Crowdfunding: analisi predittiva del lancio di progetti su Kickstarter

Alessandro Motta[1], Matteo Paparella[2], Mattia Ventola[3], Gabriele Zottola[4]
(Universita` degli Studi di Milano Bicocca, CdLM Data Science)

Abstract

La domanda alla quale si è cercato di rispondere, e sulla quale si basa il nostro lavoro è la seguente: è possibile, sulla base dei dati a nostra disposizione, riuscire a prevedere l'esito di una campagna di un progetto, in particolare come successo o fallimento? Kickstarter permette ad un progetto di concretizzarsi grazie al sostegno di una community di investitori. Una volta raggiunto il tetto di obiettivo di finanziamento, la campagna ha esito positivo, rendendola quindi classificabile come "successo". Partendo dagli attributi a nostra disposizione, abbiamo ragionato sulle informazioni che avremmo precedentemente al lancio di un progetto. Utilizzando gli attributi più adatti al nostro problema e creando degli attributi nuovi, da quelli di partenza, abbiamo utilizzato vari classificatori, per predire l'esito di una campagna di un progetto. Siamo giunti così a trovare il modello ottimale per individuare i possibili esiti di una campagna, partendo da informazioni note solo al lancio della campagna stessa. La nostra analisi predittiva, potrebbe ritenersi molto utile, per coloro che decidono di lanciare un progetto su Kickstarter, e vorrebbero capire, sotto quali condizioni, il progetto potrebbe ottenere più probabilità di successo, che è l'obiettivo primario di chi lancia una campagna di crowdfunding.

Indice

1. Introduzione

Per Crowdfunding si intende un'attività di raccolta di capitali grazie ad un numero elevato di investitori che, tramite piattaforme web, finanziano progetti "non-profit" e "for profit". In questo secondo caso si parla più precisamente di Crowdinvesting.
Queste piattaforme web di crowdfunding permettono agli investitori di non essere solo dei "benefattori" economici, ma permettono ai fondatori delle startup di confrontarsi con gli utenti per ottenere dei feedback critici per il loro progetto.
Il tema del crowdfunding è un tema che negli ultimi anni ha interessato e incuriosito moltissime persone e investitori di qualsiasi parte del globo. Kickstarter, la piattaforma americana di crowdfunding più famosa al mondo, ha sviluppato una modalità di investimento peer-to-peer (da pari a pari) anche per i finanziamenti "for-profit".
Periodicamente, vengono diffusi dati che ci dovrebbero fare capire l'impatto che il Crowdfunding ha oggi nell'economia mondiale.
Nel mondo il crowdfunding si concentra principalmente negli Stati Uniti e in Cina, mentre nel continente europeo il fenomeno

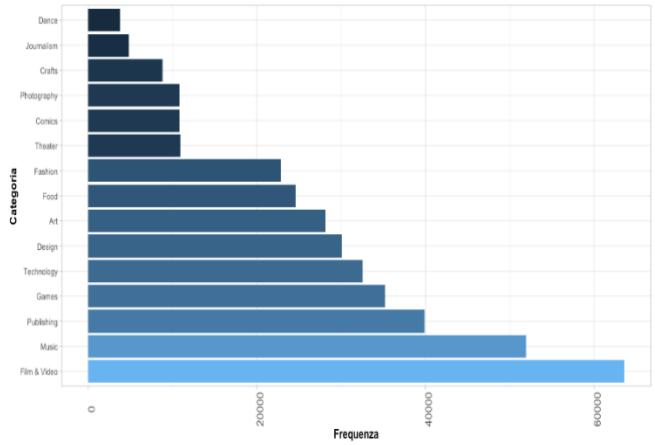
[1] Universita degli Studi di Milano Bicocca, CdLM Data Science, matricola 812309
[2] Universita degli Studi di Milano Bicocca, CdLM Data Science, matricola 812561
[3] Universita degli Studi di Milano Bicocca, CdLM Data Science, matricola 812475
[4] Universita degli Studi di Milano Bicocca, CdLM Data Science, matricola 812363

assume dimensioni contenute, concentrate principalmente in Gran Bretagna.
Secondo i dati rilevati dal Cambridge Centre for Alternative Finance, dopo uno studio effettuato nel 2017, in Europa la quantità di finanziamenti derivanti dal Crowdfunding nel mercato europeo nel 2016 si aggira intorno agli €8 miliardi di finanziamenti (di cui il 73% sono attribuibili alla Gran Bretagna). Nelle Americhe i volumi di finanziamenti sono all'incirca € 32 miliardi (di cui il 98% negli USA), mentre in Asia €220 miliardi (di cui il 99% in Cina).
Su Kickstarter si possono finanziare progetti per qualsiasi categoria: film, giochi, musica, arte, design e tecnologia e molto altro. Questa piattaforma è molto semplice: chi avvia la campagna su Kickstarter decide l'obiettivo (in denaro) del finanziamento e decide la scadenza del progetto. Ognuno può contribuire al progetto inserendo la somma di denaro che vuole, oppure scegliendo uno dei pacchetti che prevedono una sorta di "ricompensa", in base alla somma versata. I finanziatori però non ricevono delle partecipazioni alla società futura in base a quanto hanno investito. Solitamente i finanziatori vengono "ripagati" dal loro contributo con delle copie di ciò che viene realizzato o un'esperienza legata al progetto. Kickstarter ha ospitato circa 250'000 progetti con oltre 4 miliardi di dollari raccolti complessivamente per i progetti.

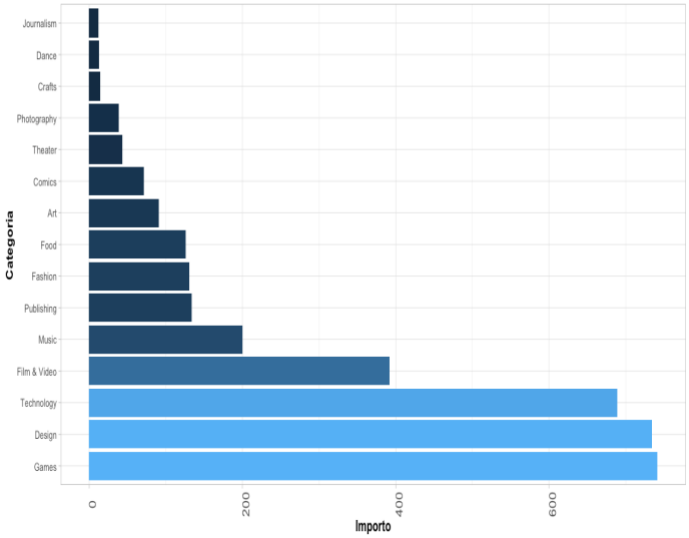
2. Dataset e Preprocessing

2.1 Analisi Esplorativa

Kaggle ci ha fornito due dataset concernenti i dati riguardanti i progetti per il 2016 e il 2018. Per la nostra analisi abbiamo deciso di tenere in considerazione solamente il dataset riferito all'anno 2018 per due motivi principali:
1) Unire i due dataset avrebbe reso il dataset finale di grandi dimensioni, rallentando così il lavoro sulla piattaforma Knime;
2) Il file del 2018, essendo più recente, ci ha permesso di effettuare un'analisi sulle macro-categorie più attuali, visto che il dataset dell'anno 2016 possedeva macro-categorie obsolete.
Il dataset del 2018 che è stato utilizzato contiene i progetti dal 1° gennaio al 31 Dicembre.
Il dataset contiene indicazioni riguardanti le macro-categoria e sub-categoria del progetto, il giorno di lancio e la deadline del progetto, e la quantità di denaro che è posta come obiettivo per il finanziamento. È fornito anche lo Stato in cui è stato lanciato del progetto, la quantità di finanziatori e se il progetto sia andato a buon fine o è fallito. Prima di effettuare le analisi sui modelli predittivi abbiamo analizzato il dataset. Per effettuare i seguenti grafici è stato necessario l'utilizzo del Software R e delle librerie *ggplot* e *tidyverse*.
Il primo attributo che è stato preso in considerazione per questa analisi è stato "main category". Delle categorie principali è stato analizzato il volume totale dei progetti.
Come mostrato dal grafico, le categorie con il più alto numero di progetti sono state "Film e Video" (63'585 progetti), "Music" (51'918), "Publishing" (39'874) e "Games" (35'231). Mentre le categorie che hanno totalizzato il numero più basso di progetti sono "Dance" (3'768), "Journalism"(4755), "Crafts"(8809) e Photography(10'779).



Successivamente abbiamo preso in considerazione l'attributo "usd_pledged" che rappresenta il totale dell'importo impiegato per ciascuna categoria.
Il seguente grafico mostra il totale, in milioni di dollari, di investimenti per ogni categoria principale.



Come mostra il grafico, la categoria che presenta il maggiore volume di investimenti è "Games" con 742 milioni di dollari complessivi. Segue "Design" con 735 milioni di dollari e "Technology" con 689 milioni. Si nota inoltre che le tre categorie che nel grafico precedente di trovavano ad essere le categorie con più progetti complessivi ("Film e video", "Music" e "Publishing"), non sono però quelle che richiedono un investimento più esiguo visto che si trovano nelle posizioni più indietro. Le categorie con una minore qualità di progetti si ritrovano, ovviamente, ad essere anche le ultime per importo complessivo investito.

2.2 Preprocessing

Partendo da queste analisi abbiamo deciso di eliminare gli attributi ridondanti o poco informativi. Abbiamo mantenuto invece diversi attributi importanti quali:

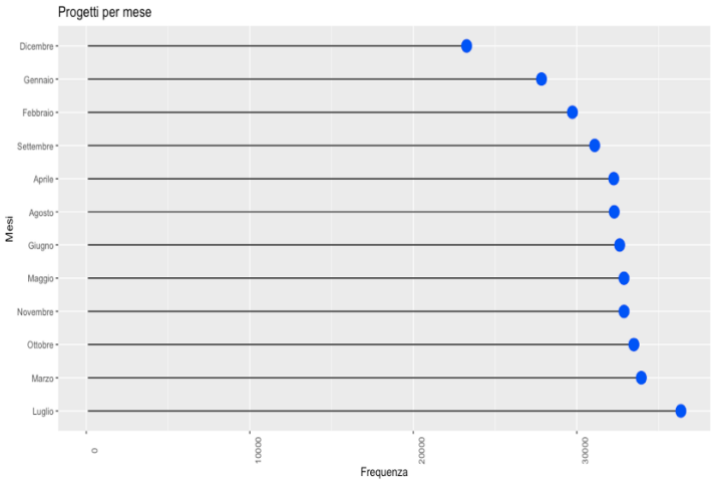
- "State";
- "Main category";
- "Deadline";
- "Usd goal real";
- "Launched";

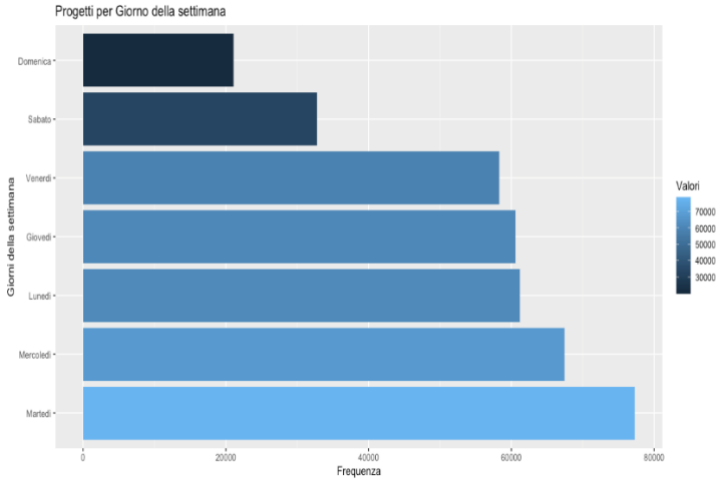
Una volta selezionate gli attributi più adatti per il nostro obiettivo di predizione, abbiamo deciso di creare da questi ultimi ulteriori attributi. Partendo dalle date di lancio e di termine della nostra campagna, abbiamo calcolato la differenza di tempo tra il termine e l'inizio della campagna stessa. Abbiamo così creato l'attributo "project_duration", con il numero di giorni di presenza della campagna su Kickstarter, considerando che probabilmente più giorni una campagna rimane attiva, più ci potrebbero sostenitori, e quindi investimenti. Dal giorno di lancio, abbiamo estrapolato:

- launched_month, che ci permette di sapere in quale mese il progetto è stato lanciato;
- launch_day, che ci permette di sapere il giorno della settimana nel quale il progetto è stato lanciato;

- weekend, che ci permette di sapere se la campagna è iniziata nel weekend o meno;

Considerando che i progetti presentano un ampio tasso di sottoscrizioni in tempi brevi dal lancio, dato che il progetto con il passare del tempo, viene superato da progetti più recenti, quindi obbligando chi ha lanciato la campagna a pagare kickstarter per mantere il progetto nelle prime posizioni di ricerca.
Sempre con l'ausilio di R abbiamo analizzato, tramite i seguenti grafici quali i giorni e i mesi in cui venivano lanciati più progetti e in quali ne fossero lanciati di meno.





Dal grafico si può vedere che il numero maggiore di progetti viene lanciato nella prima parte della settimana e poco nel weekend. Per i mei i dati si bilanciano molto con un picco a Luglio.

Per poter avere una predizione sul successo e sul fallimento finale del progetto abbiamo eliminato tutte le righe che presentavano nell'attributo "State" (Vedo appendice) tutte le modalità che non fossero "Successo o "Fallito". Le modalità eliminate sono: "Cancelled","Live","Suspended","Undefined" (12% dei progetti totali).
Per quanto riguarda i Missing Values, il dataset preso in considerazione non ne presenta.

3. Modelli

3.1 Utilizzo dei modelli

Per lo studio sono stati implementati diversi modelli di classificazione per poter individuare la tecnica più adatta a prevedere la possibilità di successo di un progetto.
In particolare, sono stati utilizzati:

Modelli probabilistici: modelli che, data un'osservazione in input, permettono di prevedere la distribuzione di probabilità su un insieme di classi possibili, invece che mostrare unicamente la classe con la maggior probabilità di appartenenza dell'attributo. In particolare, i classificatori utilizzati per l'analisi sono le Bayesian Network, utilizzando il metodo K2 e TAN, Naïve Bayes e Naïve Bayes Tree, il quale genera un albero decisionale avente classificatori Naïve Bayes sulle foglie ed infine A1DE.

Modelli euristici: modelli che permettono la classificazione sfruttando i decision trees, i quali ad ogni nodo scelgono l'alternativa migliore in funzione delle informazioni disponibili. Per l'analisi sono stati utilizzati i modelli Random Forest Learner, il Decision Tree Learner, utilizzando per lo splitting l'indice di Gini e con 8 decision trees ed il J48.

Modelli di regressione: il metodo di regressione selezionato è unicamente la Regressione Logistica, poiché a differenza della Regressione Lineare essa permette di prevedere il valore di una variabile dicotomica (state) utilizzando un insieme di variabili esplicative.

3.2 Holdout

Il dataset a nostra disposizione lo abbiamo suddiviso, secondo il metodo holdout, in 67% per la parte di training set e 33% per la parte di test set. Abbiamo utilizzato il training set per addestrare i nostri classificatori, per poi passare ad una validazione successiva con il test set.
Abbiamo deciso di utilizzare come modelli: Random Forest Learner, NBTree (3.7), Bayesnet (3.7) K2, Bayesnet (3.7) TAN, NaiveBayes (3.7), Decision Tree Learner, J48, A1DE.
Ad ogni modello, abbiamo impostato gli attributi in modo da renderli performanti con i nostri modelli, utilizzando un flow di nodi per modificare la tipologia di attributi.

3.3 Cross-validation

Ogni classificatore è stato poi sottoposto ad un processo di cross validation: è stata effettuata un'ulteriore suddivisione del training set in 5 sottoinsiemi utilizzando il campionamento casuale semplice e utilizzando 4 di essi come training set e la rimanente porzione come test set. I risultati ottenuti sono la media delle 5 iterazioni effettuate.

4. Analisi e Valutazione

4.1 Metriche

Per poter valutare le performance dei diversi modelli e per poterli confrontare tra loro, si è scelto di impiegare per l'analisi diverse metriche:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

L'*Accuracy* indica la percentuale di osservazioni positive (TP) e negative (TN) correttamente classificate, sul totale delle previsioni (TP + TN + FP + FN).
In generale, valori elevati di *Accuracy* indicano una previsione corretta ma è utile considerare anche altre metriche per poter avere una visione più precisa di quale modello riesca a prevedere al meglio le istanze.

$$Error = \frac{FN + FP}{TP + TN + FP + FN}$$

L'*Error*, definito anche come (*1-Accuracy*), indica la percentuale di osservazioni classificate erroneamente sul totale.

$$Recall = \frac{TP}{TP + FN}$$

La *Recall* indica la porzione di osservazioni positive correttamente classificate dal modello.
Valori alti di *Recall* indicano che sono state classificate correttamente la maggior parte delle osservazioni positive.

$$Precision = \frac{TP}{TP + FP}$$

La *Precision* indica la porzione di osservazioni effettivamente positive rispetto al totale delle osservazioni predette come tali dal modello.
Valori alti di *Precision* indicano che poche osservazioni sono state classificate erroneamente come positive.

In funzione del modello scelto, *Precision* e *Recall* possono risultare in conflitto tra loro, poichè non tutti sono in grado di massimizzare contemporaneamente entrambe le metriche.
Per ovviare a questo problema si utilizza la *F_1-measure*, calcolata come la media armonica tra Precision e Recall:

$$F_1\text{-}measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

Valori alti di *F_1-measure* indicano che entrambe le metriche hanno valori alti.

Come ulteriore strumento di valutazione della correttezza delle previsioni è stata utilizzata la *Curva di ROC*, un grafico avente in ascissa la *FPR* (*False Positive Rate*), cioè la percentuale di osservazioni negative erroneamente classificate come positive e in ordinata il *TPR* (*True Positive Rate*), cioè la percentuale di osservazioni positive effettivamente previste come tali. Si tendono a preferire le Curve di ROC che si discostano maggiormente dalla diagonale.

Dalla curva di ROC è possibile ricavare un'ulteriore metrica, l'*AUC* (*Area Under Curve*), la quale rappresenta il valore dell'area sottesa dalla Curva di ROC.
Valori elevati di AUC rispecchiano un elevato discostamento della Curva di ROC dalla diagonale.

4.2 Valutazione Holdout e Cross Validation

La prima analisi dei risultati è stata effettuata sulla metrica "accuracy" del training set e test set di ogni modello, per verificare casi di underfitting o overfitting. Il valore dell'indice si discosta, per ogni classificatore, in modo irrilevante, con una differenza massima di 0.05 punti tra le due partizioni del dataset potendo così considerare validi i modelli scelti.
Si è proceduto dunque a confrontare tutti i valori calcolati tra i modelli.
In base ai risultati che abbiamo ottenuto, considerando l'obiettivo del nostro problema, ossia effettuare un'ottima predizione di successo o fallimento di un progetto, possiamo trarre le prime osservazioni:
Notiamo dei valori di precision non troppo alti, ma che comunque ci permettono di considerare i modelli NBTree (3.7), BayesNet(3.7) TAN, J48 e A1DE come quelli più interessanti per questa metrica di valutazione;
Per quanto riguarda l'error, possiamo notare che il classificatore decision tree learner, presenta il più alto valore di error. Gli altri modelli di classificazione si attestano su dei valori simili.
possiamo constatare per l'accuracy dei valori alti per quanto riguarda i classificatori con Bayes, in particolare NaiveBayes (3.7), BayesNet (3.7) TAN, inoltre anche il modello A1DE presenta un'accuracy leggermente migliore;
osservando la metrica di recall, i modelli si attestano intorno allo stesso valore, eccetto il modello random forest learner che presenta un indice piuttosto basso, oltre anche al decision tree learner.

	Recall	Precision	Error	F-Measure	Accuracy
Random Forest Learner	0.402	0.585	0.371	0.402	0.629
NBTree (3.7)	0.532	0.598	0.343	0.532	0.657
BayesNet (3.7) K2	0.532	0.598	0.343	0.532	0.657
BayesNet (3.7) TAN	0.518	0..608	0.342	0.518	0.658
Naive Bayes (3.7)	0.532	0.598	0.343	0.532	0.657
Decision Tree Learner	0.475	0.493	0.412	0.475	0.588
J48	0.512	0.591	0.35	0.512	0.65
A1DE	0.533	0.603	0.341	0.533	0.659
Simple Logistic (3.7)	0.165	0.556	0.393	0.254	0.607
Logistic (3.7)	0.164	0.556	0.392	0.254	0.607

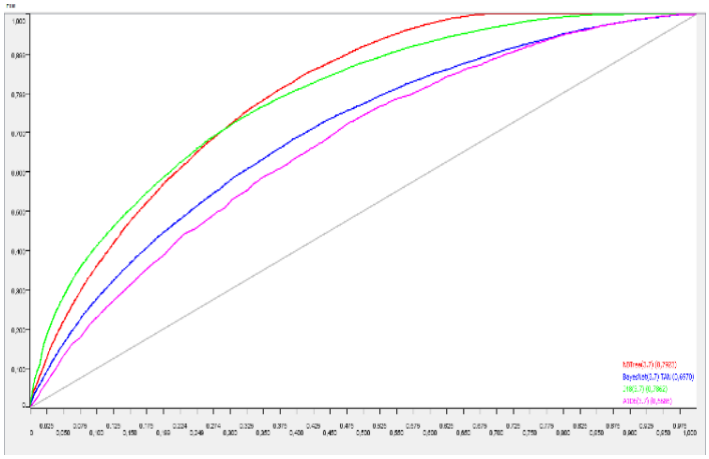
4.2 Valutazione Adaptive Boosting

Abbiamo selezionato come modelli di classificazione NBTree (3.7), BayesNet (3.7) TAN, J48 e A1DE, ponendo l'attenzione sulle metriche di valutazione di precision e f-measure.
Focalizzandoci sul nostro obiettivo, sono stati testati i modelli precedentemente selezionati applicando l'algoritmo di Adaptive Boosting5, per verificare la reazione delle metriche di valutazione dei nostri classificatori al fine di un'ottimizzazione in favore dei record classificati in modo errato.
Di seguito mostriamo i valori delle metriche, successivamente al metodo di Adaptive Boosting:

	Recall	Precision	Error	F- Measure	Accuracy	AUC
NBTREE (3.7)	0.667	0.64	0.354	0.653	0.646	0.797
BayesNet (3.7) TAN	0.666	0.546	0.359	0.6	0.641	0.699
J48	0.749	0.599	0.304	0.665	0.696	0.779
A1DE	0.66	0.525	0.379	0.585	0.621	0.68

Possiamo constatare un leggero abbassamento dei valori di accuracy dei nostri modelli, con un aumento dei risultati di recall e di precision, che hanno condizionato anche i valori di f-measure. In particolare, possiamo notare come i valori di f-measure dei modelli J48 e NBTree (3.7) risultano essere nettamente migliorati, rendendoli i modelli più idonei per il nostro obiettivo. Inoltre, i valori di precision di questi ultimi, risultano superiori rispetto a BayesNet (3.7) e A1DE.

Inoltre, per un ulteriore confronto, sono stati rappresentati i vari modelli con la ROC curve, che ci ha permesso di osservare che i classificatori NBTree (3.7) e J48 risultano avere una parabola più alta rispetto agli altri due classificatori. Come ulteriore conferma dell'affermazione precedente, l'AUC (area sottesa della curva) presenta valori più alti nei modelli prima citati (J48 = 0.779 e NBTree (3.7) = 0.797). Osserviamo, ponendo l'attenzione sul valore dell'AUC, che il modello NBTree (3.7) presenta un valore più alto.
Di seguito rappresentiamo i valori della ROC curve:



5. Conclusioni

Con il dataset ottenuto si è sperimentato l'utilizzo di tecniche e algoritmi di machine learning con il fine di classificare un progetto come un successo o un fallimento, in base alle caratteristiche date prima del lancio sulla piattaforma "Kickstarter".
Una prima criticità del lavoro si è trovata nella presenza di attributi poco soddisfacenti che hanno portato ad una analisi più approfondita del dataset.
Tramite un processo di data exploration si è potuto mostrare, come visto nel capitolo 2 "dataset e preprocessing", una rilevante incidenza di alcuni attributi sull'esito di un progetto. L'approfondimento dell'analisi ha quindi permesso di estrapolare nuove features e creare un dataset che potesse racchiudere il numero maggiore di informazioni che è possibile conoscere a priori del lancio di una campagna di crowfounding.
Si è poi deciso di testare differenti modelli e di diverse tipologie per poter avere una visione più ampia e una varietà di risultati che potessero permettere di trarre conclusioni. I modelli utilizzati, insieme alle tecniche usufruite, hanno però portato a metriche poco soddisfacenti. Con i valori ottenuti è nata l'esigenza di selezionare unicamente i modelli con le misure desiderate maggiori e di testarli, nuovamente, sotto diverse modalità.
Tramite l'utilizzo di un algoritmo differente (AdaBoost, citato nel capitolo 4) e una nuova fase di test è stato possibile ottimizzare, anche se lievemente, i risultati, prendendo quindi in considerazione i modelli NBTree (3.7) e J48. I classificatori individuati infatti mostrano una precision che si aggira intorno al 60% e un livello di f-measure poco oltre la soglia prima citata.
Si conclude quindi che, l'applicazione di tecniche e algoritmi di machine learning, sono sì in grado di raggiungere l'obiettivo prefissato dall'analisi ma non con una accuratezza elevata. La difficoltà di raggiungere una elevata precisione si ipotizza essere causata da determinati attributi che possono incidere in modo rilevante sull'esito di un progetto, ma che è difficile conoscere a priori.

[5] L'algoritmo formula H ipotesi tramite l'algoritmo di ensemblee boosting, ogni ipotesi è un albero decisionale diverso perché costruito usando parametri differenti. A ciascuna ipotesi assegna un determinato peso per misurare l'efficacia ed elabora a termine l'ipotesi finale sulla somma dei pesi delle ipotesi più alte dalle altre.

6. Appendice

Attributi presenti nel Dataset:

ID = l'id del progetto sulla piattaforma di kickstarter;

NAME = il nome del progetto che andrà ad apparire sulla piattaforma di kickstarter. È composto da una serie di parole che descrivono il progetto;

CATEGORY = la categoria nel quale il progetto andrà ad essere posizionato.

MAIN_CATEGORY = la categoria di una campagna;

CURRENCY = la valuta della moneta con il quale è stato lanciato il progetto;

DEADLINE = il giorno che viene posto come termine della campagna del progetto;

GOAL = l'obiettivo di raccolta a livello economico del mio progetto;

LAUNCHED = la data del lancio della nostra campagna del progetto;

PLEDGED = l'ammontare di denaro che è stato raccolto dalla campagna;

STATE = lo stato nel quale il progetto è terminato, in particolare può essere live, canceled, life, success, failed

BACKERS = il numero di investitori di ogni singolo progetto;

COUNTRY = la nazione dalla quale è arrivato il caricamento del progetto;

USD_PLEDGED = l'ammontare di denaro raccolto in USD (con la conversione effettuata da KS);

USD_PLEDGED_REAL = l'ammontare di denaro raccolto in USD (con la conversione di fixed.io api);

USD_GOAL_REAL = l'obiettivo di raccolta in USD.