Predviđanje da li se isplati odobriti zaposlenom u Amazonu pristup resursu

Autor: Igor Ljucović

Datum: 8.9.2025.

1. Uvod

Amazonov **dataset sadrži 32.769 opservacija (redova) i 10 kolona** od kojih je 1 ciljna kolona (**"ACTION" - vrednost koju treba predvideti**, da li dozvoliti zaposlenom pristup resursu ili ne), 1 kolona je ID samog resursa, 1 kolona je ID menadžera zaposlenog, a ostalih 7 kolona su informacije vezane za samog zaposlenog.

Napomena: ***Sve promenljive su predstavljene kao ID-evi, bez semantičkog značenja.*** *To je Amazon uradio zbog privatnosti podataka (na primer, svako pojavljivanje za `ROLE_TITLE` koje je `project_manager` će biti zamenjeno odgovarajućim ID-em `176508`).*

Cilj je da se, na osnovu istorijskih podataka o zahtevima zaposlenih za pristup određenim resursima, predvidi da li će novi zahtev (pristup resursu, npr. nekom veb sajtu) biti odobren (1) ili odbijen (0).

Zaposleni često traže pristup različitim aplikacijama, najbolje bi bilo obezbediti im pristup samo onim aplikacijama koje će im povećati produktivnost na poslu jer:

1. **Pogrešno odobravanje resursa može dovesti do sigurnosnih propusta ili do odugovlačenja umesto završavanja posla**
2. **Pogrešno odbijanje može smanjiti produktivnost zaposlenih**

Zbog velikog broja zahteva, ručna evaluacija nije skalabilna. Ako se pogrešno odobri/zabrani pristup, izgubiće se dosta vremena dok menadžer, odnosno odgovorni zaposleni, odobri/zabrani pristup resursu usred radnog vremena, što nije efikasno. Zato se koriste algoritmi mašinskog učenja koji na osnovu prethodnih primera uče obrasce i pomažu u automatskom odlučivanju. Pre upotrebe modela je potrebno adekvatno pripremiti podatke, a nakon pokretanja modela je potrebno evaluirati njegove performanse kako bismo mogli da dođemo do zaključka.

2. Priprema podataka

Kolona “ROLE_CODE” sadrži iste informacije kao i kolona “ROLE_TITLE”, tako da možemo smanjiti dimenzionalnost problema za 1 kolonu tako što **uklonimo kolonu “ROLE_CODE”**.

Kreiranje manjeg podskupa dataset-a nije bilo potrebno iz razloga što su modeli nisu imali problema sa vremenskom efikasnošću pri radu sa Amazonovim dataset-om od oko 32.000 redova.

Normalizacija podataka (npr. skaliranje vrednosti na opseg 0–1) **nije bila potrebna** iz razloga što su sve kolone kategorijskog tipa, odnosno nisu brojčane. Međutim, bilo je potrebno uraditi **one hot encoding pre upotrebe MLP** (neuronske mreže).

Otklanjanje redova koji odstupaju po nekoj karakteristici (**outlieri**) jeste bilo potrebno iz razloga što su u kolonama **“ROLE_TITLE” i “ROLE_FAMILY_DESC”** postojale određene vrednosti koje su se pojavljivale na samo 1, 2 ili 3 mesta i one su predstavljale šum, odnosno, nije postojalo dovoljno opservacija (redova) kako bi modeli mogli da dođu do korisnih zaključaka za predviđanje na osnovu njih. Iz Amazonovog dataseta je uklonjeno 0.23% redova jer imaju outlier vrednosti za “ROLE_TITLE” kolonu, od 343 mogućih vrednosti je smanjeno na 290, odnosno, uklonjeno je 53 najređe pojavljivanih vrednosti za tu kolonu. Zatim smo od preostalih redova uklonili čak 12.0362% redova jer imaju outlier vrednosti za „ROLE_FAMILY_DESC“ kolonu, od 2358 mogućih vrednosti je smanjeno na 700, odnosno, uklonjeno je 1658 najređe pojavljivanih vrednosti za tu kolonu. 12.03% jeste veliki broj uklonjenih kolona, ali ova promena je pozitivno uticala na performanse svih modela, što i ima smisla, jer kolone sa velikom kardinalnošću (velikim brojem mogućih vrednosti kod kategorijskim promenljivih, odnosno kolona) mogu da prouzrokuju da model uči iz šuma, ako postoji samo 1 do 3 reda sa određenom vrednošću, model neće imati dovoljno podataka kako bih mogao da dođe do zaključaka na osnovu tih redova, već će ga oni samo zbuniti, i zato smo ih i uklonili. Još jedna kolona koja ima jako veliku kardinalnost, zapravo najveću od svih kolona, je “MGR_ID” sa čak 2540 različitih vrednosti u preostalih 28.000 redova. Iz te kolone se zapravo nije isplatilo izbacivati outlier-e, odnosno modeli bi imali lošije performanse nakon izbacivanja tih outlier-a, jer je “MGR_ID” u velikom broju modela najkorisnija, ili 1 od najkorisnijih kolona za donošenje zaključaka, a osim toga, ne postoji nijedan menadžer koji se pojavljuje znatno češće ili ređe nego drugi, tako da je model imao mnogo manji broj vrednosti koje se previše retko pojavljuju da bi se nešto iz njih naučilo.

3. Modeli

Razmatrano je 7 modela mašinskog učenja, od kojih su se 3 modela loše pokazala, 1 model osrednje i 3 modela odlično, u zavisnosti od toga koje parametre najviše cenimo pri predviđanju. S obzirom da se naš dataset sastoji od isključivo kategorijskih promenljivih, modeli koji mogu da rade sa kvalitativnim (kategorijskim) promenljivama će se bolje pokazati (modeli koji koriste šume i/ili boosting) od onih koji najbolje rade samo sa kvantitativnim (kontinualnim, numeričkim) promenljivama (kao što je linearna regresija).

Parametri modela su bili podešavani tako da se optimizuje vrednost metrika, gde se najviše gledalo da se maksimizira ROC-AUC vrednost jer se rečeno da je ona najvažnija na Amazonovoj stranici za ovaj dataset na Kaggle-u.

Moć predviđanja	Model	Model	Model	Komentar
Loša	Logistic Regression			Previše jednostavan
Osrednja	Cat Boost	Neural Network	Logistic Regression with Random Forest stacking	Ostali modeli su bolji u barem nekim aspektima
Odlična	Random Forest	Decision Tree	Light GBM	Potencijalno najbolji model u zavisnosti od naših potreba

4. Evaluacija

Main metrics:

1. **ROC-AUC** – koliko model generalno bolje predviđa od najjednostavnijeg mogućeg modela koji uvek predviđa najčešći mogući slučaj (da treba dati pristup resursu)
2. **PRECISION** - od svih zaposlenih kojima je model dodelio pristup, koliko njih je zaista trebalo da dobije pristup
3. **SPECIFICITY** - od svih zaposlenih kojima nije trebalo dati pristup, koliko ih je model tačno odbio
4. **RECALL** - od svih zaposlenih kojima je trebalo dati pristup, koliko ih je model zaista prepoznao.
5. **F1** - Kombinacija preciznosti (Precision-a) i osetljivosti (Recall-a), ako je bar 1 od te 2 metrike niska, biće niska i ova metrika
6. **TIME** - vreme (u sekundama) potrebno da se model istrenira i da odluči o pristupu resursu zaposlenom
7. **OVERALL RANK** - Koji je model po redu ako svih 6 metrika posmatramo kao jednako važne gde model dobija 1 poen za metriku ako je najgori od svih 5 modela, a 5 poena za metriku ako je najbolji od svih 5 modela u njoj (najbolji model je rank 1)
8. **SUBJECTIVE RANK** - Koji model je najbolji po mom mišljenju, najviše sam vrednovao Specificity, odnosno moć modela da što tačno predvideti slučajeve u kojima zaposlenom nije trebalo dati pristup resursu, jer, za razliku od slučajeva kada jeste bilo potrebno dati pristup, ovde osim potencijalno smanjene produktivnosti imamo i bezbednosni rizik.

Detaljan prikaz performansi 5 najboljih modela:

Model	ROC-AUC	PRECISION	SPECIFICITY	RECALL	F1	TIME	OVERALL RANK	SUBJECTIVE RANK
Random Forest	0.880	0.965	0.439	0.985	0.975	5.64	1 (25p)	3
Decision Tree	0.809	0.755	0.732	0.757	0.853	0.20	5 (14p)	1
Light GBM	0.875	0.937	0.646	0.954	0.965	2.68	4 (17p)	2
Cat Boost	0.874	0.940	0.608	0.960	0.967	13.99	2 (18p)	4
Neural Network	0.879	0.956	0.281	0.960	0.967	64.90	2 (18p)	5

5. Zaključak

Ne postoji objektivno najbolji model za ovaj dataset, već:

1. Ako je najvažnije da što ređe date resurs nekome kome nije trebalo da ga date (zbog bezbednosti ili smanjene produktivnosti) - **Decision Tree**
2. Ako je najvažnije da što ređe ne dozvolite pristup resursu nekome kome jeste trebao (zbog smanjene produktivnosti) - **Random Forest**
3. Ako 1 od kriterijuma nije značajno važniji od drugog - **Light GBM**

Modelima su za predviđanje najviše od koristi bile kolone “RESOURCE”, “MGR_ID” i ”ROLE_DEPTNAME”, dok su im najmanje bitne kolone bile “ROLE_ROLLUP_1” i “ROLE_FAMILY”. U tabeli ispod je prikazano 5 najkorisnijih kolona za svaki od najboljih modela za naš dataset:

Model	1. najvažnija kolona	2. najvažnija kolona	3. najvažnije kolona	4. najvažnija kolona	5. najvažnija kolona
Random Forest	RESOURCE	MGR_ID	ROLE_DEPTNAME	ROLE_FAMILY_DESC	ROLE_ROLLUP_2
Decision Tree	MGR_ID	ROLE_DEPTNAME	ROLE_FAMILY_DESC	RESOURCE	ROLE_ROLLUP_2
Light GBM	RESOURCE	MGR_ID	ROLE_DEPTNAME	ROLE_FAMILY_DESC	ROLE_TITLE