Master universitario di primo livello ADABI — Analisi dei Dati per la Business Intelligence e data science XIII edizione, anno accademico 2024/25

Studente: Giacomo Coggiola

Titolo della tesi: Dal dato grezzo al modello. Preparazione dei dati e regressione logistica per la previsione delle intenzioni turistiche in Piemonte

Abstract

Questo elaborato documenta l'attività di tirocinio curricolare svolta presso Metis Ricerche s.r.l., avente per oggetto la preparazione e la modellazione predittiva di dati provenienti da un'indagine campionaria sulle intenzioni di viaggio e vacanza degli italiani. L'indagine, condotta dall'azienda con cadenza stagionale a partire dal 2020 mediante tecniche CATI e CAWI, ha prodotto nel suo insieme sedici campagne di rilevamento e oltre 7.000 interviste, rivolte a campioni indipendenti di individui residenti in nove regioni italiane. La variabile di interesse principale è l'intenzione dichiarata di trascorrere una vacanza in Piemonte.

L'obiettivo operativo del progetto è la costruzione di una pipeline di modelli di regressione logistica in grado di simulare le risposte al questionario a partire dalle sole caratteristiche anagrafiche degli intervistati (genere, età, titolo di studio, condizione lavorativa, regione di residenza), fino alla previsione dell'intenzione di vacanza. Un tale strumento potrebbe trovare impiego nella generazione di dati sintetici utili alla pianificazione e alla valutazione delle strategie di promozione turistica.

Il lavoro si articola in tre capitoli. Il primo capitolo delinea il quadro metodologico: la distinzione tra dati primari e secondari, le dimensioni della qualità del dato in un contesto predittivo, gli strumenti della regressione logistica e i requisiti che essa impone alla base dati. Il secondo capitolo – che costituisce il nucleo centrale della trattazione – documenta la preparazione dei dati variabile per variabile: la ricodifica dei predittori categoriali in variabili dummy, la dicotomizzazione delle variabili di interesse, l'armonizzazione delle codifiche tra campagne diverse e la gestione dei valori mancanti, fino alla composizione di un'Analytical Base Table. Tale attività è stata interamente condotta in SAS Studio. Il terzo capitolo descrive l'impostazione della modellazione: la perimetrazione stagionale della base dati, la costruzione dei campioni di stima e validazione, l'architettura della pipeline sequenziale e le dimensioni di specificazione lungo le quali sono state articolate le diverse configurazioni dei modelli.

La trattazione si circoscrive alla preparazione dei dati e all'impostazione dell'architettura di modellazione, mentre i risultati quantitativi sono discussi dal collega con il quale si è svolta l'attività di tirocinio.