News ed engagement - Articoli Virali

SEZIONI

Introduzione

Questa relazione ha come obbiettivo l'analisi di un dataset arbitrario da scegliere tra quelli disponibili su Kaggle rispettando alcuni vincoli tra cui:

- Utilizzo di libreriaìe pandas, numpy
- Utilizzo d Boolean/Fancy indexing
- Presenza di grafici
- Lettura e scrittura di file

Individuato il dataset Internet news data with user engagement ci si è posta la seguente domanda:

> Qual'è l'orario migliore per pubblicare un articolo di giornale?

Si procederà inizialmente ad una scrematura del dataset fornito da kaggle per ridurre il più possibile la mole di dati su cui lavorare.
Sarà poi studiato l'andamento delle interazioni sia in gruppo che divise per fonte.
A riflettere la bontà dell'orario di pubblicazione sarà esclusivamente lo user engagement.

Per l'analisi si è tratta forte ispirazione dal notebook Melboune Real Estate Market comprehensive Analysis

Preparazione dataset

notebook: Preparazione Dataset

Si inspeziona il dataset con lo scopo di alleggerirlo il più possibile. Vengono inoltre gestiti i pochi casi di dati mancanti.

1. Scrematura

Già da un primo squardo al dataset si nota che molte delle colonne non saranno necessarie alla nostra analisi, essendo noi interessati alla relazione tra l'orario di pubblicazione e la risposta dei lettori. Il titolo viene mantenuto per identificare facilmente gli articoli e per avere un po' di contesto.
Le colonne *source_id* e *source_name* possono essere rese ridondanti. A questo scopo viene creta una tabella di dettaglio che mappa *source_id* a *source_name*. Esaminando questa tabella ci si rende conto della presenza di una riga erronea, che viene eliminata.

Si aggiunge una colonna rappresentante il numero di articoli presenti per ogni fonte alla tabella di dettaglio. In questo modo è possibile valutare quanto la singola testata giornalistica sia rappresentata nel dataset. ESPN, con 82 articoli, risulta essere è significativamente meno presente rispetto alle altre fonti e potrebbe essere eliminata. Prima di lasciarla andare però ci si riserva di osservarene l'engagement.

2. Gestione dei null

Si traccia una heatmap che mette in evidenza la distribuzione dei valori nulli. Fortunatamente appartengono tutti agli stessi tre articoli, che vengono perciò scartati.

Analisi dei dati

notebook: Analisi Dati

Distribuzione interazioni

La distribuzione delle interazioni su tutti gli articoli, non divisi per testata, risulta pesantemente inclinata verso lo zero: un numero ridotto di articoli conta la vasta maggioranza delle interazioni di ogni tipo. Definiamo questi articoli "virali".
Risulta inoltre evidente che la colonna *comment_plugin* presenta un numero di interazioni molto ridotto rispetto alle altre; si procede quindi a contare gli articoli con engagement non nullo, per ciascun tipo di interazione. Con solo 50 aricoli che presentano un'interazione ed un massimo di 15 interazioni per articolo si può impotizzare che *comment_plugin* sia irrilevante per la nostra indagine.\

Studiando la distribuzione delle interazioni divise per fonte, si conferma l'ipotesi fatta in precedenza in quanto la maggior parte delle testate non hanno engagement per *comment_plugin*, che viene quindi scartata.
In questa fase si nota anche che la testata ESPN, oltre ad essere sottorappresentata per numero di articoli, non presenta interazioni in nessuna categoria. Si sceglie quindi di eliminarla dal dataset.

Si realizza quindi un istogramma raffigurante la distribuzione delle interazioni per fonte e per tipo, mettendo in evidenza l'articolo "migliore", ovvero quello con engagement maggiore. Guardando questi grafici si ipotizza che la grande differenza di interazioni tra l'articolo migliore ed il resto degli articoli sia legata ad una notizia o avvenimento di particolare interesse.
Se così fosse introdurremmo un bias nella nostra indagine tenendo questi articoli in quanto è ragionevole ipotizzare che questi siano pubblicati il prima possibile.

Viene quindi stilata una lista dei titoli ed orari di pubblicazione degli articoli migliori per ogni fonte, divisi per tipo di interazione.
Da questa lista si scopre l'articolo migliore di ogni testata riguarda una notizia diversa rispetto a quello delle altre. Anzi, l'articolo migliore per una testata cambia spesso a seconda del tipo di interazione.