CdLM Data Science

Università degli studi di Milano-Bicocca

Corso di Data Science Lab

Anno Accademico 2022-2023

Campaign Click

Emanuela Elli (892901)

Armando Epifani (826153)

Gloria Giorgetti (826226)

Francesco Gregori (889206)

Indice

Introduzione

Tecniche di Machine Learning applicate alle informazioni su vari utenti possono permettere a un'azienda di aumentare l'efficacia delle proprie campagne pubblicitarie su internet, stabilendo le caratteristiche di coloro che più saranno interessati al prodotto mostrato nell'inserzione. Nel seguente report verrà quindi mostrato come, partendo da un dataset di informazioni su vari utenti raccolte tramite cookie, può essere creato un modello di classificazione che predica se un utente cliccherà sull'inserzione o meno e la confidenza sulla predizione stessa. L'obiettivo finale dell'analisi è stabilire a quali utenti inviare l'inserzione, in modo che essa sia vista dalla maggior parte degli utenti che potrebbero risultare interessati e quindi cliccare, anche a discapito di doverla inviare a molti utenti che non lo saranno.

Nel dataset utilizzato viene riportato, oltre al numero di volte in cui l'utente ha cliccato sull'inserzione, il numero di volte in cui l'inserzione è stata mostrata all'utente. Un diverso tipo di analisi avrebbe potuto quindi portare alla creazione di un modello di regressione in grado di stabilire, date le informazioni dell'utente, il rapporto tra il numero click fatti sul numero di volte in cui l'inserzione è stata vista. Dal rapporto sarebbe stato possibile stabilire il grado di interesse dell'utente verso il prodotto pubblicizzato e mostrare quindi l'inserzione agli utenti più interessati.
Si è scelto tuttavia di addestrare un modello che stabilisca se un utente cliccherà o meno, senza considerare il numero di volte in cui lo farà o quante volte l'inserzione dovrà essere mostrata perché ciò avvenga. Questa scelta è stata dettata dalla scarsità di utenti con numero di click diverso da zero all'interno del dataset (circa 0.3%), che avrebbe reso meno complesso l'addestramento di un modello di classificazione rispetto a un modello di regressione. L'inserzione verrà pertanto mostrata a utenti con alta confidenza associata alla predizione del click e non in base al grado di interesse mostrato dall'utente.

In aggiunta al modello di classificazione è stata svolta una analisi sugli utenti considerati sospette frodi. Nel dataset sono infatti presenti utenti segnalati come sospetti da un motore dedicato: ci si aspetta che tali utenti mostrino una diversa distribuzione del rapporto del numero di click sul numero di visualizzazioni dell'inserzione rispetto agli utenti non sospetti. Sono quindi state realizzate delle analisi che hanno permesso di stabilire se le due distribuzioni coincidessero o differissero.

Preparazione dei dati

Il dataset mostra 1416 variabili per ogni utente. Esse contengono, oltre all'informazione su quante volte l'utente ha cliccato, quante volte ha visto l'inserzione e se l'utente è considerato sospetto, anche informazioni sul tipo di sistema operativo, browser e dispositivo utilizzati e se l'utente ha comprato il prodotto pubblicizzato o meno. Sono inoltre presenti informazioni sulle pagine visitate dall'utente e che contenevano l'inserzione pubblicitaria, come la percentuale di pagine la cui lunghezza del testo cade in dati intervalli, la percentuale di pagine visitate in dati momenti della giornata, la percentuale di pagine che trattano di uno specifico argomento e la percentuale di pagine che trasmettono un certo sentimento o emozione.

Nel dataset sono presenti 82654 utenti, 273 dei quali mostrano un numero di click fatti diverso da zero. Dal dataset sono stati rimossi gli utenti sospetti e gli utenti con numero di visualizzazioni dell'inserzione pari a zero, ovvero quelli a cui la pubblicità è arrivata per e-mail. Gli utenti mantenuti sono quindi quelli non sospetti e che hanno ricevuto la pubblicità tramite web, in modo da non inficiare sul risultato dell'analisi: gli utenti sospetti potrebbero essere bot il cui comportamento dipende da come sono stati programmati e non da un reale interesse; invece, per gli utenti la cui pubblicità è arrivata tramite e-mail non si può supporre mostreranno lo stesso comportamento anche in caso vedano l'inserzione in una pagina web. Sono stati inoltre rimossi circa 20 mila utenti che presentano valori mancanti per la lunghezza dei testi delle pagine visitate e per le categorie degli argomenti trattati nelle stesse.

Si è quindi passati a stabilire quali variabili mantenere per l'analisi e quali non considerare.
Per prima cosa si è deciso di rimuovere tutte quelle variabili che mostrano valore nullo per ogni utente, ovvero la colonna che indica se l'utente ha comprato il prodotto pubblicizzato e quelle relative al sentimento espresso nelle pagine visualizzate dall'utente. Anche le colonne relative alle emozioni trasmesse dalle pagine visitate sono state rimosse: pur non mostrando valore nullo per ogni utente e per ogni variabile, gli utenti che presentano valori mancanti sono più di due terzi del totale (71.5%).

Le colonne relative al momento della giornata in cui le pagine sono state visualizzate sono in tutto 15 e si dividono in due gruppi: time1 e time2. In time1 vengono mostrate le percentuali delle pagine visualizzate rispettivamente di mattina, pomeriggio, sera e notte sia dei giorni lavorativi che del fine settimana; in time2 invece vengono mostrate le percentuali delle pagine visualizzate rispettivamente di mattina, di mattina presto, a mezzogiorno, nel pomeriggio, durante la sera, di notte e durante le ore normalmente dedicate al sonno.
I due gruppi di colonne presentano quindi lo stesso tipo di informazione rappresentata in modo diverso. Le colonne del gruppo time2 mostrano, tuttavia, informazioni mancanti per circa 3 mila utenti; si è pertanto deciso di mantenere solamente le variabili relative a time1, rimuovendo quelle di time2.

Le colonne che stabiliscono l'argomento delle pagine visualizzate si dividono in quattro gruppi:

- categories1, formato da 26 argomenti;
- categories2, formato da 359 argomenti;
- categories3, formato da 827 argomenti;
- admants, formato da 44 argomenti.

Per ogni argomento in ogni gruppo è mostrata la percentuale di pagine su quell'argomento visitate dall'utente, inoltre la somma delle percentuali in ogni gruppo è 100%.

I gruppi categories1, categories2 e categories3, per come sono definiti, riportano la medesima informazione a un diverso livello di aggregazione: ogni argomento di categories1 è suddiviso in più argomenti di categories2 e ogni argomento di categories2 è suddiviso in più argomenti di categories3. Per questo motivo sono state mantenute solamente le colonne del gruppo categories2, in modo da avere degli argomenti più specifici di quelli contenuti in categories1 e non avere un alto numero di colonne con pochi valori considerevolmente diversi da zero, come sarebbe avvenuto mantenendo gli argomenti del gruppo categories3.

Per il gruppo admants, quasi due terzi delle colonne associate agli argomenti mostrano una correlazione superiore a 0.7 con almeno una colonna di categories1. Le informazioni riportate nel gruppo admants sono quindi in larga parte presenti nelle informazioni del gruppo categories1 e di conseguenza nel gruppo categories2. Per questo motivo le colonne del gruppo admants non sono state considerate nell'analisi.

Come ultima operazione è stata aggiunta una variabile che assume valore "clicker" se l'utente ha cliccato sull'inserzione almeno una volta oppure "non clicker" se l'utente non ha mai cliccato sull'inserzione. Vengono quindi rimosse le colonne con l'informazione sul numero di click fatti e il numero di volte che l'utente ha visto l'inserzione.

Analisi descrittiva

Il dataset processato è formato da 62379 utenti e 395 variabili, contenenti le seguenti informazioni:

- Sistema operativo (7 variabili);
- Browser (10 variabili);
- Tipo di dispositivo (una variabile);
- Momento della giornata in cui sono state visitate le pagine (8 variabili);
- Lunghezza delle pagine visitate (9 variabili);
- Argomento delle pagine visitate (359 variabili);
- Variabile che indica se l'utente ha cliccato o no (una variabile).

Le variabili legate al sistema operativo indicano se un utente ha utilizzato uno specifico sistema operativo per visualizzare la pagina con l'inserzione, oppure no. Vengono considerati 7 sistemi operativi diversi, tra i quali il più utilizzato risulta essere "Windows", come mostra la figura 1. Nel primo grafico è riportato per ogni sistema operativo la frazione di utenti che cliccano rispetto al numero totale di utenti che cliccano; nel secondo, è riportata la stessa informazione per gli utenti che non cliccano: tra le due distribuzioni non si osservano significative differenze a livello di andamento.

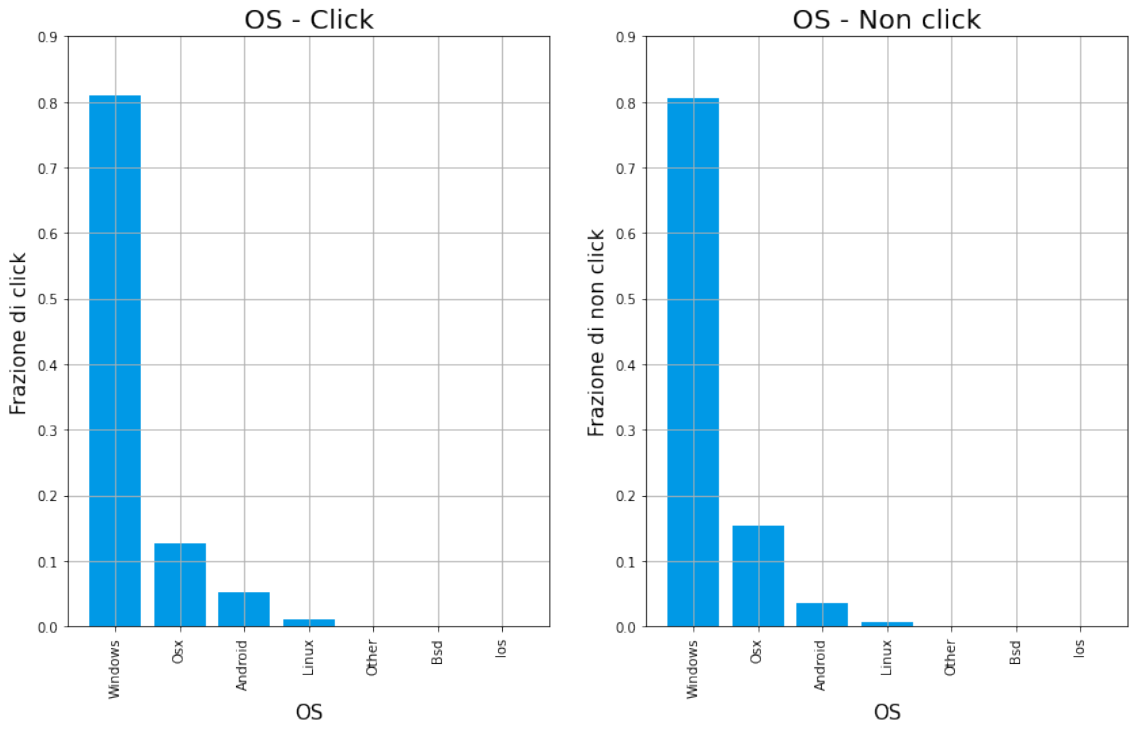

Figura 1: Frazione di utenti che cliccano e di utenti che non cliccano nelle variabili relative ai sistemi operativi. La frazione è ottenuta come rapporto tra il numero di utenti che cliccano (o di utenti che non cliccano) per la data variabile e il numero totale di utenti che cliccano (o di utenti che non cliccano).

Le variabili legate al browser indicano se un utente ha utilizzato uno dei 10 browser considerati per visualizzare la pagina con l'inserzione, oppure no. I due grafici della figura 2 mostrano le distribuzioni per gli utenti che cliccano e per quelli che non cliccano: si osservano delle differenze

nei valori, soprattutto tra “Chrome”, che assume valore pari a 0.58 per gli utenti che cliccano e a 0.65 per quelli che non cliccano, e “Edge”, che assume valori 0.24 e 0.16 rispettivamente.

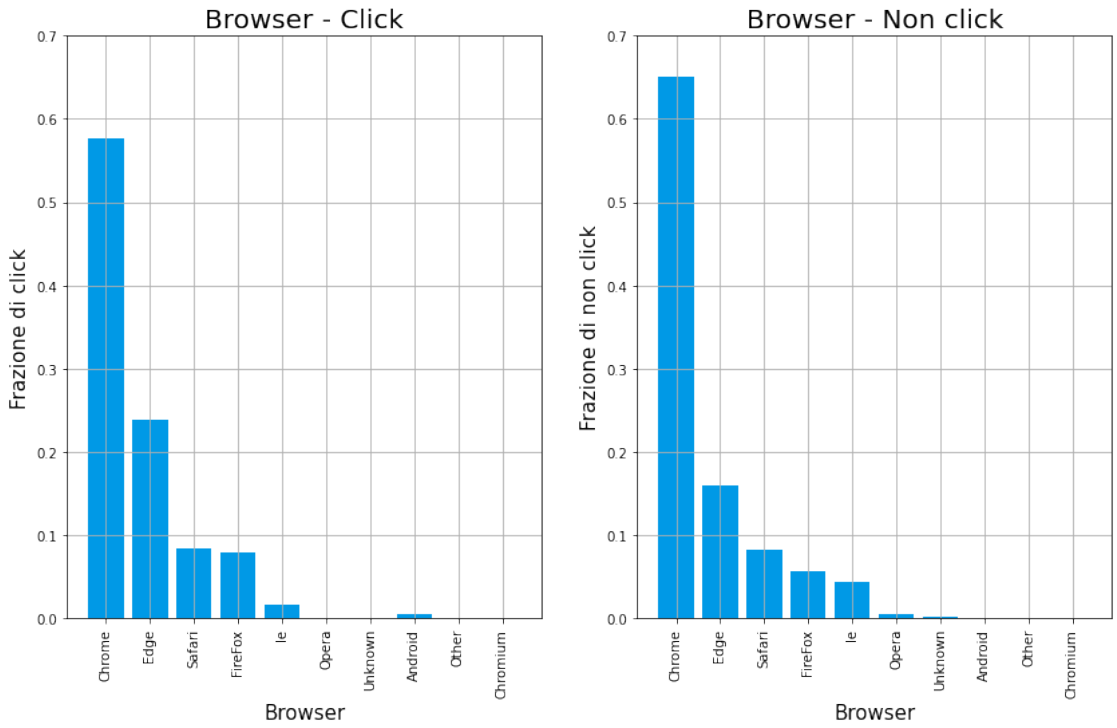

Figura 2: Frazione di utenti che cliccano e di utenti che non cliccano nelle variabili relative ai browser. La frazione è ottenuta come rapporto tra il numero di utenti che cliccano (o di utenti che non cliccano) per la data variabile e il numero totale di utenti che cliccano (o di utenti che non cliccano).

La variabile che contiene le informazioni relative al tipo di dispositivo può assumere tre valori, legati rispettivamente ai tipi di dispositivo considerati. Nei due grafici della figura 3 è mostrata la frazione degli utenti che cliccano e di quelli che non cliccano rispetto alle modalità della variabile: in entrambe le distribuzioni si può notare una netta prevalenza di utenti che utilizzano come dispositivo “Desktop e laptop”.

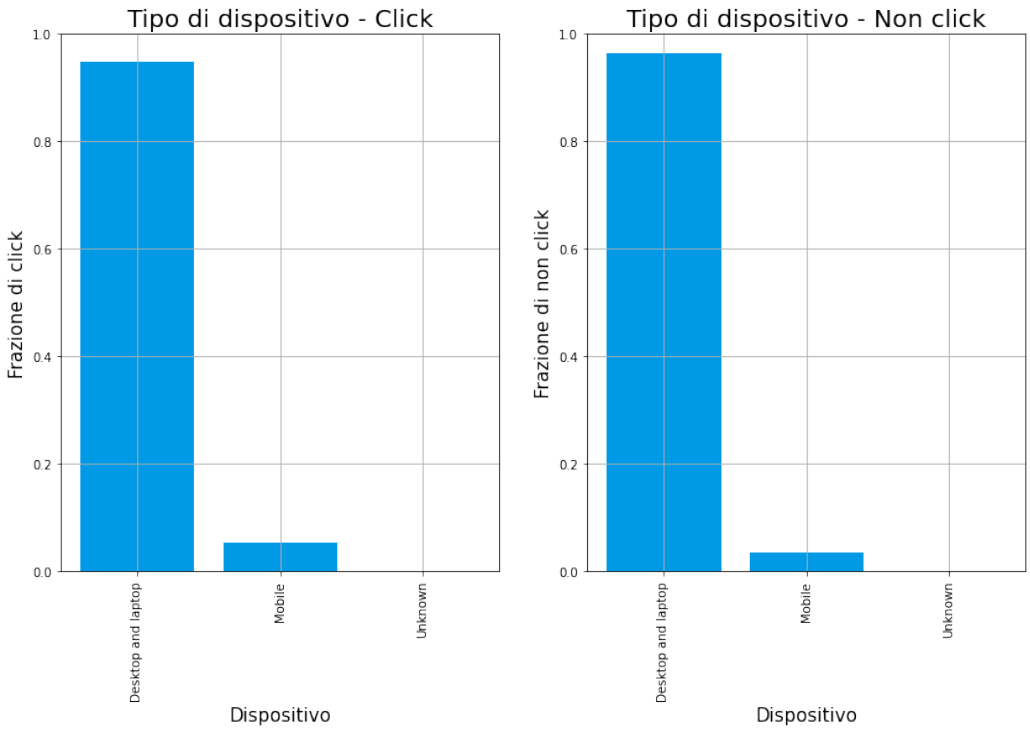

Figura 3: Frazione di utenti che cliccano e di utenti che non cliccano nelle modalità della variabile legata al tipo di dispositivo. La frazione è ottenuta come rapporto tra il numero di utenti che cliccano (o di utenti che non cliccano) per la data variabile e il numero totale di utenti che cliccano (o di utenti che non cliccano).

Come precedentemente esposto, sono presenti gruppi di variabili relative al momento della giornata in cui sono state visitate le pagine, alla lunghezza delle pagine visitate e all'argomento delle pagine visitate. Ogni gruppo è formato da colonne, i cui valori sono percentuali che sommano a 100 per ogni utente. Per analizzare queste variabili si è quindi considerato il valore medio di ognuna, separando gli utenti che cliccano da quelli che non cliccano.

Per le variabili legate al tempo si è scelto di raggruppare assieme le quattro legate ai giorni infrasettimanali e le quattro legate al weekend, in modo da poter analizzare meglio le differenze nelle distribuzioni nei due diversi momenti della settimana.
Nei giorni infrasettimanali, gli utenti che cliccano, tendono a visitare maggiormente le pagine in cui è contenuta l'inserzione nel pomeriggio o alla sera, come mostrato nei primi due grafici di figura 4. Si può anche notare come le medie legate alla mattina e alla sera non siano simili nelle due distribuzioni.
Nel weekend invece, gli utenti che cliccano tendono a visitare maggiormente i siti solo nel pomeriggio e non si osservano grosse differenze tra questi e gli utenti che non cliccano, come mostrato nel terzo e quarto grafico della figura 4.
Si osserva invece una grande differenza se si confrontano le distribuzioni legate ai diversi momenti della settimana: ad eccezione della mattina, le medie legate agli altri momenti della giornata sono più alte nei giorni infrasettimanali, piuttosto che nel weekend.

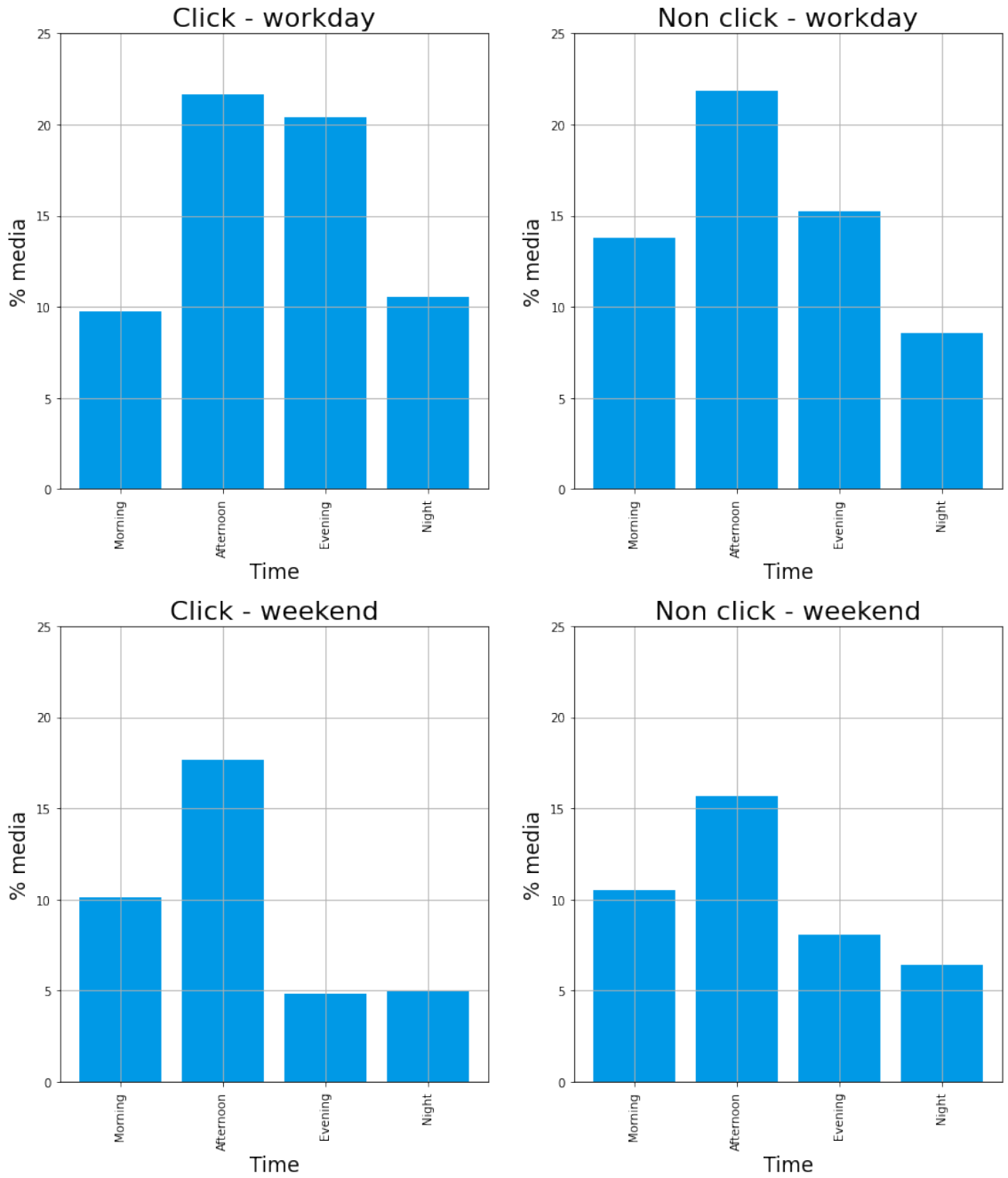

Figura 4: Media delle variabili legate al momento della giornata. I due grafici in alto raffigurano rispettivamente la media per gli utenti che cliccano e per quelli che non cliccano nei giorni infrasettimanali; il terzo e il quarto grafico raffigurano la stessa informazione relativamente al weekend.

Le variabili legate alla lunghezza mostrano delle differenze nelle due distribuzioni, come evidenziato dai grafici in figura 5: gli utenti che cliccano tendono a visualizzare maggiormente pagine dalla lunghezza compresa tra i 51 e i 100 caratteri, mentre per gli utenti che non cliccano si osserva un picco in corrispondenza di pagine dalla lunghezza minore di 50 caratteri.

Infine, visto il gran numero di variabili relative all'argomento, si è scelto di tenere in considerazione per l'analisi solo quelle con la media più alta. Nei grafici di figura 6 sono mostrate le due distribuzioni: si può notare come l'andamento generale sia circa simile, con l'eccezione della categoria "games", che mostra una media sensibilmente più alta per gli utenti che cliccano.

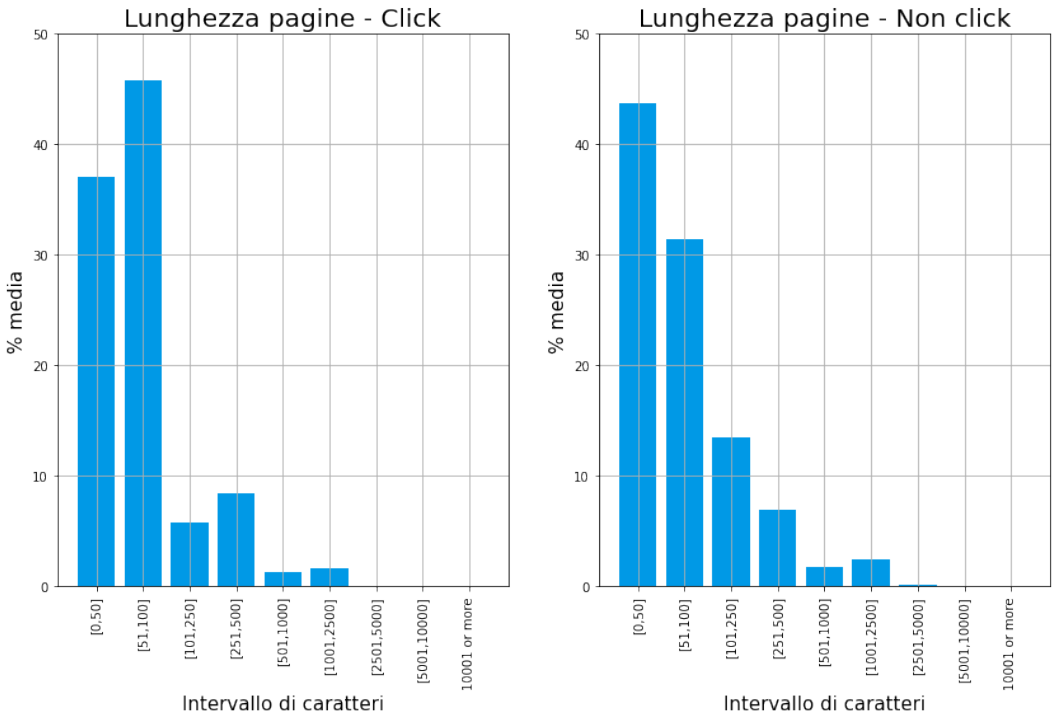

Figura 5: Media delle variabili legate alla lunghezza delle pagine visitate per utenti che cliccano e per utenti che non cliccano.

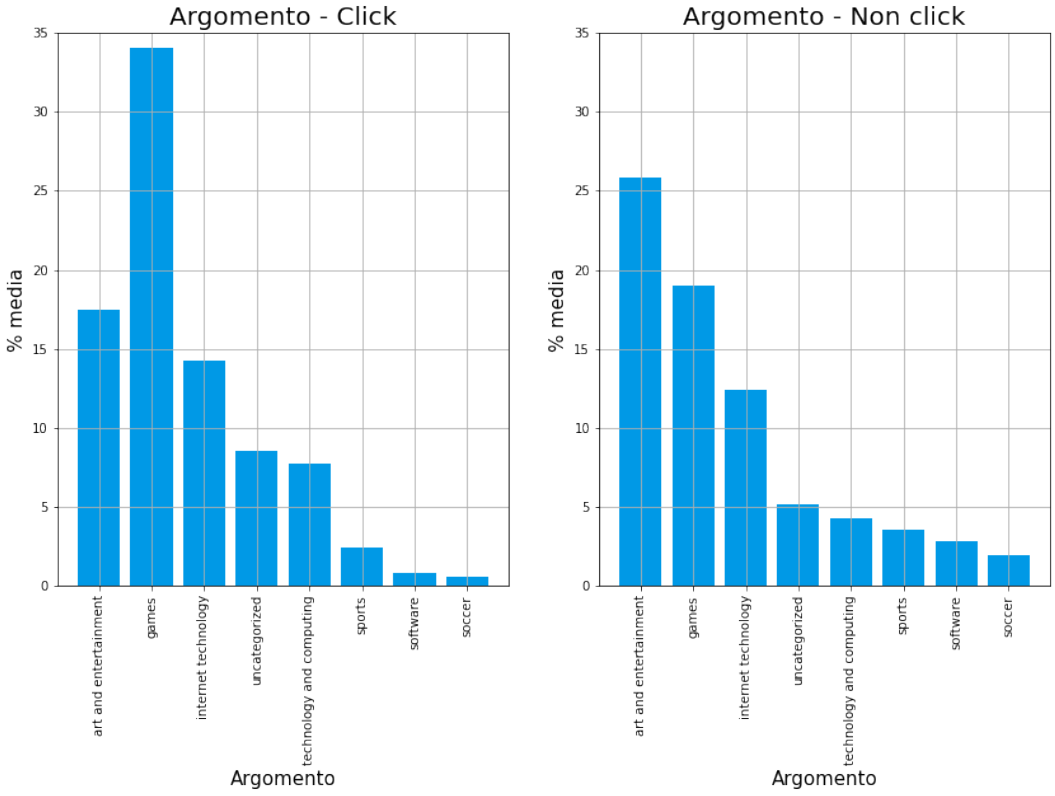

Figura 6: Media delle variabili legate all'argomento delle pagine visitate per utenti che cliccano e per utenti che non cliccano.

Analisi degli utenti sospetti

All’interno del dataset sono stati rilevati 152 utenti identificati come sospetti di frode, su 82654 utenti totali. Si ha pertanto una percentuale di utenti sospetti di frode pari a 0.18%, rispetto al totale del numero di utenti nel dataset, mentre la restante parte è relativa agli utenti non sospetti.

Per ogni utente è stato calcolato, all’interno di una nuova variabile, il rapporto tra il numero di click effettuati rispetto al numero di volte in cui è stata vista l’inserzione. In questo modo è stato possibile verificare se le distribuzioni di tali rapporti fossero identiche o differenti a livello statistico all’interno dei due gruppi. Per compiere questo tipo di analisi, è stato effettuato il test U di Mann-Whitney, che permette di verificare l’ipotesi nulla per cui due distribuzioni sono identiche. Il p-value ottenuto dal test, pari a 0.41, è significativamente maggiore di 0.05 e ne consegue che le medie dei due gruppi analizzati non sono significativamente differenti. Il test stabilisce quindi che non c’è evidenza di una differenza significativa tra le due distribuzioni, ovvero la distribuzione di utenti sospetti e non sospetti si può considerare statisticamente identica.

A seguito dei risultati del test, per verificare ulteriormente il funzionamento del motore, sono stati analizzati i conteggi degli utenti sospetti per le diverse tipologie di dispositivo, sistema operativo e browser. Questa analisi è stata svolta per capire se gli utenti sospetti di frode presenti all’interno del dataset seguissero un pattern, ovvero se il motore stabilisse il sospetto sulla base del software o dell’hardware utilizzati da un utente.

Per la tipologia di sistema operativo utilizzato si nota che un alto numero di utenti sospetti di frode utilizza sistemi operativi poco conosciuti, categorizzati come “other”, a differenza di quanto si osserva per gli utenti non sospetti. Risulta inoltre evidente che il numero degli utenti sospetti all’interno della categoria “other" è più elevato di quelli non sospetti, come mostra la figura 7.

Per quanto riguarda le tipologie di dispositivo si può notare che il maggior numero di utenti sospetti tende ad utilizzare maggiormente dispositivi sconosciuti, identificati come “unknown”, come mostra la figura 8. Si può notare la differenza con gli utenti non sospetti, i quali tendono invece a utilizzare maggiormente dispositivi noti, ovvero “desktop e laptop” e “mobile”.

Per i browser utilizzati, come mostra la figura 9, il numero di utenti sospetti risulta distribuito equamente tra le variabili, a differenza di quanto si osserva per gli utenti non sospetti che utilizzano maggiormente “Chrome”. Di conseguenza non ci sono evidenze grafiche tali per cui sia possibile ipotizzare che uno specifico browser sia correlato all’identificazione del sospetto da parte del motore.

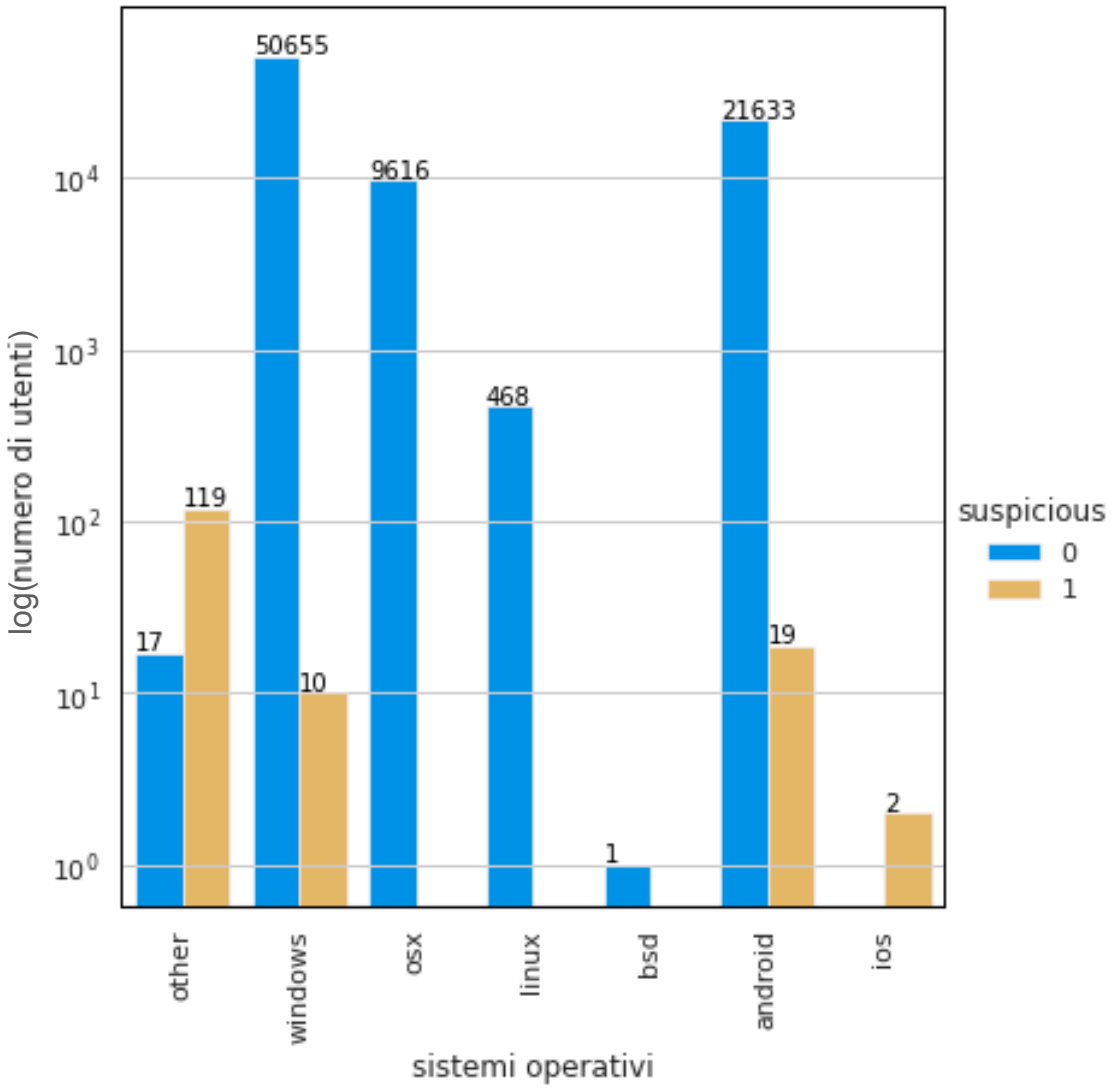

Figura 7: Grafico in scala logaritmica del numero di utenti sospetti (suspicious pari a 0) e non (suspicious pari a 1) nelle variabili relative ai sistemi operativi.

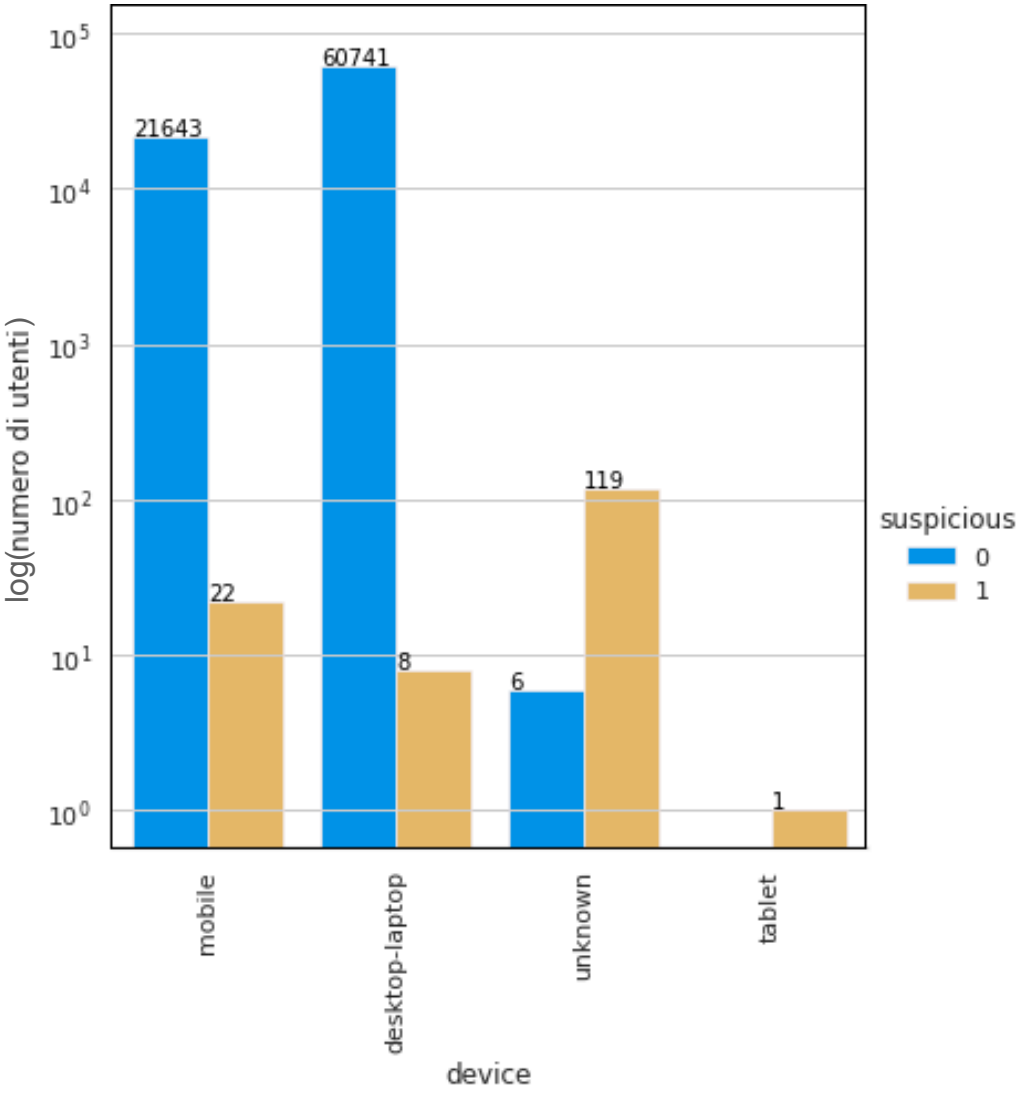

Figura 8: Grafico in scala logaritmica del numero di utenti sospetti (suspicious pari a 0) e non (suspicious pari a 1) nelle modalità della variabile legata al tipo di dispositivo.

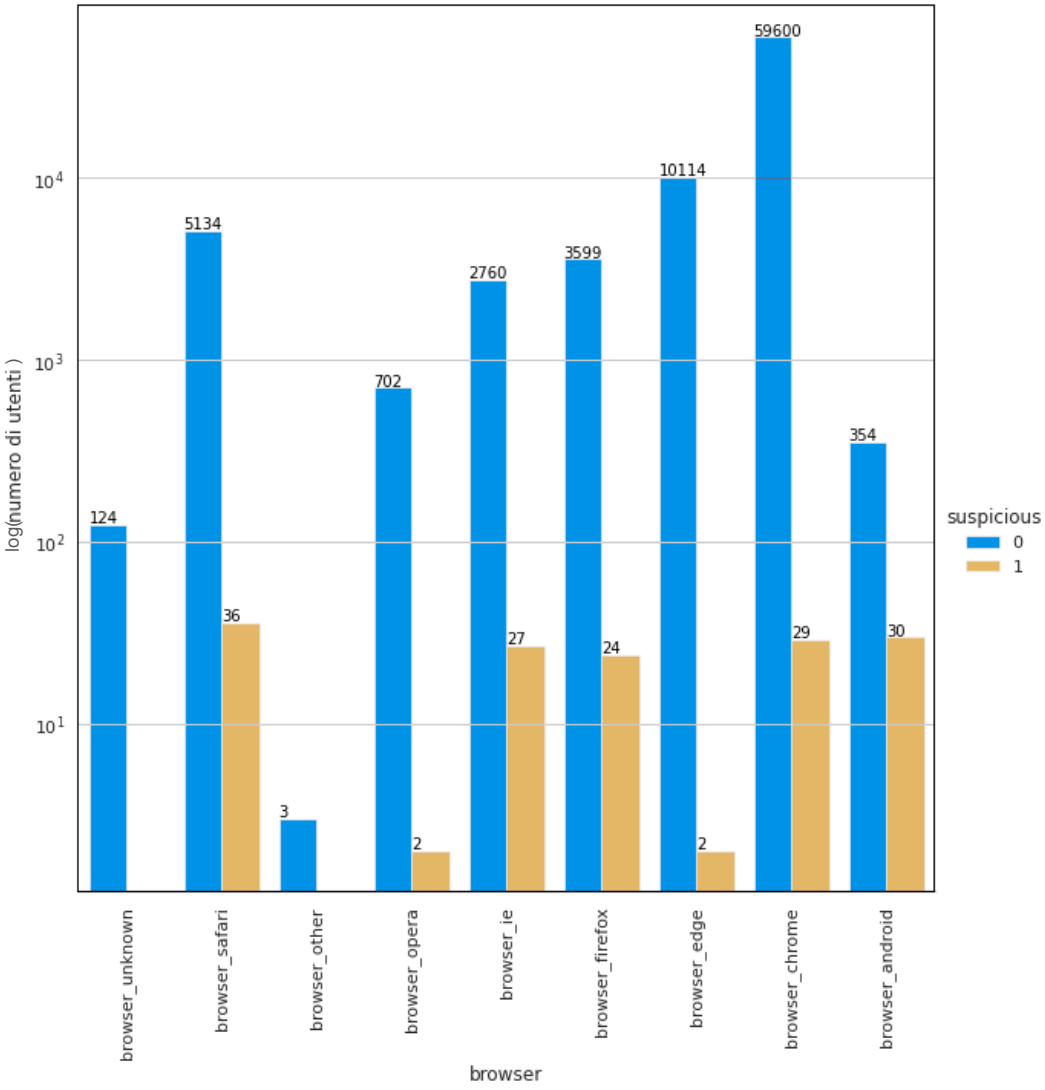

Figura 9: Grafico in scala logaritmica del numero di utenti sospetti (suspicious pari a 0) e non (suspicious pari a 1) nelle variabili relative al browser.

Modello

Il modello sviluppato è un modello di classificazione binaria in grado di stabilire, dalle informazioni degli utenti, quali tra questi cliccheranno sull'inserzione, quindi appartengono alla classe "clicker", e quali no, quindi appartengono alla classe "non clicker". Il modello fornisce inoltre per ogni utente la confidenza che questo appartenga alla classe "clicker".

Il modello è stato appreso grazie all'impiego di un classificatore Support Vector Machine (SVM) con kernel lineare. Per l'addestramento è stato utilizzato un training set formato dal 70% dei dati, mentre il restante 30% del dataset è stato utilizzato per valutare le performance del modello nell'assegnazione della confidenza e stabilire come avverrà la classificazione.
Date le caratteristiche del dataset, anche il training set risulta sbilanciato (circa lo 0.3% degli utenti appartiene alla classe "clicker"): per questo motivo è stato applicato un algoritmo di Random Oversampling ai soli dati di training, in modo da avere uno stesso numero di istanze per entrambe le classi. L'addestramento di un modello con dati significativamente sbilanciati avrebbe infatti portato a classificare tutte le istanze con la classe maggiormente rappresentata, ovvero quella degli utenti che non cliccano. Al contrario, l'obiettivo finale di questa analisi è identificare il maggior numero di utenti che cliccheranno, anche a costo di classificare erroneamente gli utenti che non lo faranno.

Valutazione delle performance del modello

Il funzionamento del modello è stato valutato sulla base dei valori di confidenza sulla predizione della classe "clicker" assegnati agli utenti del dataset non impiegato in fase di addestramento: analogamente al set usato per l'apprendimento, anche questo è sbilanciato e il numero di utenti che cliccano è pari a 53 su 18714 (0.3% circa). La confidenza sulla predizione della classe "clicker" viene assegnata dal modello tramite una decision function: i valori di questa funzione rappresentano la distanza di ogni predizione dall'iperpiano che separa la regione degli utenti classificati come "clicker" da quella degli utenti classificati come "non clicker" e cadono nell'intervallo [-1, 1]. Una volta mappati i valori di confidenza nell'intervallo [0, 1], gli utenti sono stati suddivisi prima in due, poi in quattro e infine in otto gruppi sulla base rispettivamente della mediana, dei quartili e degli ottili della distribuzione dei valori di confidenza. Per ognuno dei gruppi è stato valutato il numero di utenti che realmente cliccano che vi sono contenuti: i risultati sono mostrati nelle figure 10, 11, e 12. In tutti i grafici, il gruppo associato ai valori di confidenza più alti è sempre quello più popolato. Si può inoltre notare, nei gruppi costruiti a partire dalla mediana e dai quartili, un andamento nel numero di utenti che realmente cliccano decrescente al diminuire della confidenza dei gruppi. Questo andamento non è osservabile all'interno dei gruppi costruiti a partire dagli ottili, per i quali si ha comunque che i quattro associati a valori di confidenza maggiori sono più popolati dei quattro associati a valori di confidenza minori. Da queste osservazioni si può concludere che il modello funziona correttamente: pur essendoci utenti che realmente cliccano a cui sono assegnati valori bassi di confidenza, alla maggior parte di essi vengono associati valori alti.

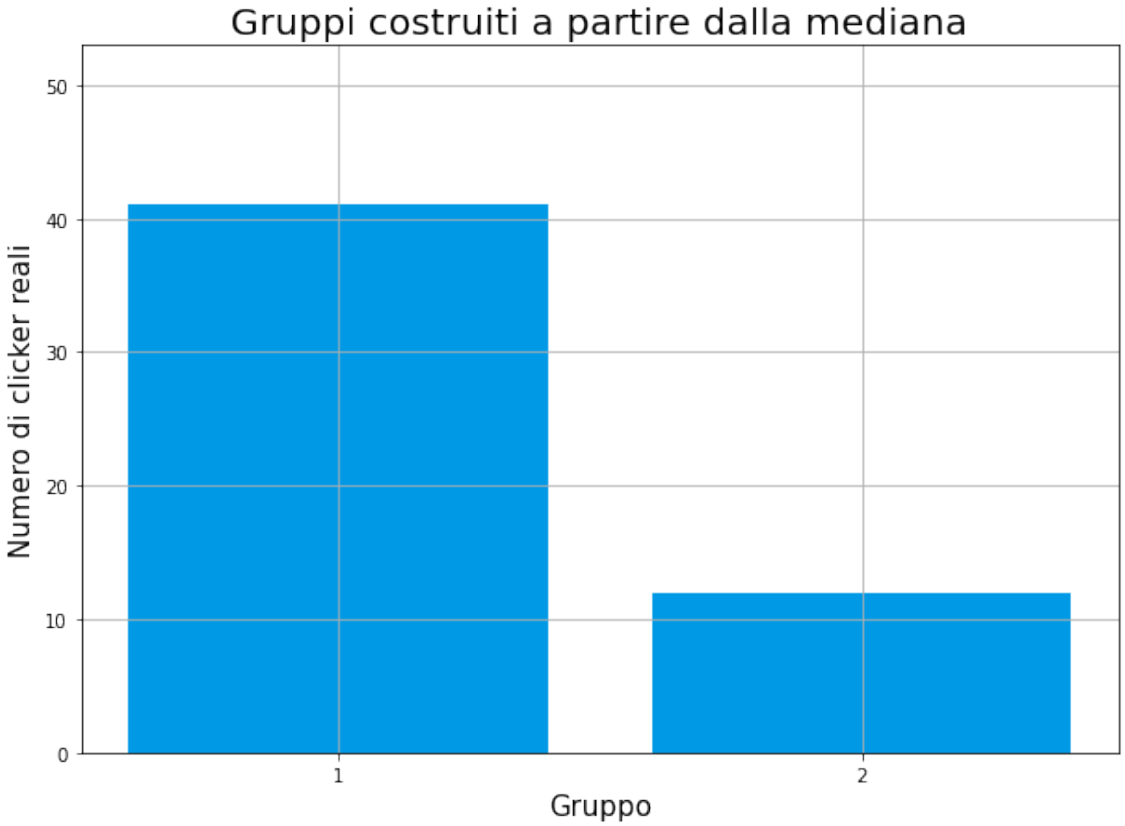

Figura 10: Numero di utenti che realmente cliccano nei gruppi costruiti a partire dalla mediana della confidenza ($q_{0.5}$= 0.46).

Figura 11: Numero di utenti che realmente cliccano nei gruppi costruiti a partire dai quartili della confidenza ($q_{0.75}$ = 0.55, $q_{0.5}$ = 0.46 e $q_{0.25}$ = 0.35).

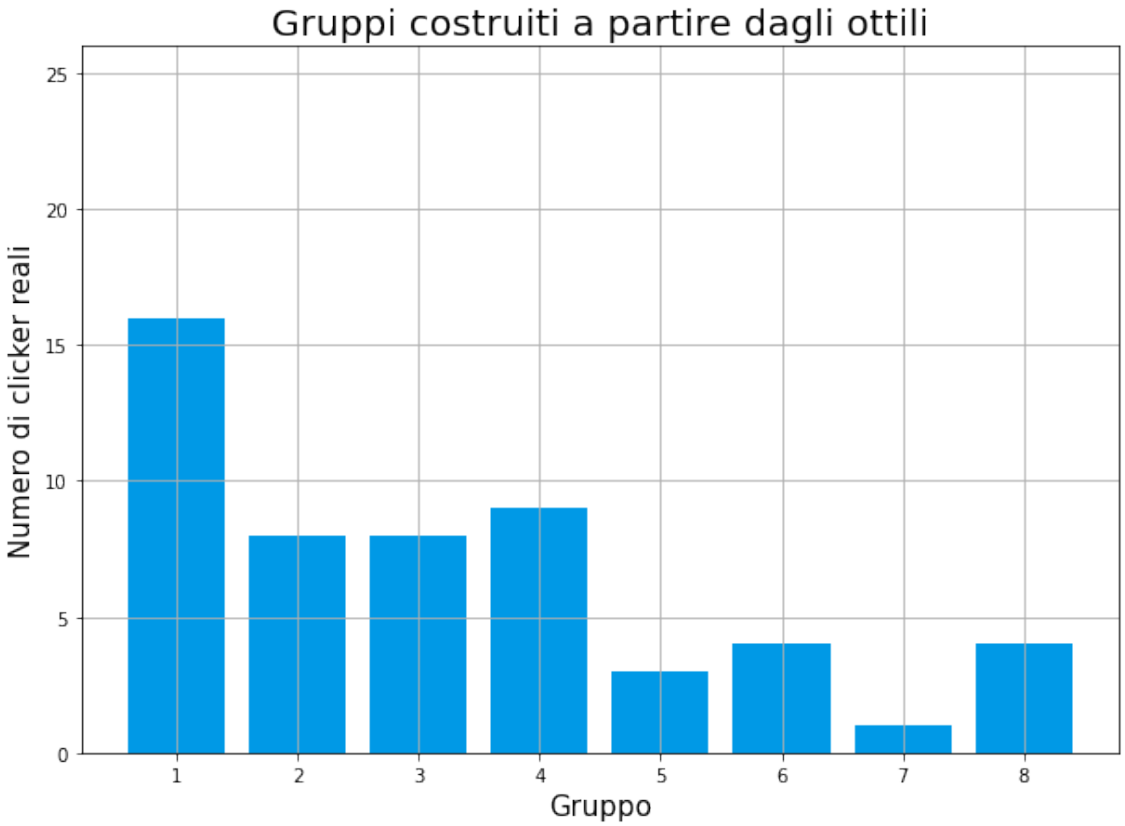

Figura 12: Numero di utenti che realmente cliccano nei gruppi costruiti a partire dagli ottili della confidenza ($q_{0.875}$ = 0.57, $q_{0.75}$ = 0.55, $q_{0.625}$ = 0.50, $q_{0.5}$ = 0.46, $q_{0.375}$ = 0.42, $q_{0.25}$ = = 0.35, $q_{0.125}$ = 0.20).

Classificazione e performance

Il modello addestrato, come discusso precedentemente, associa a ogni utente la confidenza che questo appartenga alla classe “clicker”. Questa confidenza può essere utilizzata per costruire la curva ROC del modello, mostrata in figura 13: variando la soglia della confidenza oltre la quale viene assegnata agli utenti la classe “clicker”, viene mappata la frazione dei “clicker” correttamente identificati in funzione della frazione di “non clicker” predetti erroneamente.

Nella curva ROC in figura 13, costruita a partire dai dati non utilizzati in fase di addestramento, il punto con coordinate rispettivamente di 0.38 e 0.64 è associato a una soglia per la confidenza pari a 0.5. Un modello di classificazione utilizza di default tale valore di soglia, in caso di confidenze distribuite tra 0 e 1, per restituire la predizione della classe di appartenenza di ogni utente. Scegliendo quindi la soglia di default, il modello sarebbe in grado di identificare correttamente il 64% degli utenti appartenenti alla classe “clicker” e il 62% degli utenti appartenenti alla classe “non clicker”.

Scopo dell’analisi è quello di identificare correttamente un alto numero di “clicker”, anche a costo di classificare erroneamente alcuni dei “non clicker”. La soglia di default non è pertanto stata valutata sufficiente per adempiere a tale scopo e si è deciso di abbassarla a 0.464: gli utenti con una confidenza di appartenere alla classe “clicker” superiore a questa soglia verranno classificati dal modello come “clicker”, viceversa gli utenti con una confidenza inferiore saranno classificati come “non clicker”. La soglia è stata scelta sfruttando le informazioni presenti nella curva ROC,

individuando la soglia che permettesse di avere meno della metà degli utenti della classe “non clicker” erroneamente classificati, pur portando a classificare correttamente un buon numero di utenti della classe “clicker”. Grazie a questa nuova soglia vengono classificati correttamente il 77% degli utenti della classe “clicker” e il 52% degli utenti della classe “non clicker”.
Basandosi sui risultati di questa classificazione, un’azienda sarà in grado di mostrare l’inserzione a meno di metà degli utenti potenzialmente non interessati e più di tre quarti degli utenti potenzialmente interessati. Poiché la maggior parte degli utenti appartiene alla classe “non clicker”, l’azienda riuscirà quindi a scartare circa la metà di tutti i possibili utenti e non mostrare loro l’inserzione. L’azienda dovrà tuttavia accettare che l’inserzione non sarà vista da tutti gli utenti potenzialmente interessati.

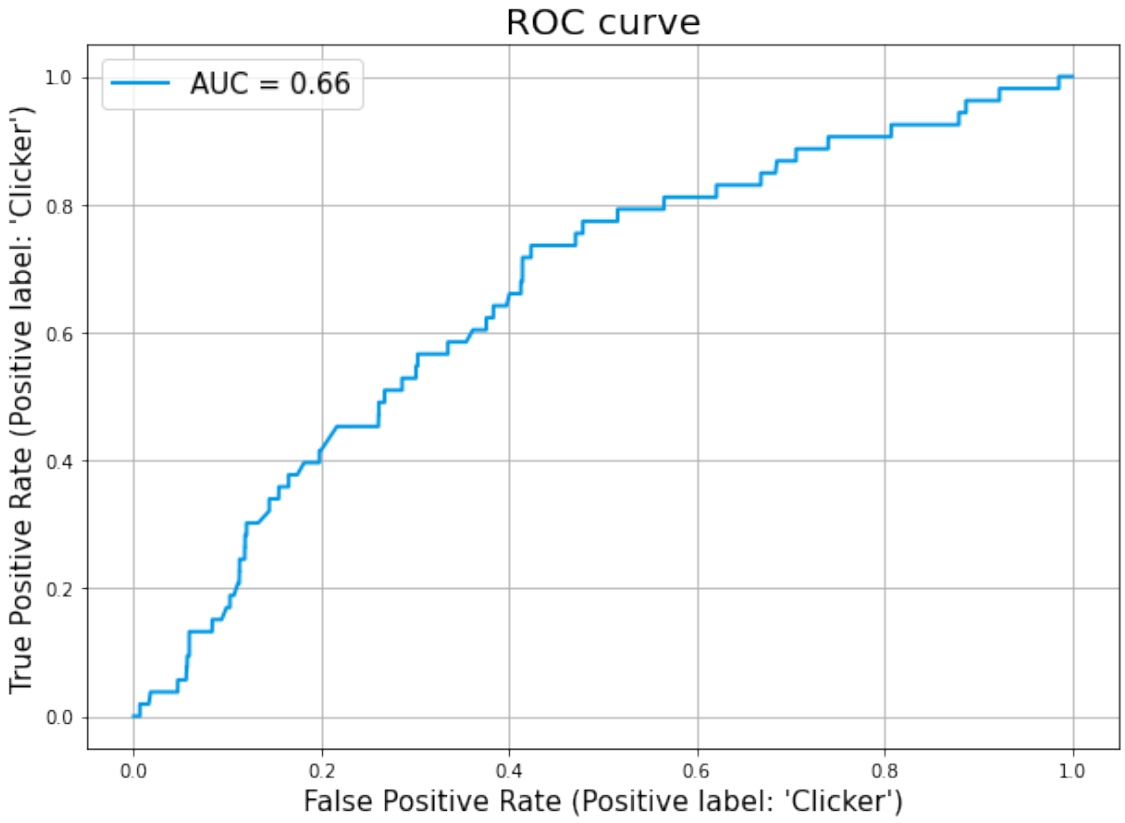

Figura 13: Curva ROC del modello ottenuta dal 30% del dataset non utilizzato per l’addestramento.

Svolta la classificazione come stabilito, per il dataset di utenti non utilizzato per l’addestramento si ottengono le performance mostrate nella matrice di confusione in figura 14. Si può notare come i risultati siano in accordo con le percentuali precedentemente riportate, in particolare 41 utenti della classe “clicker” su 53 sono correttamente identificati. L’accuratezza di questa classificazione è pari a 0.52, circa 10 punti percentuali più bassa rispetto a quella ottenibile utilizzando la soglia di default: questo è dovuto all’aumento del numero di “non clicker” classificati in modo errato che si sono ottenuti dall’abbassamento della soglia per la confidenza.

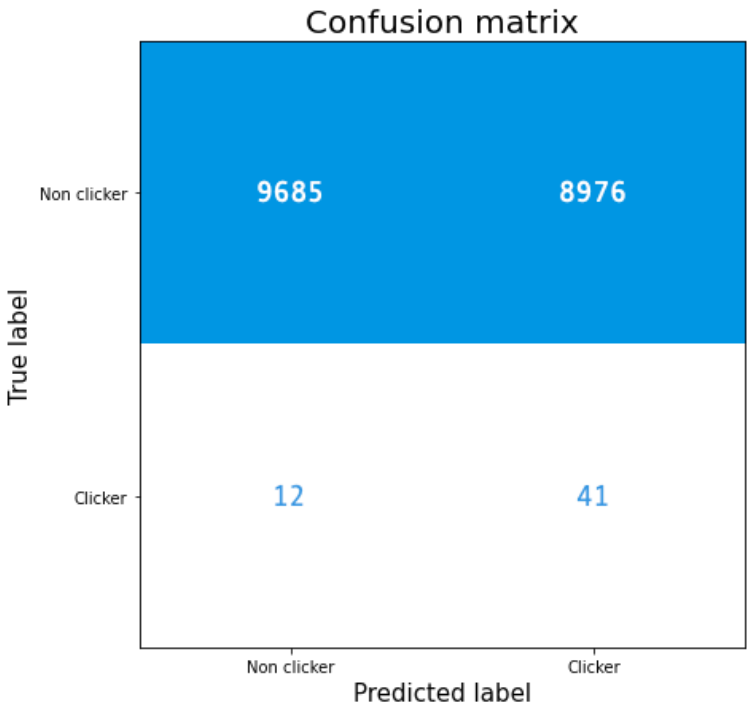

Figura 14: matrice di confusione ottenuta dal 30% del dataset non utilizzato per l'addestramento.

Conclusione

Il dataset analizzato presenta informazioni relative a diversi utenti raccolte tramite i cookie di una campagna pubblicitaria. Tra queste informazioni è riportato se e quante volte l'utente ha cliccato sull'inserzione mostrata, rendendo quindi possibile la realizzazione di un modello che stabilisca se, date le caratteristiche dell'utente, questo cliccherà sull'inserzione o meno. All'interno del dataset gli utenti che hanno cliccato almeno una volta sono lo 0.3% del totale: il dataset risulta pertanto fortemente sbilanciato.

Del dataset sono stati mantenuti 62379 utenti, ovvero quelli che non presentano valori mancanti per più variabili, non sono ritenuti sospetti dal motore dedicato e non hanno ricevuto la pubblicità tramite e-mail. Le variabili mantenute sono invece 395 delle 1416 di partenza: sono state rimosse quelle con valori pari a zero per tutti gli utenti e quelle la cui informazione è in larga parte presente anche nelle variabili mantenute.
Il dataset così processato è stato suddiviso in due set, contenenti rispettivamente il 70% e il 30% degli utenti. Il primo set è stato utilizzato per addestrare una SVM: per fare in modo che lo sbilanciamento non influisse sulle prestazioni, al set è stato applicato un algoritmo di Random Oversampling, in modo che il numero di utenti che cliccano e che non cliccano fosse uguale.
Il secondo set è stato utilizzato per valutare come il modello associa agli utenti che realmente cliccano la confidenza di appartenere alla classe "clicker". Gli utenti sono stati suddivisi in gruppi sulla base della mediana, dei quartili e degli ottili della distribuzione dei valori di confidenza e per ogni gruppo è stato contato il numero di utenti che realmente cliccano. Dai grafici nelle figure 10, 11, 12 si può desumere che il funzionamento del modello è adeguato: si osserva che dai due gruppi costruiti a partire dalla mediana, quello associato a valori di confidenza più alti contiene più del triplo di utenti che realmente cliccano rispetto al gruppo associato a valori di confidenza più bassi; inoltre, dagli otto gruppi costruiti a partire dagli ottili, quello associato ai valori di confidenza più alti contiene circa il doppio di utenti che realmente cliccano rispetto al secondo gruppo più popolato.

Una volta stabilito che il modello associa confidenze più alte agli utenti che realmente cliccano, si sono potute valutare le performance della classificazione. Tramite la curva ROC costruita sulla base del secondo set è stato possibile determinare la soglia della confidenza oltre la quale il classificatore associa a un utente la classe "clicker" e al di sotto della quale viene invece assegnata la classe "non clicker". La soglia scelta è pari a 0.464 poiché permette di classificare erroneamente circa la metà degli utenti che non cliccano, riuscendo tuttavia a classificare correttamente più di tre quarti degli utenti che cliccano. Le performance della classificazione sono quindi mostrate nella matrice di confusione in figura 14: su 53 utenti che realmente cliccano, 41 di questi vengono associati alla classe corretta, mentre per gli utenti che non cliccano, il 52% su un totale di 18661 viene classificato correttamente.
Poiché la maggior parte degli utenti appartengono alla classe "non clicker", un'azienda che decida di utilizzare il modello addestrato per stabilire a quali utenti mostrare l'inserzione potrà scartare circa metà di tutti i possibili utenti. L'azienda potrà quindi mostrare l'inserzione alla restante metà di utenti, sapendo che tra questi sono presenti circa tre quarti degli utenti che cliccheranno e quindi potenzialmente interessati al prodotto pubblicizzato.

Nel caso l’azienda voglia mostrare l’inserzione a un più alto numero di utenti potenzialmente interessati, sarà necessario abbassare la soglia di confidenza oltre la quale all’utente viene assegnata la classe “clicker”. L’azienda dovrà tuttavia accettare che aumenterà anche il numero di utenti “non clicker” classificati erroneamente come “clicker”: se ad esempio si scegliesse una soglia che permetta di inviare l’inserzione al 90% dei “clicker”, l’inserzione verrà inviata anche a circa 74% dei “non clicker”.

Sitografia e bibliografia

[1] Seaborn
[2] MatPlotLib
[3] Sklearn
[4] Oversampling and under sampling methods for imbalanced classification
[5] Marco Fattore, Metodi di riduzione della dimensionalità