ANNO ACCADEMICO 2021/2022
DATA SCIENCE PER L'ECONOMIA E LE IMPRESE
DATA ANALYSIS FOR BUSINESS DECISIONS

Assignment 3

Regressione logistica per l'analisi di una campagna di marketing

A CURA DI: RICCARDO ROCCHI, BENITO DEL PESCHIO, CHIARA ANTOGNOZZI E FEDERICA VOLPI

Obiettivo

Lo scopo della seguente analisi è quello di prevedere se un potenziale cliente aderirà o meno a una campagna marketing effettuata da un'azienda che offre depositi a termine. L'analisi di quest'ultima ha permesso di determinare la sua efficacia sui clienti finali, analizzando i fattori coinvolti, quali informazioni proprie del cliente, caratteristiche sulla campagna di marketing e differenti indici economici.

Metodologia

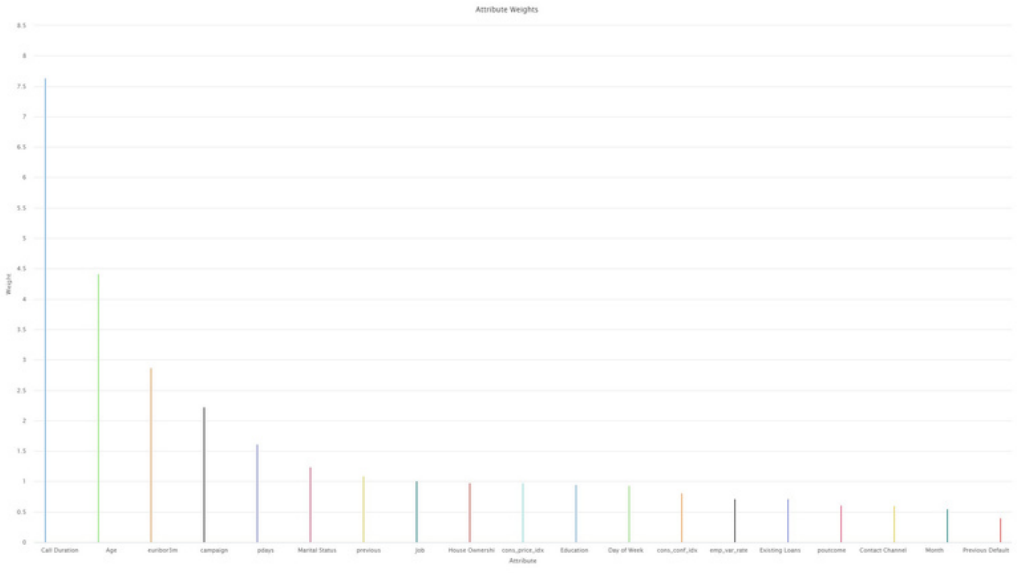
La variabile target è di tipo binario, assume cioè valore 1 nel caso in cui la proposta abbia successo, ossia il cliente accetta e decide di aprire un contratto, 0 viceversa. Il dataset risulta essere fortemente sbilanciato verso il valore 1 (circa il 88,7% contro 11,3% della classe minoritaria).

Analisi delle variabili

Nello specifico, il primo step effettuato è stato analizzare le variabili, cercando di individuare quelle che a primo impatto sembrano essere le più rilevanti:

1. Variabile 'age': risulta che gli individui maggiormente contattati sono quelli della fascia 30-50, seppur i maggior aderenti alla campagna, in proporzione, si concentrino nelle estremità, ovvero persone molto giovani o anziane;
2. variabile 'month': riferita al mese in cui è avvenuta la chiamata, questa variabile presenta valori che si concentrano maggiormente nei mesi di Maggio, Giugno, Luglio, Agosto e Novembre, sebbene i mesi dove risultano esserci meno chiamate siano quelli in cui si riscuote in proporzione maggior successo;
3. variabile 'empvarrate': si riferisce alla variazione del tasso di occupazione nazionale. All'aumentare della variazione, si nota una minore propensione a sottoscrivere il contratto;
4. variabile 'euribor3m': Euribor è l'indice di riferimento dei mutui a tasso variabile. All'aumentare dei valori assunti dall'indice, la propensione a sottoscrivere il contratto diminuisce;
5. variabile 'education': a seguito del raggruppamento dei valori 'basic' perché semanticamente simili, si nota come, tra coloro che hanno frequentato l'università, la campagna di marketing ha riscosso maggior successo.

Features Selection



Il grafico di cui sopra mostra i pesi che le features assumono nell'analisi. La tecnica utilizzata per discriminarne l'importanza delle variabili è stata quella del Random Forest, che ha permesso di individuare come rilevanti la durata della chiamata, l'età del cliente, il tasso Euribor e il numero di volte di precedenti contatti con il cliente.

L'analisi

Essendo il dataset fortemente sbilanciato verso il valore 1 della variabile target, per evitare di costruire modelli con bias sono state utilizzate due differenti tecniche di bilanciamento delle classi: Metacost e undersampling. Queste hanno condotto all'implementazione di due modelli, entrambi basati sulla regressione logistica come classificatore.

	Regressione Logistica	
	Metacost	**Undersampling**
Solver	IRLSM	Coordinate Descent
Link	Logit	Logit
F-measure	55.99%	51.74%

Per la valutazione delle performance delle previsioni out of sample si ricorre all' F1-measure, metrica che permette di valutare la bontà del modello. Dai risultati ottenuti la regressione logistica con applicazione del metacost risulta essere la più efficace.

Considerazioni finali

Dall'analisi effettuata emergono diverse considerazioni:

- sembrerebbe più efficace rivolgersi a individui molto giovani o anziani che hanno un livello d'istruzione universitario;
- le campagne pubblicitarie dovrebbero essere condotte prevalentemente nei periodi il cui sia il tasso d'occupazione sia l'Euribor sono bassi;
- dal momento che i mesi in cui sono state effettuate meno chiamate (Marzo, Settembre, Ottobre, Dicembre), sono in termini relativi quelli che hanno riscosso maggior numero di adesioni, sarebbe consigliabile investire maggiormente in questi periodi.