosi da 297 a 242 variabili.
Successivamente è stata costruita una funzione che rilevasse le variabili a varianza vicino lo zero (near zero variance). Una volta rilevate, queste variabili sono state eliminate poiche non apportavano grande contributo alla previsione finale. Il dataframe è passato da 242 a 147 variabili.
Eliminate tutte le variabili che presentavano una correlazione superiore a 0.99

2. Missing values

   I missing non sono stati imputati in maniera diretta, ma sono state create delle dummy che identificavano qualora l'osservazione era mancante o meno.

3. Feature engineering

   In modo tale da utilizzare la variabile 6 (feature_6), è stata utilizzata una tecnica di encoding facendola diventare continua.

   Le variabile che presentavano meno di 5 livelli sono state trasformate in Object cioè discretizzate.

4. Feature selection

   Intrinseca nel modello

5. Final model

   IL modello finale scelto è stato un' Ensambe di due modelli (GredientBusting e Regressione Logistica) con oversampling della classe minore, modificando il threshold da 0.5 a 0.45.

6. Model tuning and evaluation

   Cross-validation stratificata con RandomOverSampler, k=5 e ripetuta 2 volte. Valutazione del tuning parameter attraverso F1 measure.

   Il tuning è stato effettuato solo per il modello GredientBusting ottenendo i seguenti parametri:

   learning rate=0.01, max depth= 5, max features= 0.1, min samples leaf= 10, n estimators= 300

7. Python packages

   Pandas, Numpy, Category_encoders, Sklearn, Collections, Imblearn