KARALIS

In merito al problema di decisione sull'erogazione dei finanziamenti sottoposto dall'Azienda Findomestic Banca/S.R.l. (denominata Findomestic), nonché alla richiesta di "Proof of Concept", si sono percorse due strade, così da poter, da un lato, offrire una soluzione che possa ovviare al problema di *class imbalance* presente nel dataset, portando al risultato oggettivamente migliore, e dall'altro, proporre un sistema alternativo, qualora l'azienda non ritenesse opportuno affidarsi a un modello che parta dal presupposto di sbagliare, seppur in maniera non troppo pesante, gran parte delle previsioni.

1. Analisi oggettiva;
2. Analisi alternativa;

Analisi oggettiva

Tenendo conto della documentazione consegnataci, ovvero della matrice di perdita (Tabella 1), la miglior soluzione oggettiva proponibile non può che essere una risposta unimodale.

	Previsto		
Vero	Regolare	Contenzioso	Recupero
Regolare	0	50	10
Contenzioso	4000	0	30
Recupero	10	20	0

Tabella 1: Matrice di perdita.

La strada seguita prende spunto dalla *teoria dei giochi non collaborativa*, pertanto si consideri ogni singola previsione come un fattore di rischio calcolabile nel seguente modo:

dati i costi relativi ad una modalità **c1**, **c2** e gli errori commessi prevedendo erroneamente quella medesima modalità **x'**, **x'',** allora

$$rischio_{modalità} = c1_{modalità}x' + c2_{modalità}x''$$

Applicando la formula per ciascuna modalità si ottengono i seguenti risultati:

$$rischio_{regolare} = 4000_{regolare}x^{'} + 10_{regolare}x^{''}$$

$$rischio_{contenzioso} = 50_{contenzioso}x^{'} + 20_{contenzioso}x^{''}$$

$$rischio_{recupero} = 10_{recupero}x^{'} + 30_{recupero}x^{''}$$

Considerando che il rischio è più grande quanto più è maggiore di zero, è pacifico che la scelta più rischiosa sia quella di effettuare previsioni sulla modalità “regolare”: un solo errore potrebbe comportare un costo analogo al compiere dagli 80 ai 400 errori qualora si fossero scelte altre strade.

Ora, considerando che, nel training set affidatoci, le modalità sono così distribuite:

Regolare = 96.105% Contenzioso = 3.135% Recupero = 0.76%

e che, a causa di ciò, i modelli tendono a produrre come spiegazione dei dati una variabile unimodale pari a “Regolare”, accettando di commettere un ME del ~3.9%, si può notare come tale tasso di errore sia di molto superiore a quello che si otterrebbe usando una variabile unimodale pari a “Recupero”.

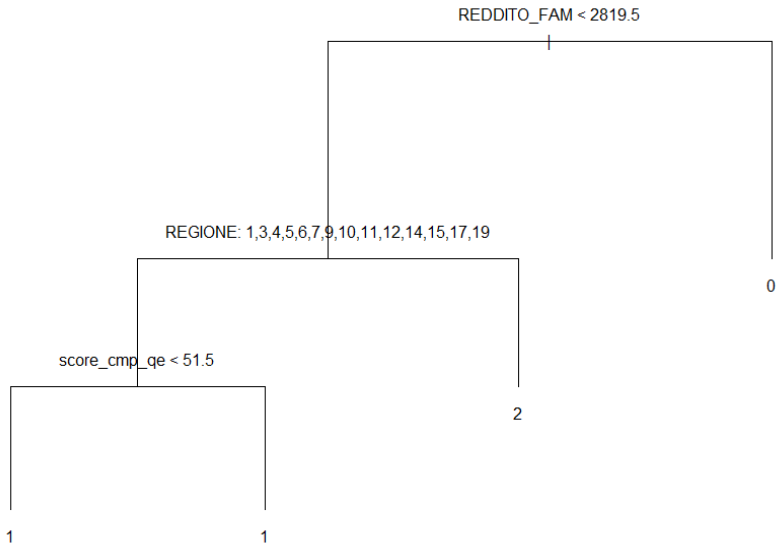
Secondo il ragionamento precedente, infatti, una spiegazione di questo tipo, è dunque la miglior soluzione oggettiva dal punto di vista statistico per quanto ci è dato conoscere, infatti minimizza il costo complessivo.

Analisi alternativa

Tenendo tuttavia conto della difficoltà per l’azienda di accettare un modello che parte “perdente”, accettando di prevedere modalità quasi sempre errate nell’ottica di ridurre i costi in base ai pesi segnalati, la soluzione alternativa che proponiamo consiste in un pretrattamento dei dati seguito dall’applicazione di un modello ad “alta spiegabilità”.

Per far fronte al problema di class imbalance e, dunque, poter consentire ai modelli di operare correttamente, si sono in primis rimosse tutte le osservazioni con valori mancanti, procedendo poi con un’operazione di bilanciamento: considerando che vi erano esclusivamente 304 variabili di tipo “Recupero”, si sono rimosse tutte le altre modalità per poi inserirne un campione selezionato casualmente con ampiezza pari a 304 per ciascuna.

In questo modo, tutte le modalità sono diventate equiprobabili e, trattandosi comunque di un campione sufficientemente grande, si è ritenuto applicabile il modello degli alberi.



Il risultato, nonostante presenti un ME del 58.11%, riesce ad avere un peso (dato dalla matrice di perdita) enormemente più basso di qualsiasi altro modello adoperato senza tale trattamento dei dati.