Predlog projekta

Definicija projekta

Predviđanje uspešnosti finansiranja crowdsourcing projekta objavljenog na Kickstarteru na osnovu naziva, kategorije, analize dostupnog sadržaja koji opisuje projekat u vidu slika i teksta i video materijala, kao i količine i vrste traženih finansijskih sredstava za realizaciju projekta. Ideja ovog projekta se zasniva na analizi svih dostupnih parametara koji definišu jedan Kickstarter startup projekat u cilju korišćenja dobijenih rezultata radi sticanja uvida u potencijalne korake koji dovode do povećanja uspešnosti novih projekata.

Motivacija

Svaki startup projekat zahteva početni kapital kako bi zaživeo. Veliki broj preduzetnika nije u mogućnosti da ispuni zahteve početnog kapitala, stoga su prinuđeni da traže investitore koji će da ulože novac u njihovu ideju. Jedan od načina kako je moguće pronaći potreban novac, jeste korišćenjem crowdsourcing platformi, od kojih je najpopularnija Kickstarter. Kroz povećanje popularnosti same platforme, raste i broj projekata koji ne uspevaju da sakupe dovoljno sredstava da bi pokrili prethodno projektovane troškove. Cilj ovog projekta je predviđanje uspešnosti Kickstarter projekta u trenutku njegovog objavljivanja.

Relevantna literatura

KickPredict: Predicting Kickstarter Success

Kevin Chen, Brock Jones, Isaac Kim, Brooklyn Schlamp
Dept. of Computing and Mathematical Sciences California Institute of Technology 2013.

Svrha ovog projekta je bila da se razvije sistem predviđanja da li će Kickstarter projekat biti uspešan pre njegovog završetka. Kako bi se to izvelo, korišćen je SVM model obučen na velikom skupu podataka o Kickstarter projektima. Skup analiziranih podataka je preuzet direktno sa sajta Kickstarter, kao i sa portala Kicktraq koji prati statistiku Kickstarter projekata. Skup podataka je sadržao sledeće informacije:

- Uplaćeni iznos tokom vremena

- Broj projekata iza kojih stoji kreator
- Ukupan broj projekata koje je kreirao kreator
- Da li je kreator povezan na Fejsbuk
- Ciljani iznos finansiranja
- Dužina projekta
- Broj slika prisutnih na stranici projekta
- Broj karaktera u opisu projekta
- Da li stranica projekta ima video ili ne
- Da li stranica projekta ima Youtube video ili ne
- Ako je prisutan Youtube video, broj pregleda tokom vremena
- Koliko puta je Tvitter veza projekta bila podeljena

Evaluacija rešenja vršena je na osnovu podele podataka na skup za obučavanje (19000 Kickstarter projekata) i testni skup (1000 Kickstarter projekata). Postignuta je preciznost predviđanja uspešnosti projekta u nultom danu od 67%. Zaključeno je da su najznačajnije stavke koje utiču na uspešnost projekta:

- Broj projekata iza kojih stoji kreator
- Ukupan broj projekata koje je kreirao kreator
- Da li stranica projekta ima video ili ne
- Ukupan cilj potreban da bi projekat dobio sredstva

U našem projektu će biti korišćen SVM, kao i tehnike kreiranja kompletnog skupa podataka. Kako se projekti često razlikuju po količini teksta i broju slika korišćenim za promociju, analiziranjem njihovog odnosa bi potencijalno bilo moguće dodatno napraviti razliku između uspešnih i neuspešnih projekata,i time potencijalno povećati preciznost predikcije.

http://courses.cms.caltech.edu/cs145/2013/blue.pdf

Predicting Crowdfunding Success with Optimally Weighted Random Forests

Fahad Sarfaraz Ahmad, Devank Tyagi, *Simran Kaur*
Delhi Technological University New Delhi, India 2017

Cilj ovog projekta bilo je kreiranje softvera za predikciju uspešnosti Kickstarter projekta koristeći Random Forest algoritam. Skup podataka koji je korišćen je preuzet sa sajta Kicktraq i sadržao je podatke o projektima od 01.12.2009. do 20.03.2017. Za parsiranje podataka korišćena je Nokogiri biblioteka.
Skup podataka je sadržao:

- Kategorija projekta
- Ciljana suma finansiranja
- Indeks čitljivosti datog opisa projekta
- Broj recenica u datom opisu projekta

- Broj recenica u datom opisu nagrade
- Broj slika na stranici
- Broj nagrada koje je dao korisnik
- Broj veb stranica povezanih sa projektom
- Broj saradnika na projektu
- Broj facebook prijatelja koje kreator ima
- Broj projekata koje je kreirao kreator
- Broj projekata koje je kreator podrzao

Kako bi se ispitalo rešenje i ocenili modeli predikcije korišćene su sledeće mere tokom empirijskog testiranja algoritma:

- Tačnost
- Preciznost
- F-mera

Tokom evaluacije rešenja predikcije random forest algoritmom zaključeno je da tačnost iznosi 94.29%, preciznost 94.5%, a F-mera 94.3%. Nasuprot prethodno navedenom istraživanju, broj projekata koje je kreator prethodno objavio nema veliki uticaj na uspešnost određenog projekta.

Kako se zaključci ovog i prethodno navedenog istraživanja ne poklapaju, biće dodatno ispitan uticaj broja prethodno započetih projekata kreatora, kao i efikasnost random forest algoritma na drugom skupu podataka kako bi se stekao dodatni uvid u pouzdanost predikcije.

https://ieeexplore.ieee.org/abstract/document/8286110

Success Prediction using Random Forest, CatBoost, XGBoost and AdaBoost for Kickstarter Campaigns

Siddharth Jhaveri, Ishan Khedkar, Yash Kantharia, Shree Jaswal
Department of Information Technology, St. Francis Institute of Technology, Mumbai, India 2019

U ovom radu cilj je bio da se vrši predikcija uspešnosti Kickstarter kampanje da bi pomogla budućim vlasnicima kampanja da lakše utvrde da li je njihov plan dovoljno dobro prezentovan. Rađeno je na osnovu istorijskih podataka svih kampanja od 2014 do 2019.

Bitni parametri koji su korišćeni su:

- Kratak opis u naslovu projekta
- Trajanje kampanje finansiranja
- Glavna kategorija projekta
- Podkategorija projekta
- Valuta koju kreator traži
- Lokacija projekta
- Ciljana suma

- Dužina naziva

Algoritmi su primenjivani nad kreiranim skupom kako bi se utvrdila njihova efikasnost, skup podataka je podeljen na trening skup i testni skup razmerom 90:10. Testni skup je podeljen po kategorijama projekata i klasifikacija je vršena za svaku kategoriju. Mere koje su korišćene tokom empirijskog testiranja algoritma:

- Tačnost
- Preciznost i
- Recall

Metode korišćene u ovom radu su Random Forest sa težinskim faktorima, AdaBoost, XGBoost i CatBoost, gde je utvrđeno koji je od njih najefikasniji. Zaključeni je da je najpouzdaniji CatBoost koji daje tačnost od 83.33%.

U ovom radu je naveden značaj lokacije projekta, kao i njegova kategorija, što će dodatno biti korišćeno tokom obučavanja algoritma predikcije. Takođe za svaku kategoriju biće pravljen poseban model predikcije.

https://ieeexplore.ieee.org/abstract/document/8819828

Skup podataka

Za ovaj projekat potrebno je obezbediti skup podataka nad kojim će se vršiti analize. Planirano je da se skup podataka preuzme sa sledećeg linka:

https://www.kaggle.com/datasets/kemical/kickstarter-projects

Iz ovog skupa podataka za svaki projekat moguće je uzeti naziv, glavnu i sporednu kategoriju, potrebnu količinu novca za uspešni projekat (goal), takođe i količinu koju je projekat dobio (pledged). Nalaze se informacije o datumima objavljivanja i roka do kog je planirano prikupljanje sredstava.
Pored navedenih stavki, skup podataka sadrži i status uspešnosti (successful, failed, canceled, live), na osnovu čega će podaci za obučavanje predickionog algoritma biti klasifikovani.

Skup podataka se sastoji od 378661 instanci.

Ciljno obeležje skupa podataka jeste “state” i ono podrazumeva 6 kategorija: failed (52%), successful (35%), dok ostalih 13% čine “live”, “canceled”, “suspended” i “undefined” od kojih su nam najznačajnije “failed” i “successful”. Odnos broja neuspelih projekata u odnosu na ostale govori da inicijalni skup podataka nije balansiran.

Pored toga, planirano je i preuzimanje opisa (tekst, slike i video materijal) projekata koji se nalaze u skupu podataka sa sajta, kao i imena njihovih osnivača, odnosno broja prethodno objavljenih projekata istog osnivača sa sajta: https://www.kickstarter.com/.

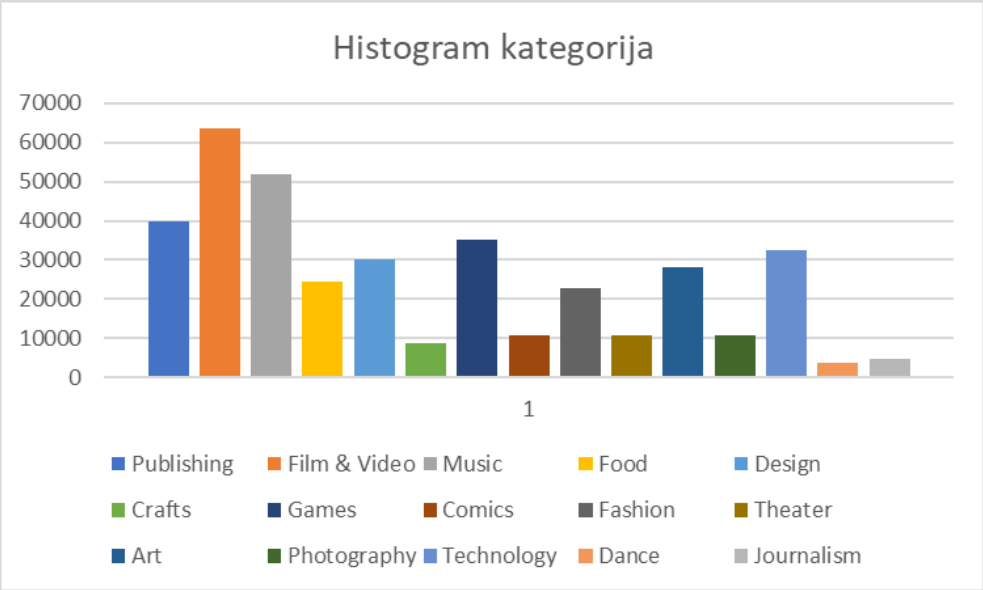
Softver

Za izradu projekta će biti korišćen PyCharm radno okruženje i python programski jezik. Za manipulisanje podacima Pandas biblioteka za analizu podataka. Svi podaci će biti čuvani u csv fajlu.

Metodologija

S obzirom da deo podataka planiranih za korišćenje nije sadržan u skupu podataka, potrebno ih je prethodno pripremiti za predikcioni algoritam. Koraci metodologije:

- Priprema dataseta, koja podrazumeva preuzimanje tekstualnog opisa i slika sa Kickstarter platforme za projekte sadržane u datasetu preuzetog sa Kaggle platforme, kao i uklanjanje projekata koji su otkazani ili čije prikupljanje još uvek traje. Projekti će biti podeljeni po kategorijama, i za svaku kategoriju će se obučavati drugi predikcioni model na sličan način kao što je urađeno u navedenom radu. Takođe biće izabran deo projekata koji su bili neuspešni da bi odnos uspešnih i neuspešnih projekata koii su analizirani bio 1:1.
- Postoji 15 glavnih kategorija i one su: Publishing, Film & Video, Music, Food, Design, Crafts, Games, Comics, Fashion, Theater, Art, Photography, Technology, Dance, Journalism.



- Naredni korak jeste obučavanje modela i optimizacija parametara nad validacionim skupom podataka. Algoritmi koji će se koristiti su SVM, Random Forest, XGBoost i Bagging koristeći SVM kao bazni klasifikator.

  Tokom analize teksta biće uzeta u obzir dužina opisa projekta. Takođe će se računati indeks čitljivosti.

  Kako je česta pojava da slike sadrže raznovrstan sadržaj informacija koje opisuju kickstarter projekat (Tekst, grafici, animacije, makete proizvoda), sadržaj slike bi bilo isuviše komplikovano analizirati, stoga će biti uzet u obzir broj slika, njihove dimenzije kao i odnos broja slika i dužine tekstualnog opisa.

  Iz istog razloga će se za analizu videa koristiti samo dužina njegovog trajanja.

- Testiranje modela nad test podacima
- Poređenje rezultata različitih algoritama, kao i poređenje sa drugim rešenjima iz radova navedenih u relevantnim literaturama
- Interpretacija uticaja različitih obeležja na uspešnost projekta.

Plan

Plana rada na ovom projektu obuhvata sledeće bitne tačke:

- Prikupljanje podataka
- Transformacija podataka

- Kreiranje modela
- Provera modela
- Vizualizacija dobijenih rezultata

Metod evaluacije

Krajnji rezultat dobijen obradom podataka biće predikcija da li će projekat biti finansiran u celosti. Uspešnost predikcije će se evaluirati nad skupom podataka koji će se biti podeljen na trening, validacioni i test skup. Mere korišćene za evaluaciju će bii tačnost, preciznost i F-mera. Učenje algoritama će se vršiti nad 70% od ukupnog skupa podataka, dok će ostatak da se podeli 50/50 na validacioni skup podatak i na testni skup podataka. Nakon evaluacije modela biće odrađena analiza grešaka izdvajanjem podskupa primera na kojima model greši i nad njima će biti izvršena ručna analiza da bi se utvrdili uzroci nastanka grešaka.

Tim

Tim čine : Stefan Krstić (E2 35/2022), Luka Stupar (E2 3/2022)