Project: Airline Satisfaction

Path: start2impact University Master Data Science Final Project

Author: Giacomo Abramo

Questo progetto parte da un dataset e la sua analisi non è definita in ogni dettaglio, proprio come un vero progetto di data science. Non ci sono istruzioni precise, ma semplicemente una linea guida che riteniamo essere abbastanza standard e generale per ogni problema. Il tuo compito sarà seguire ogni punto e commentarlo, mostrando di aver capito cosa tu stia facendo. Se alcuni punti si rivelano impossibili, spiegane la motivazione. Troverai probabilmente alcuni modelli o tecniche che non hai propriamente studiato. Non preoccuparti: la sperimentazione, la scoperta e il continuo aggiornamento fanno parte del gioco.

Prima di tutto, dividi il dataset in una parte di training e una parte di test. Inizia ad analizzare il dataset. Presta attenzione a che dataset utilizzare per ogni punto (training, test o tutto?). Tra le varie cose da fare, considera anche: Sono presenti degli outlier? Se sì, che percentuale? Trovi qui (https://machinelearningmastery.com/model-based-outlier-detection-and-removal-in-python/) alcuni modi per l'identificazione degli outlier. Ci sono variabili categoriche? Come pensi di trattarle? Le varie feature potrebbero avere valori molto diversi. Questo non è un bene per quasi qualsiasi algoritmo di ML, quindi valuta se applicare una Standardizzazione al dataset, come per esempio StandardScaler(). Fai attenzione a escludere la colonna delle label! Chiediti poi le seguenti domande: Ci sono delle feature che presentano valori mancanti? Come ti comporti? Commenta le tue analisi in modo chiaro. Controlla se il dataset è bilanciato? Quanto sono correlate le variabili? Ci sono dei casi di multicollinearità (correlation coefficient uguale a 1)? Sono tutte le variabili necessarie o posso selezionarne un sottoinsieme e trascurare le altre? Studia bene la correlation matrix ed effettua poi un test del Chi-square per vedere quali feature sono più importanti. Oltre al Chi-square prova la Mutual Information, il T-test e confronta se le feature selezionate sono le stesse.

Una volta compreso il tipo di task da risolvere (nel nostro caso è una classificazione binaria), uno step importante che determina tutte le valutazioni future è la scelta delle metriche con cui andremo a misurare le performance. La prima è l'accuracy, ma non è l'unica a essere rilevante. Cerca di comprendere quali siano le altre e definiscile una a una. D'ora in poi per valutare la bontà dei modelli userai l'accuracy.

Pensa ora a come approcciare e impostare il problema della soddisfazione dei clienti della compagnia aerea. Scegli almeno tre modelli, tra cui RandomForest e AdaBoost. Di ogni modello scelto leggi la documentazione su sklearn. Ora proviamo due strade diverse. La strada che chiamiamo 1) consiste nell'utilizzare tutte le feature disponibili; mentre la strada che chiamiamo 2) consiste nel selezionare le feature (3 o 4) che si sono dimostrate più rilevanti nel T-test di cui sopra. Per più rilevanti si intendono le feature che hanno un p-value sotto una certa soglia. Esegui dunque la seguente procedura per entrambe le strade, prima con la 1) poi con la 2). Effettua uno spot check (https://machinelearningmastery.com/spot-check-classification-machine-learning-algorithms-python-scikit-learn/) (qui un'altra risorsa) (https://machinelearningmastery.com/spot-check-machine-learning-algorithms-in-python/) sui modelli scelti e selezionare i due migliori modelli usando come metodo di evaluation la k-fold cross validation. Lo spot check non vuole essere un training completo ma giusto una procedura veloce per vedere se esiste una tipologia di modelli che spiccano più degli altri. Effettua

poi su questi due migliori modelli un tuning degli iperparametri, per esempio utilizzando una grid search e sempre la cross validation come metodo di valutazione. Dovresti ora essere nella condizione di conoscere i migliori iperparametri per entrambi i modelli selezionati. Valuta ora i due modelli sul test set, finora inutilizzato. Capirai quale dei due è più performante.

Confronta ora le strade 1) e 2) e otterrai così il miglior modello. Fai una discussione finale sul problema e sulla soluzione trovata.