Master Universitario in: "ANALISI DATI PER LA BUSINESS INTELLIGENCE E DATA SCIENCE" A.A. 2021/2022

Titolo della tesi: Ottimizzazione dello sconto per le polizze RC Auto

Autore: Marco Vita

Abstract

Il presente lavoro ha come obiettivo il miglioramento dell'algoritmo che calcola lo sconto proposto per la garanzia RC delle polizze auto in fase di preventivazione.

La soluzione ipotizzata prevede la creazione di un'API che riceva le chiamate dall'applicativo di emissione polizze nella fase di compilazione del preventivo e restituisca una percentuale di sconto ottimizzata e personalizzata in base alle caratteristiche del cliente e veicolo oggetto di copertura.

Viene utilizzato un modello di Machine Learning per la stima della probabilità di conversione del preventivo in polizza. Il modello è stato costruito in Python utilizzando l'algoritmo Random Forest.

Per il training del modello sono stati estratti circa 100mila preventivi che contengono dati anagrafici del cliente, dati tecnici del veicolo, composizione della polizza (garanzie aggiuntive, massimali…) e dati dell'attestato di rischio (storico dei sinistri). Sono stati poi agganciati al portafoglio polizze per ricavare la variabile target che indica se il preventivo è stato convertito o meno (flag 1 o 0).

I preventivi vengono arricchiti con variabili aggiuntive calcolate da due modelli attuariali: premio di rischio, ovvero importo necessario a compensare il costo stimato dei sinistri, e premi dei competitor, stimati grazie a modelli costruiti a partire da analisi di mercato.

Il modello di conversion è un Random Forest di classificazione binaria dove la classe 1 è minoritaria (15%). Come risultato per i passi successivi non verrà utilizzata la classe predetta ma un altro output restituito: la probabilità che la classe sia 1.

Per ottimizzarne i parametri è stato eseguito un processo di grid search: si definiscono più combinazioni di valori e per ognuna si esegue un training del modello, selezionando infine il set di parametri che ottiene il miglior risultato per la metrica target.

Integra inoltre un sistema di cross validation per utilizzare ad ogni esecuzione una porzione del dataset di training come dev set.

L'accuratezza complessiva del modello misurata sul test set è 0.87 con un'area sotto la curva ROC pari a 0.9

L'algoritmo di ottimizzazione prevede la definizione dei corridoi, ovvero estremo superiore e inferiore per il premio da proporre al cliente. Il minimo è fissato pari al premio di rischio con il contributo di alcuni coefficienti moltiplicativi/additivi per le spese di compagnia. Il massimo è calcolato come media dei 5 premi dei competitor più bassi.

Tra questi due estremi viene calcolato un range uniforme di importi a cui applicare modello di conversion (probabilità che la classe sia 1).

Con i dati ottenuti nel passaggio precedente è possibile calcolare il profitto atteso per ognuno degli importi che fa parte del range. Viene selezionato come premio finale quello a cui corrisponde il maggior profitto atteso e convertito in percentuale calcolando il rapporto rispetto alla tariffa di partenza.

Il data product finale è un'API pubblicata sull'ambiente Azure Cloud utilizzando un App Service Container che si appoggia a CosmosDB, database NoSQL efficiente per dati non strutturati in forma tabellare utilizzato per memorizzare i parametri necessari. Per il training del modello si sfrutta l'ambiente Databricks per la parte di ETL e Azure ML per monitorare il processo di training, risultati e metriche.

Il risultato ha soddisfatto gli obiettivi iniziali e pone le basi per ulteriori sperimentazioni dei modelli di ML all'interno dei processi aziendali.