Google Analytics Customer Revenue Prediction

1. Uvodni opis problema

1.1. Uvod u problem

Enciklopedija 21. stoljeća je Google. Ljudi ga koriste za izvor gotovo svih informacija. Počevši od povijesnih, preko geografskih pa sve do aktualnih u raznim sferama, ali i za svakodnevnu upotrebu i zabavu. Nije čudno da se zato Google proslavljuje iz dana u dan posjetima njihovim stranicama. Tvrtka želi pružiti kupcima svoj brand kako bi još više izazvala interes za svoje proizvode i promovirala se. U cilju je tvrtke imati što veću potražnju i prihod od prodaje proizvoda 's potpisom'. Njih prodaje u Google trgovini- tzv. 'Google Merchandise store'. Budući da za većinu tvrtki vrijedi da samo mali postotak kupaca (oko 20%) donosi do čak 80% prihoda, želimo napraviti model koji bi precizno predviđao buduće vrijednosti prihoda po kupcu na temelju raznih podataka o kupcima. Tako bi Google trgovina i njihovi marketinški djelatnici mogli znati na koji način i u kojem smjeru (što se tiče kupaca) ulagati u promotivne strategije.

1.2. Skup podataka

Ovaj je problem bio predstavljen u obliku Kaggle natjecanja dostupnog na linku ovdje, stoga ćemo naš projekt bazirati upravo na strategijama i metodama koje su se koristile u pobjedničkom rješenju. Za rješavanje ovog problema koristimo skup podataka train_v2.csv i test_v2.csv. Dokument train_v2.csv sadrži transakcije korisnika u vremenskom razdoblju od 1. kolovoza 2016. godine do 30. travnja 2018. godine, a u dokumentu test_v2.csv nalaze se korisničke transakcije od 1. svibnja 2018. godine do 15. listopada 2018. godine. Podaci su organizirani tako da sadrže stupce koji se nalaze ispod Data Fields (atributa). Svaki redak u dataset-u predstavlja jedan posjet trgovini.

Zadani atributi:

fullVisitorId –jedinstveni identifikator za svakog korisnika Google Merchandise Store-a.
channelGrouping –kanal preko kojeg je korisnik došao do trgovine.
date –datum na koji je korisnik posjetio trgovinu.
device –specifikacije uređaja kojim je korisnik pristupio trgovini.
geoNetwork –informacije o geografiji korisnika.
socialEngagementType –tip društvene angažiranosti, može biti ili "Socially Engaged" ili "Not Socially Engaged".
totals –skup stupaca koji sadrži agregirane vrijednosti tijekom sesije.
trafficSource –sadrži informacije o izvoru prometa s kojeg je sesija nastala.
visitId –identifikator za trenutnu sesiju. Dio vrijednosti koja je obično pohranjena kao _utmb cookie. Jedinstven samo korisniku, a za potpuno jedinstveni ID, trebalo bi koristiti kombinaciju fullVisitorId i visitId.

visitNumber –broj sesije za trenutnog korisnika. Ukoliko je ovo prva sesija, postavlja se na 1.
visitStartTime –vremenska oznaka (izražena kao POSIX vrijeme).
hits –ovaj su redak i ugnježđena polja ispunjeni za bilokakve tipove "hit-ova" (to obuhvaća sve podatke koji se učitavaju na web stranici poput slika, gumbova, i slično). Pruža zapis svih posjeta stranici.
customDimensions –ova sekcija sadrži sve prilagođene dimenzije na razini korisnika ili sesije, koji se postavljaju za sesiju. Ovo je ponavljajuće polje i ima ulaz za svaku postavljenu dimenziju.

Postoji više stupaca koji sadrže JSON dokumente različitih detaljnosti. U jednom od takvih JSON stupaca, *totals*, jedan od podstupaca transactionRevenue sadrži informacije o prihodu koje pokušavamo predvidjeti. Ovaj podstupac postoji samo u training dataset-u. Budući da je više atributa u JSON dokumentu konstantno i/ili ne utječu na rezultat prediktivnog modela, pri izradi modela ćemo prvo raditi data engineering pa ćemo takve atribute maknuti iz zadanog csv-a. Bez prethodnog micanja atributa, u train_v2.csv dokumentu imamo na raspolaganju 1 708 337 redaka, a u test_v2.csv 401 589 redaka, odnosno posjeta trgovini koje analiziramo. Kad se stupci koji su u JSON formatu raspakiraju, dobijemo 60 stupaca u train setu, što uz 1 708 337 redaka u tom setu čini data frame od ukupno 102 500 220 podataka.

2. Cilj i hipoteze istraživanja problema

Cilj istraživanja je napraviti model koji bi predviđao prihod od svakog pojedinog kupca. Ciljna funkcija će stoga biti prirodni logaritam zbroja transakcija po korisniku, tj. :

$$target_{user} = \ln(\sum_{i=1}^{n} transaction_{user_i} + 1)$$

Dakle, za svaki jedinstveni ID u skupu test_v2.csv ćemo napraviti predviđanje.

Neke hipoteze koje se nameću su jesu li pojedini atributi značajni za model ili ne. Neki atributi intuitivno nas navode na sljedeće hipoteze:

- atribut '*date*' je značajan u prediktivnom modelu (zbog pretpostavke o povećanju kupovine u vrijeme blagdana)
- atribut '*date*' nije značajan u modelu
- atribut '*geoNetwork*' je značajan u prediktivnom modelu (pretpostavljamo da bi najveći trošak mogao biti kod korisnika iz Sjeverne Amerike, budući da je tamo sjedište trgovine)
- atribut '*geoNetwork*' nije značajan u modelu

Također, hipoteza je činjenica koja vrijedi u većini tvrtki, a to je da tek 20% korisnika čini prihod od 80%.

3. Pregled dosadašnjih istraživanja

Bazirat ćemo se na pobjedničkom rješenju Konstantina Nikolaeva na linku ovdje. On je koristio linearnu regresiju kao glavnu metodu učenja modela, ali ne direktno na dobivenim podacima train_v2.csv. Većina drugih natjecatelja koristila je samo linearnu regresiju. Način na koji je Nikolaev poboljšao rezultate, odnosno nedostatke same regresije, je klasifikacija skupa podataka train_v2.csv. Greška u nekim istraživanjima je i način na koji su natjecatelji dijelili naveden skup. Nikolaev je pregrupirao train skup i predvidio vjerojatnost vraćanja pojedinog kupca u Google trgovinu. Zatim je predvidio iznos transakcija za one kupce koji su se vratili.

4. Materijali, metodologija i plan istraživanja

Problem ćemo pokušati riješiti pozivajući se na gore navedeno rješenje Konstantina Nikolaeva. Dakle, nećemo analizirati direktno metodom linearne regresije, nego ćemo prvo koristiti klasifikaciju za train skup. Podaci se nalaze na Kaggleovoj stranici, na linku ovdje. Na temelju vizualne analize podataka, očekujemo da će već biti moguće vidjeti istinitost hipoteze o 80%/20% pravilu, a i koliko su značajni pojedini atributi. Alati koje ćemo koristiti su jezici Python i/ili R u Jupyter notebooku. Pokušat ćemo u treniranju modela za klasifikaciju i regresiju, kao i Nikolaev, koristiti jednu od boosting metoda, a to je tzv. 'light GBM' algoritam. Preciznost dobivenog modela bit će mjerilo uspješnosti, a ona će se mjeriti pomoću root means square error (RMSE), što je često korištena mjera razlika između vrijednosti koje predviđa model i promatrane vrijednosti-opservacije.

5. Očekivani rezultati predloženog projekta

Očekujemo da će konačni rezultat našeg projekta dati precizno predviđanje troškova kupaca u određenom periodu te da ćemo na temelju rezultata moći potvrditi ili opovrgnuti hipoteze postavljene u ovom prijedlogu, tj. da će nam model omogućiti da analiziramo koji atributi zaista utječu na troškove, a koji imaju zanemariv utjecaj. Ovakav prediktivni model zasigurno je od velike koristi kompanijama poput Google-a koje su orijentirane na online trgovinu i koje žele maksimalno iskoristiti marketinške alate koji su im na raspolaganju kako bi povećale svoje prihode.

6. Popis literature

Kaggle natjecanje: https://www.kaggle.com/c/ga-customer-revenue-prediction/overview
Pobjedničko rješenje Kaggle natjecanja: https://www.kaggle.com/c/ga-customer-revenue-prediction/discussion/82614
The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman
Razne (Jupyter notebook) bilježnice, dostupne na Kaggleu:
https://www.kaggle.com/julian3833/1-quick-start-read-csv-and-flatten-json-fields
https://www.kaggle.com/ogrellier/create-extracted-json-fields-dataset
https://www.kaggle.com/sudalairajkumar/simple-exploration-baseline-ga-customer-revenue
https://www.kaggle.com/shivamb/exploratory-analysis-ga-customer-revenue
https://www.kaggle.com/kailex/group-xgb-for-gstore-v2
https://www.kaggle.com/julian3833/2-quick-study-lgbm-xgb-and-catboost-lb-1-66
https://www.kaggle.com/smasar/tutorial-preprocessing-processing-evaluation#4-BASIC-REGRESSION
Online tečaj: https://www.coursera.org/learn/machine-learning Andrew Ng
Modeling Caries Experience: Advantages of the Use of the Hurdle Model:
http://www.elisedusseldorp.nl/pdf/HofstetterDusseldorpetal2016.pdf
End-to-end Machine Learning project: https://github.com/ageron/handson-ml/blob/master/02_end_to_end_machine_learning_project.ipynb