Predikcija uspešnosti mobilne aplikacije na osnovu metapodataka sa Google Play prodavnice

Luka Savkov E2 63/2025, Stefan Bogdanović R2 40/2024

1. Definicija problema

Potrebno je prikupiti i obraditi podatke o mobilnim aplikacijama sa Google Play prodavnice kako bi se analiziralo koji faktori najviše utiču na njihovu ocenu. Cilj projekta je modelovati uspeh aplikacije kroz zadatak regresije, gde će se na osnovu metapodataka vršiti predikcija prosečne korisničke ocene. Model treba da omogući uvid u to kako različiti parametri, poput kategorije i cene, diktiraju finalni rejting softvera pre nego što on postane dostupan široj javnosti.

2. Motivacija problema rešavanog u projektu

Analizom trendova na tržištu mobilnih aplikacija moglo bi se ustanoviti koje karakteristike doprinose boljem rangiranju, što bi imalo direktne primene za razvojne timove i kompanije. Ovakva analiza bi se mogla iskoristiti u svrhe optimizacije resursa tokom procesa razvoja, estimacije tržišnog uspeha, i smanjenja finansijskog rizika pri lansiranju novih proizvoda. Time se direktno utiče na poslovnu strategiju, ali i na opšte zadovoljstvo korisnika kroz kvalitetniji finalni proizvod.

3. Relevantna literatura

3.1 Predicting Android App Success Before Google Play Store Launch [pdf]

Tema rada: Rad se bavi analiziranjem faktora uspeha mobilnih aplikacija pre njihovog samog objavljivanja na prodavnici, modelujući problem kao predikciju marketinškog uspeha. Fokus istraživanja je na utvrđivanju korelacije između metapodataka dostupnih programerima i finalne ocene korisnika.

Podaci: Skup podataka obuhvata aplikacije sa Google Play Store-a i fokusira se na „interne" karakteristike kao što su naziv, kategorija, tip aplikacije i veličina instalacionog fajla.

Metodologija: U radu je primenjen linearni regresioni model nad prikupljenim metapodacima. Model je treniran tako da na osnovu ulaznih parametara generiše predviđenu kontinuiranu vrednost ocene.

Evaluacija rešenja: Rešenje je evaluirano merenjem odstupanja (greške) između predviđenih vrednosti i realnih ocena zabeleženih na prodavnici nakon lansiranja.

Zaključak: Rezultati pokazuju da metapodaci pružaju dovoljno informacija za validnu procenu uspeha. U našem projektu ćemo, po ugledu na ovaj rad, staviti prioritet na regresionu analizu „pre-launch" atributa.

3.2 Predicting Mobile Apps Performance using Machine Learning [pdf]

Tema rada: Tema rada je sistematsko poređenje različitih algoritama mašinskog učenja u svrhu predviđanja performansi aplikacija na osnovu njihovih karakteristika. Cilj je identifikovati koji algoritmi najbolje modeluju uspeh aplikacije.

Podaci: Skup podataka koji se koristi u radu sastoji se od preko 10.000 aplikacija sa Google Play Store-a sa atributima koji uključuju cenu, kategoriju i tip aplikacije.

Metodologija: Za formiranje prediktivnog modela testirano je šest različitih algoritama, uključujući linearnu regresiju i k-najbližih suseda (k-NN). Svaki model je prošao proces treninga i testiranja nad istim skupom podataka.

Evaluacija rešenja: Evaluacija je sprovedena poređenjem performansi modela u predviđanju kontinuiranih vrednosti rejtinga korisnika.

Zaključak: Kako su se k-NN i Random Forest pokazali kao najstabilniji modeli, planiramo da ih testiramo kao primarne prediktore u našem radu.

3.3 Analysis of Google Play Store Data set and predict the popularity of an app [pdf]

Tema rada: Rad se bavi analizom faktora koji utiču na popularnost aplikacija na Google Play prodavnici i razvojem prediktivnih modela za procenu rejtinga. Fokus je na razumevanju korelacije između metapodataka poput cene i kategorije i finalnog uspeha kod korisnika.

Podaci: Za istraživanje je korišćen Google Play Store skup podataka koji sadrži informacije o preko 10.000 aplikacija, uključujući broj recenzija, veličinu i uzrasnu kategoriju.

Metodologija: Autori su primenili proces čišćenja podataka i inženjeringa atributa, nakon čega je korišćen Random Forest Regressor za rešavanje problema regresije. Model je treniran da predvidi kontinuiranu vrednost ocene aplikacije.

Evaluacija rešenja: Preciznost modela je evaluirana upotrebom standardnih regresionih metrika MAE (Mean Absolute Error) i RMSE (Root Mean Square Error).

Zaključak: Rezultati su pokazali da Random Forest model pruža najveću preciznost. Naš projekat će se osloniti na ovaj rad pri izboru metrika evaluacije i primeni Random Forest modela kao ključnog alata za predikciju.

4. Skup podataka

Podaci su preuzeti sa Kaggle platforme iz skupa pod nazivom „Google Play Store Apps“ koji sadrži informacije o preko dva miliona aplikacija. Rezultujući skup podataka nakon čišćenja sadržaće sledeće atribute: kategorija (npr. Productivity, Games), veličina aplikacije (u megabajtima), cena, tip (besplatna ili plaćena), uzrasna grupa (Content Rating), prisustvo reklama (Ad Supported), opcija kupovine unutar aplikacije (In-App Purchases) i verzija Androida koju aplikacija zahteva. Ciljno obeležje je Rating, odnosno prosečna ocena korisnika koju model treba da predvidi kao kontinuiranu vrednost.

5. Metodologija

Nakon prikupljanja podataka, analiziraćemo da li postoje trendovi u uspešnosti aplikacija u zavisnosti od širokog spektra faktora. Ulaz u algoritme neće biti ograničen samo na kategoriju i veličinu, već će obuhvatiti i ekonomske faktore (cena i prisustvo reklama), kao i ciljanu demografiju. Metodologija uključuje detaljno pretprocesiranje gde ćemo numeričke vrednosti normalizovati, dok ćemo kategoričke varijable transformisati u format pogodan za mašinsko učenje. Centralni deo metodologije predstavlja razvoj sopstvene arhitekture Veštačke neuronske mreže (ANN) zasnovane na fundamentalnim matričnim operacijama, bez oslanjanja na gotove biblioteke visokog nivoa. Implementiraćemo potpuno povezan višeslojni perceptron sa ručno definisanim algoritmima propagacije unapred i unazad, kako bismo modelovali složene nelinearne veze u podacima. Radi validacije ovog pristupa, performanse naše mreže uporedićemo sa standardnim modelima kao što su Linear Regression, k-Nearest Neighbors (k-NN) i Random Forest Regressor.

6. Metod evaluacije

Za evaluaciju predikcije ocene aplikacije koristiće se metrike MAE (Mean Absolute Error) i RMSE (Root Mean Square Error), gde ćemo računati prosečno odstupanje predviđene ocene od stvarne ocene zabeležene u bazi podataka. Podelu podataka izvršićemo u razmeri 80:10:10, gde ćemo 80% podataka koristiti za trening, 10% za validaciju parametara modela i preostalih 10% za finalno testiranje. Ključni deo evaluacije biće analiza stabilnosti naše neuronske mreže na neviđenim podacima i njeno direktno poređenje sa osnovnim modelima, čime ćemo osigurati objektivnu meru kvaliteta predloženog rešenja.