UNIVERSITÀ
DI PARMA

Dipartimento di Ingegneria e Architettura

Corso di Laurea in Ingegneria Informatica, Elettronica e delle Telecomunicazioni

---

Tecniche di Machine Learning nell'analisi di preventivi di aziende grafico-editoriali

Machine Learning techniques applied to the analysis of graphic-publishing companies' quotes

Relatore:
Prof.ssa Monica Mordonini

Tesi di Laurea di:
Vincenzo Fraello

Correlatore:
Prof. Michele Tomaiuolo

ANNO ACCADEMICO 2020/2021

Il seguente elaborato riassume la tesi derivante dal tirocinio svolto presso l'azienda *Logica s.r.l*: una software house che realizza sistemi informativi e presta consulenza organizzativa e gestionale. Tra i tanti moduli software realizzati dall'azienda è presente un software che consente di realizzare preventivi: *Proto-PREV*. La tesi si focalizza sull'ottimizzazione del processo di preventivazione di prodotti di aziende del settore grafico-editoriale sfruttando i benefici derivanti dall'uso di tecniche di *Machine Learning* (ML). Pertanto, gli obiettivi che la tesi si propone di raggiungere sono: (i) realizzare un sistema che sia in grado di fornire una previsione con un certo intervallo di confidenza circa la commissione del lavoro; (ii) fornire al produttore delle possibili informazioni sulle variabili che rendono competitivo il suo preventivo, per fare eventuali scelte tecnologiche/organizzative che lo portino a colmare eventuali *gap* ed ottimizzare la produzione.

Logica ha diverse tipologie di clienti, tutti in ambito grafico-editoriale, ma con struttura produttiva specifica e di conseguenza con una tipologia di prodotto finito differente. La tipologia di prodotto che è stata analizzata in questo progetto sono le *shopping bag in carta con manici in corda* che vengono prodotte e vendute per il mercato europeo.

La struttura ospitante ha fornito la base di dati contenente le tabelle relative alle diverse parti che compongono un preventivo: (i) *Testata*: contiene i dati generali del lavoro per il quale si realizza il preventivo (es. cliente); (ii) *Prodotti*: contiene i dati di dettaglio del prodotto finito (es. numero di copie); (iii) *Parti*: contiene le componenti del prodotto (es. pagine, copertina); (iv) *Impianti*: contiene informazioni tecniche sul processo di stampa; (v) *Capitolato*: contiene informazioni sui dati produttivi; (vi) *Riepilogo*: contiene i dati per la definizione di un prezzo di vendita. Ogni preventivo può essere richiesto dal cliente in diverse *varianti*, che possono differire tra loro in relazione al numero di copie richiesto o sulla base delle specifiche tecniche del prodotto stesso. Il problema di ottimizzazione è stato affrontato tramite tecniche di apprendimento automatico supervisionato e non supervisionato. Gli strumenti che sono stati utilizzati sono principalmente librerie del linguaggio di programmazione *Python*: *NumPy*, *Pandas*, *Matplotlib*, *Seaborn* e *Scikit-learn*. Un ulteriore strumento è stato lo *Structured Query Language*; fondamentale per l'estrazione dei dati dal database.

Ogni tabella della base di dati è dotata di un numero differente di campi. Quindi, già durante il processo di estrazione dei dati è stato necessario effettuare delle osservazioni che hanno evitato la presenza di dati non utili all'analisi o, ancora, dati che avrebbero potuto alterare o falsare i risultati finali. Alcuni esempi sono: (i) le analisi sono state eseguite sul prodotto in cui l'azienda è maggiormente specializzata poiché, pur considerando il settore in

cui l'azienda risulta essere più performante, il numero di preventivi non approvati è maggiore rispetto a quelli approvati. Tra le N varianti del preventivo, solo una sarà approvata e, le altre N-1 non saranno approvate. Se si prendessero in considerazione settori in cui l'azienda è meno focalizzata si avrebbero classi maggiormente sbilanciate; (ii) i campi sono stati selezionati anche in merito al *problema della granularità delle varianti*. Le varianti dei preventivi, come detto in precedenza, possono variare in relazione a: numero di copie del prodotto finito o caratteristiche tecniche a livello granulare. Durante l'analisi, dunque, si potrebbe incorrere nel problema che le diverse varianti del medesimo preventivo differiscano fra loro per parametri granulari che, se non considerati, potrebbero far in modo che lo stesso preventivo sia contemporaneamente approvato e non approvato. Allo stesso tempo, si ha l'esigenza di produrre un sistema che in futuro si possa adattare alle diverse tipologie di prodotto. Di conseguenza, non si possono considerare caratteristiche troppo specifiche. Dunque, in cooperazione con gli esperti del dominio, sono stati identificati i principali parametri granulari che distinguono le varianti del prodotto finito. Da tale analisi si evince che i parametri in questo contesto siano: la carta, lo spessore della carta, la coprenza e le colorazioni. Inoltre, si è utilizzata un'euristica tale per cui le varianti differiscano maggiormente in relazione al numero di copie; (iii) sono stati considerati solamente i preventivi il cui codice cliente è diverso da quello dell'azienda stessa. Questa considerazione è fondamentale perché l'azienda produce dei prodotti necessari per l'uso interno che non venderà ad altri clienti del settore. Se si considerassero pure questi preventivi, si accrescerebbe il numero di preventivi approvati (i quali non possono mai essere non approvati) il che, altererebbe il risultato finale dell'analisi.

Dopodiché, sono state eseguite le classiche operazioni sui Big-data: (i) analisi dei valori mancanti; (ii) operazioni di standardizzazione e normalizzazione; (iii) analisi univariata, bivariata, multivariata e della correlazione lineare; (iv) trattazione dei valori anomali; (v) codifica delle variabili categoriche; (vi) bilanciamento delle classi; (vii) processo di selezione delle features. Tutte le operazioni precedentemente citate sono state eseguite ciclicamente e in ordine non lineare in relazione a: risultati ottenuti dagli algoritmi di ML, andamenti dei grafici prodotti durante le analisi univariata, bivariata e multivariata.

Gli algoritmi di ML utilizzati sono: (i) Decision Tree Classifier; (ii) K-nearest neighbors; (iii) Categorical Naïve Bayes; (iv) K-means; (v) SVM, GaussianNB, BernoulliNB, MLPClassifier.

Le metriche utilizzate per valutare la bontà degli algoritmi sono: matrice di confusione, accuracy, precision, recall e curva ROC. Gli algoritmi che hanno avuto prestazioni più elevate sono: Decision Tree Classifier, K-nearest neighbors, Categorical Naïve Bayes. La motivazione è stata compresa attraverso la fase di rappresentazione dei dati effettuata a valle del processo di features selection.

Infine, mediante l'algoritmo *t-Distributed Stochastic Neighbor Embedding* (t-SNE) si è giunti al miglioramento dei *gap produttivi*. Grazie alla sua capacità di riduzione delle dimensionalità non lineare è stato possibile rappresentare il dataset in maniera chiara: i preventivi approvati sono nettamente distinguibili da quelli non approvati. Questo ha permesso di comprendere il perché gli algoritmi riescano a percepire il pattern sottostante alla classe minoritaria nonostante lo sbilanciamento delle classi e, inoltre, osservando il grafico a dispersione, si comprende che non solo sono presenti dei preventivi che non vengono mai commissionati, ma si riscontra che tra quelli che vengono commissionati, a volte, qualcosa causa la perdita di potenziali lavori. Di conseguenza, dopo aver individuato i tipi di preventivi maggiormente commissionati e, condotto uno studio sui *gap* in tali contesti, l'azienda può apportare dei miglioramenti mirati che provocano un effettivo accrescimento del processo produttivo.

Il processo di analisi discusso in questa tesi non è stato utile al solo fine di creare un sistema di diagnostica intelligente in grado di fornire predizioni sulle commissioni dei preventivi. Di fatto, è stato possibile comprendere che uno dei parametri decisivi in questo specifico caso studio è il numero di copie del prodotto. Di conseguenza, è stata condotta un'ulteriore analisi sulla distribuzione di frequenza delle copie dei preventivi approvati e non approvati, al fine di poter determinare quali sono i preventivi che vengono commissionati con maggiore e minore assiduità. L'azienda è quindi in grado di intervenire in maniera diretta e mirata sulla produzione.