Unaprjeđenje poslovanja strojnim učenjem

Radečić, Dario

LIBERTAS MEĐUNARODNO SVEUČILIŠTE

ZAGREB

DARIO RADEČIĆ

ZAVRŠNI RAD

UNAPRJEĐENJE POSLOVANJA STROJNIM UČENJEM

Zagreb, rujan 2018.

LIBERTAS MEĐUNARODNO SVEUČILIŠTE

ZAGREB

PREDDIPLOMSKI STRUČNI STUDIJ

POSLOVNA EKONOMIJA

UNAPRJEĐENJE POSLOVANJA STROJNIM UČENJEM

KANDIDAT: Dario Radečić

KOLEGIJ: Poslovna informatika

MENTOR: Mihael Plećaš, mag. ing.

Zagreb, rujan 2018.

SADRŽAJ

1. UVOD 1
1.1. Problem i predmet rada 1
1.2. Ciljevi i svrha rada 1
1.3. Istraživačka pitanja 2
1.4. Izvori podataka i metodologija u radu 2
1.5. Struktura i sadržaj rada 2
2. UVOD U STROJNO UČENJE 4
2.1. Definicije i primjena strojnog učenja 5
2.2. Klasifikacije strojnog učenja 5
2.2.1. Učenje uz nadzor 5
2.2.2. Učenje bez nadzora 6
2.2.3. Pojačano učenje 7
3. MATEMATIČKA I STATISTIČKA OSNOVA STROJNOG UČENJA 8
3.1. Matematika 8
3.1.1. Linearna algebra 9
3.1.2. Diferencijalni i integralni račun više varijabli 10
3.2. Statistika 11
3.3. Vjerojatnost 12
4. STROJNO UČENJE NA PRIMJERU 13
4.1. Istraživanje i priprema podataka 14
4.2. Izrada sustava preporuka 20
4.3. Verifikacija istraživačkih pitanja 27
5. ZAKLJUČAK 27
LITERATURA 28
POPIS SLIKA 29
POPIS TABLICA 30
POPIS GRAFIKONA 31

1. UVOD

Strojno učenje je već godinama središte fokusa najvećih tehnoloških kompanija te se ne odnosi samo uobičajenu percepciju umjetne inteligencije (robotika i slično), nego se iz dana u dan sve više primjenjuje u poslovanju tvrtki koje svojim kupcima žele ponuditi pravi proizvod u pravo vrijeme s ciljem maksimiziranja zarade. Utjecaj strojnog učenja može se primijetiti na web tražilicama, web stranicama tvrtki (online trgovine), reklamama u stvarnom vremenu, e-pošti (automatsko filtriranje neželjenih poruka) i na još mnogo mjesta. Cilj rada je objasniti što je strojno učenje, navesti vrste i zahtjeve te praktičnim primjerom pokazati kako tvrtke njegovom primjenom mogu povećati dobit.

1.1. Problem i predmet rada

Problem ovog rada je financijski rezultat tvrtke, odnosno njegovo unaprjeđenje primjenom jedne od metoda strojnog učenja – sustava preporuka. Suvremeni uvjeti poslovanja i konstantni napredak tehnologije tvrtkama omogućuju da svom klijentu ponude pravi proizvod brže i bolje od konkurencije, ili čak da klijentu prodaju proizvod koji od prethodno nije ni trebao.

Predmet rada je korištenje sustava preporuka u internetskoj trgovini. Rad objašnjava kako tvrtke modernim metodama strojnog učenja mogu doći do novih kupaca te kako mogu povećati lojalnost stalnih kupaca.

1.2. Ciljevi i svrha rada

Ciljevi rada mogu se navesti kako slijedi:

- Objasniti strojno učenje
- Objasniti primjenu strojnog učenja u svakodnevnom životu
- Objasniti kako tvrtke mogu unaprijediti poslovanje primjenom metoda strojnog učenja
- Objasniti potencijalne opasnosti strojnog učenja

Svrha ovog rada je ukazati na sve veću važnost strojnog učenja, kako za pojedinca, tako i za poduzeća te objasniti zašto i kako su poduzeća koja u svom poslovanju primjenjuju metode strojnog učenja u konačnici uspješnija od onih koji ih ne primjenjuju.

1.3. Istraživačka pitanja

Temeljem navedenog problema i predmeta rada, odnosno njegovog cilja u radu se postavljaju dva istraživačka pitanja:

IP1: Što su sustavi preporuka?

IP2: Kako tvrtke mogu koristiti sustave preporuka kako bi unaprijedile poslovanje?

1.4. Izvori podataka i metodologija u radu

Teorijska poglavlja koja se odnose na uvod u strojno učenje, njegove predzahtjeve i klasifikaciju izrađena su pomoću knjiga, znanstvenih i stručnih članaka te internetskih stranica. Istraživačko poglavlje, odnosno poglavlje s primjenom u praksi izrađeno je prema *MovieLens* skupu podataka o stvarnim ocjenama i sviđanjima različitih korisnika za različite filmove.
U izradi teoretskih poglavlja primijenjena je metoda deskripcije, klasifikacije i komparacije te metode statističke analize u praktičnom dijelu rada. Statistička obrada podataka izvedena je korištenjem biblioteke *Pandas* u programskom jeziku *Python*.

1.5. Struktura i sadržaj rada

Rad je strukturiran na sljedeći način. U idućem, drugom poglavlju govorit će se općenito o strojnom učenju, njegovoj definiciji i podjeli, razvoju definicije kroz vrijeme, razlici između učenja uz nadzor, učenja bez nadzora i pojačanog učenja te za svaki navesti nekoliko jednostavnih primjera kako bi se koncept približio čitatelju.

Treće poglavlje govori o pred zahtjevima strojnog učenja - matematici i statistici. Rad ne objašnjava detaljno svaku granu matematike čije je razumijevanje potrebno za razumijevanje strojnog učenja, nego u nekoliko rečenica pojašnjava koji dijelovi iz područja linearne algebre, diferencijalnog i integralnog računa više varijabli, statistike i vjerojatnosti se koriste pri izradi modela strojnog učenja. Nadalje, treće poglavlje govori ukratko o potrebnim znanjima programskih jezika i jezika za rad s bazama podataka koji su potrebni za strojno učenje.

Četvrto poglavlje je glavni dio rada - obrada strojnog učenja na primjeru. Rad nastoji riješiti sljedeću problematiku: kako tvrtke mogu primijeniti strojno učenje te s njim povećati zaradu i

privući nove klijente. Za primjer će biti uzet *MovieLens* skup podataka o sviđanjima i ocjenama filmova od strane korisnika. Primjer je napravljen s tim skupom podataka zbog njegove lake dostupnosti, ali isti kod može se primijeniti na bilo koji drugi skup podataka, ovisno o tome čime se tvrtka bavi.

Na kraju rada nalazi se zaključak koji će ukratko sažeti sve navedeno te reći nešto o budućnosti strojnog učenja, kako će promijeniti svijet. Fokus zaključka je objasniti naprednu primjenu strojnog učenja, odnosno automobile koji voze sami, primjenu u robotici i slično. Zaključak također objašnjava opasnosti napredne primjene strojnog učenja koje može, a zasigurno u nekoj mjeri i bude dovelo do automatizacije određenih poslova.

2. UVOD U STROJNO UČENJE

Još od vremena nastanka računala, znanstvenike i inženjere je zanimalo može li ih se programirati da sami uče. Ukoliko bi takav oblik programiranja računala bio moguć, neki od najvećih svjetskih problema bi bili riješeni. Na primjer, strojno učenje može riješiti mnoge probleme u zdravstvu – računalo može učiti iz povijesnih zapisa i na temelju njih zaključiti koji oblik tretmana koristiti za svakog pojedinog pacijenta. Nadalje, kuće bi bile u stanju učiti prema uzorcima potrošnje svojih stanovnika te tako proizvesti cjenovno optimalne načine potrošnje energije. Ovdje se pojavljuje jedno ključno pitanje – što je zapravo strojno učenje?

2.1. Definicije i primjena strojnog učenja

Postoji mnogo definicija, Drew Conway i John Myles White navode slijedeće: „Možemo razmišljati o strojnom učenju kao o skupu alata i metoda koji pokušavaju prepoznati i izvući uzorke iz promatranih podataka“[1]. Drugu definiciju strojnog učenja iznosi Tom Mitchell koji kaže: „Računalni program uči iz iskustva **E** prema skupu zadataka **T** i mjerama performansi **P**. Ako su performanse programa T mjerene s P, one će se poboljšati kroz iskustvo E“[2].

Mitchellovu definiciju isprva je malo teže shvatiti, ali u suštini je vrlo jednostavna, što dokazuje slijedeći primjer (igra šaha):

E = iskustvo nastalo igranjem velikog broja igre
T = zadatak igranja igre
P = vjerojatnost da će računalni program pobijediti u slijedećoj igri.

Cilj rada nije detaljno objasniti sve moguće primjene strojnog učenja, ali valja navesti one najčešće:

1. Detektiranje prevare (transakcije karticama)
2. Rezultati pretraživanja interneta
3. Reklame u stvarnom vremenu na web stranicama
4. Predviđanja otkazivanja opreme
5. Izrada modela cijena
6. Sustavi preporuka
7. Segmentacija kupaca
8. Analiza teksta

[1] Conway, D., White, Myles J.: *Machine Learning for Hackers*, O'Reilly, 2012., str. 7 - preface
[2] Mitchell, T.: *Machine Learning*, McGaw-Hill, 1997., str 2

9. Pronalaženje uzoraka u fotografijama
10. Odvajanje neželjene elektronske pošte (spam)

Naravno, gore navedena lista primjene strojnog učenja nije finalna te prikazuje samo mali dio onoga što strojno učenje može napraviti. Fokus rada biti će na praktičnu primjenu strojnog učenja u poslovanju izmišljene tvrtke, popraćeno kratkim, jasnim i sažetim teorijskim uvodom. Navedena izmišljena tvrtka bavi se recenzijama filmova i TV serija te dopušta korisniku izradu vlastitog računa i ocjenjivanja gore navedenih filmova i TV serija. Ovdje će zadatak strojnog učenja biti prikupiti sve podatke o ocjenama korisnika te na temelju njegovih preferencija preporučiti filmove drugim korisnicima koji imaju iste ili slične preferencije.

Kako ne bi imalo smisla rad započeti s primjenom navedenog sustava preporuka, najviše zbog (ne)razumijevanja sustava, prva stvar koju rad pokriva je kratki uvodi u matematičku i statističku podlogu potrebnu za razumijevanje sadržaja. Nadalje, rad će prikazati kako se primjenom modernih programskih jezika uvelike olakšava matematička obrada podataka te objašnjava ulogu programskih jezika u primjeni strojnog učenja. Tek nakon što je sve gore navedeno obrađeno može se krenuti u dublje istraživanje sustava preporuka i njegovu primjenu u praksi.

2.2. Klasifikacije strojnog učenja

U pravilu, bilo koji program koji može biti riješen strojnim učenjem spada u jednu od tri klasifikacije učenja:

1. Učenje uz nadzor
2. Učenje bez nadzora
3. Pojačano učenje

2.2.1. Učenje uz nadzor

Kod učenja uz nadzor imamo dostupan skup podataka za kojeg unaprijed znamo kako rezultat treba izgledati. Također, poznata nam je veza između skupa podataka (ulaza) i rezultata (izlaza).

Problemi učenja uz nadzor kategorizirani su u dvije skupine – regresijski i klasifikacijski problemi. U regresijskom problemu ulazne varijable želimo provesti kroz neku funkciju, dok u klasifikacijskom problemu varijable ulaza stavljamo u diskretne kategorije.[3]

[3] *Supervised Learning*, Machine Learning online course, Coursera – Stanford University, https://www.coursera.org/learn/machine-learning, pristupano 14.06.2018.

Primjer 1 – prema zadanim podacima o veličini kuća na tržištu nekretnina, predvidi njihovu cijenu.

Primjer 1 je klasični primjer regresijskog problema – do rezultata se može doći samo ako varijablu ulaza (veličinu kuće) stavimo u neku funkciju. Primjer možemo pretvoriti u klasifikacijski problem ako postavimo pitanje hoće li se kuća prodati za manje ili više od cijene koju kupac traži. U tom slučaju klasificiramo kuće unutar dvije diskretne kategorije.

Primjer 2 – regresijski problem

Računalu je zadana fotografija osobe te prema njoj mora predvidjeti koliko godina osoba ima.

Primjer 3 – klasifikacijski problem

Računalo na temelju fotografije tumora mora predvidjeti je li tumor dobroćudan ili zloćudan.

2.2.2. Učenje bez nadzora

Učenje bez nadzora omogućuje pristupanje problemu za kojeg ne znamo kako rezultat treba izgledati. Izvlačenje strukture iz podataka je moguće, ali nije nam poznat utjecaj jedne varijable na drugu. Također, možemo izvući strukturu podataka grupiranjem podataka ovisno o vezi između pojedinih varijabli.[4]

Primjer 1 – grupiranje

Računalu je dan skup podataka o 1.000.000 različitih gena te ih ono automatski razvrstava u grupe koje su na neki način slične ili povezane različitim varijablama (uloga, lokacija…).

Primjer 2 – negrupiranje

Takozvanim *algoritmom zabave (eng. Cocktail party algorithm)* računalo iz zvučnih zapisa može prepoznati glasove te tako odvojiti glasove ljudi od pozadinske glazbe.

2.2.3. Pojačano učenje

Pojačano učenje uglavnom se koristi za robotiku, video igre i navigaciju. Pomoću njega algoritam metodom pokušaja i pogrešaka otkriva koja akcija donosi najveću nagradu.[5]

[4] *Unsupervised Learning*, Machine Learning online course, Coursera – Stanford University, https://www.coursera.org/learn/machine-learning, pristupano 14.06.2018.
[5] *Reinforcement Learning*, Machine Learning online course, Coursera – Stanford University, https://www.coursera.org/learn/machine-learning, pristupano 14.06.2018.

Ovaj tip učenja ima 3 glavne komponente:

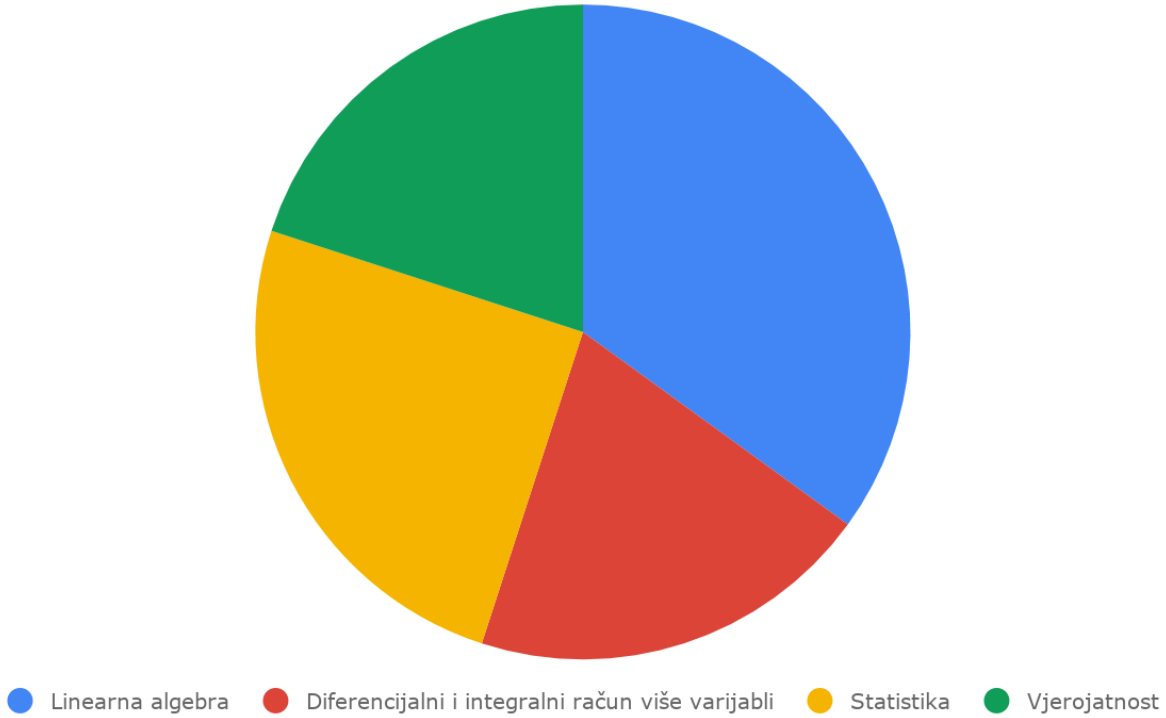
1. Agenta (onaj koji donosi odluku)
    - zadatak agenta je odabrati akcije koje maksimiziraju očekivanu nagradu nakon određenog, zadanog vremena
    - agent će doći do svog cilja puno brže ukoliko će zadatke obavljati u dobrom smjeru
2. Okruženje (sve sa čime agent interaktira)
3. Akcija (sve što agent može napraviti)

Dakle, zadatak pojačanog učenja je pronaći najbolji smjer ili način obavljanja zadataka.

3. MATEMATIČKA I STATISTIČKA OSNOVA STROJNOG UČENJA

Kao i svaka grana računarstva, strojno učenje zahtjeva razumijevanje osnova računalstva, ali kako se ono primarno zasniva na matematici i statistici, razumijevanje tih predmeta je potrebno svakoj osobi koja se želi baviti ovom granom računalstva. Što sve pojedinac treba znati navedeno je u sljedećem prikazu.

Grafikon 1. Zahtjevi strojnog učenja



Izvor: Sistematizacija autora

Naravno, netko tko se želi baviti ovim područjem ne mora nužno biti ekspert u gore navedenim područjima, ali ih mora dovoljno dobro poznavati iz razloga što zanimanje često zahtjeva izradu novih algoritama za rješavanje određenog problema, a ti algoritmi se uvelike oslanjaju na odlično poznavanje matematike. Nadalje, vrlo je važno dobro poznavanje programiranja, naročito u jednom od dva jezika koja se koriste u ovoj grani (*Python* ili *R*) te *SQL* zbog rada s bazama podataka u kojima su pohranjeni podaci iz kojih će računalo učiti.

3.1. Matematika

Kao što se iz prikaza 1 može vidjeti, matematika čini više od 50% potrebnih predznanja za strojno učenje. Dva najvažnija područja matematike su linearna algebra i multivarijabilni

račun. Iz linearne algebre strojno učenje posuđuje koncepte matrica i vektora, npr. podatke iz klasične Excel tablice može se vrlo lako prikazati u obliku matrice. Multivarijabilni račun koristi se za optimizaciju modela, najčešće algoritmom zvanim spuštanje po gradijentu (eng. gradient descent).

3.1.1. Linearna algebra

Linearna algebra je područje matematike koje se bavi vektorima, matricama i linearnim transformacijama. Kao što je već gore navedeno, služi za izradu modela te prezentaciju podataka računalu iz kojih ono može prepoznati uzorke u podacima. Osobi koja to po prvi put čuje može biti nejasno zašto se podatke uopće treba organizirati u oblik matrice jer su ionako već posloženi u bazi podataka ili Excel datoteci. Odgovor na to pitanje je vrlo jasan, svaki stupac tablice smatra se novom dimenzijom te zbog toga ne možemo vizualizirati grafikon u 50 dimenzija, jer nažalost, raspoznajemo samo tri. Ovaj odjeljak će prikazati nekoliko primjera primjene linearne algebre u strojnom učenju.

Primjer 1 – Skupovi podataka

Ispod se nalazi tablica koja prikazuje zapažanje u svakom redu i značajku zapažanja u svakom stupcu. [6]

Tablica 1. Iris flowers dataset

5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa

Izvor: Machine Learning Mastery, https://machinelearningmastery.com/examples-of-linear-algebra-in-machine-learning/, Pristupljeno 16.06.2018.

Ovi podaci su, zapravo, u obliku matrice – ključnog tipa podataka u linearnoj algebri. Nadalje, kada bi podaci bili razdvojeni u ulazne i izlazne kako bi mogli ući u nadzirani model strojnog učenja, na primjer na mjere i vrste cvijeta, rezultat bi bila matrica **X** i vektor **y**. Kao što je već ranije napomenuto, vektor je također ključan tip podataka u linearnoj algebri.

[6] *Linear Algebra in Machine Learning*, Machine Learning Mastery, https://machinelearningmastery.com/examples-of-linear-algebra-in-machine-learning/, pristupano 16.06.2018.

Primjer 2 – Slike i fotografije

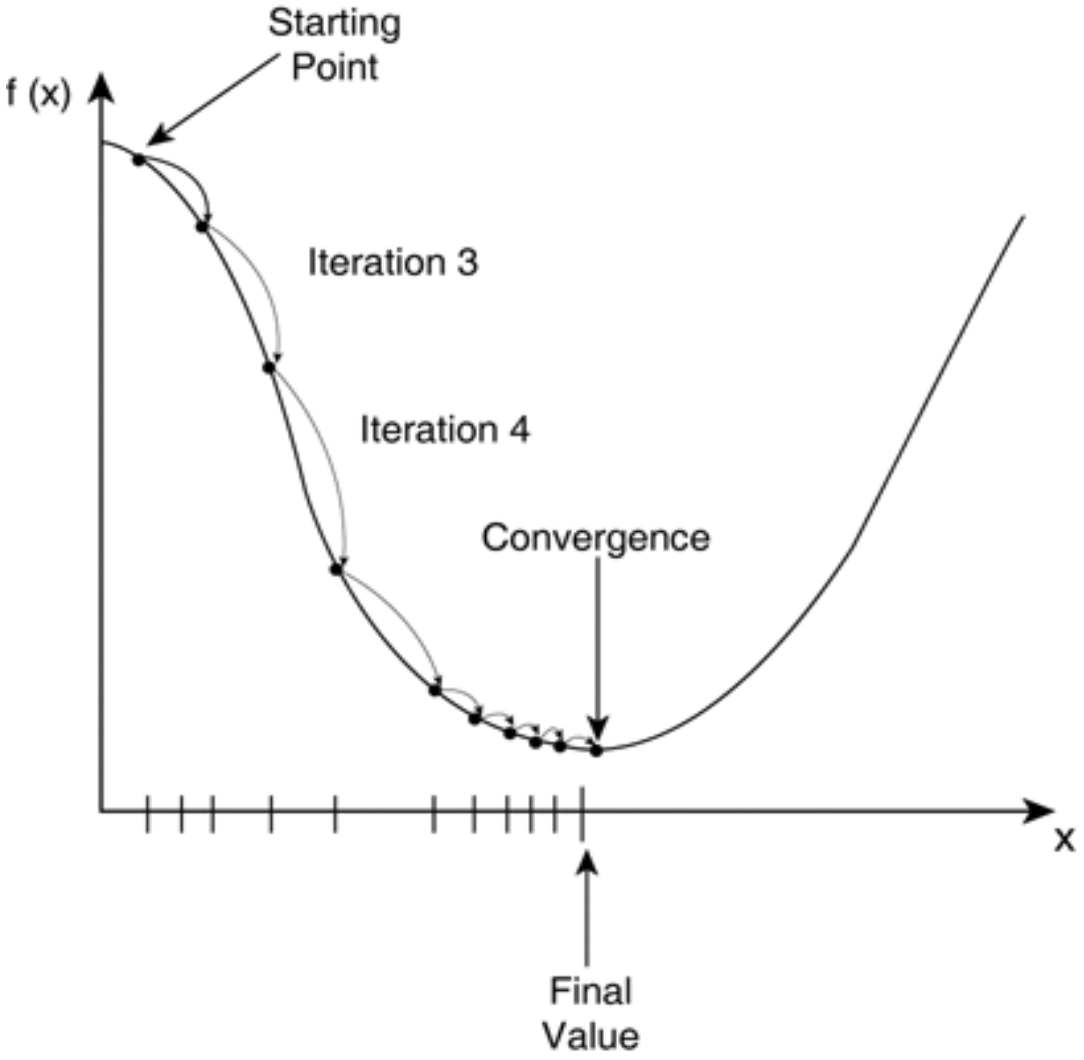
Za nekoga tko radi s fotografijama i slikama u području računalne vizije, važno je znati da je svaka slika zapravo struktura poput tablice sa širinom i visinom. U toj tablici svaki piksel u svakoj ćeliji ima jednu vrijednost za crno-bijele fotografije i tri vrijednosti za fotografije u boji (crvena, zelena i plava). To dokazuje da su fotografije čisti primjer matrice iz linearne algebre. Operacije izvršene na fotografiji, poput obrezivanja i povećanja izvršavaju se pomoću operacija koje se koriste u linearnoj algebri.

3.1.2. Diferencijalni i integralni račun više varijabli

Sam naziv „diferencijalni i integralni račun više varijabli“ (eng. multivariable calculus) poprilično dobro objašnjava njegovu definiciju. On je nastavak računa s jednom varijablom i bavi se funkcijama više varijabli. U strojnom učenju najzastupljeniji je algoritam spuštanja po gradijentu (eng. gradient descent) koji se koristi za pronalazak minimuma funkcije. Detaljnije, to je algoritam koji se koristi za pronalazak vrijednost parametara funkcije koji minimiziraju cost funkciju[7]. Cost funkcija mjeri koliko je model pogriješio u svojoj mogućnosti procjene veze između vrijednosti X i y.[8] Algoritam spuštanja po gradijentu najčešće se koristi kada se parametri ne mogu izračunati korištenjem linearne algebre i moraju se naći optimizacijskim algoritmom. On također omogućuje modelu da nauči u kojem smjeru se treba kretati kako bi se smanjile greške (razlika stvarne vrijednosti y i predviđena vrijednost y). Kako se model kreće, sve se više približava minimumu gdje daljnja optimizacija parametara rezultira minimalnim ili nikakvim promjenama – točka konvergencije. Primjer toga se može vidjeti na sljedećem prikazu:

[7] *Gradient Descent for Machine Learning*, Machine Learning Mastery, https://machinelearningmastery.com/gradient-descent-for-machine-learning/, pristupano 17.06.2017.
[8] *Machine learning fundamentals – linear regression*, Towards Data Science, https://towardsdatascience.com/machine-learning-fundamentals-via-linear-regression-41a5d11f5220, pristupano 17.06.2017

Slika 1. Primjer algoritma spuštanja po gradijentu



Izvor: Towards Data Science, https://towardsdatascience.com/machine-learning-fundamentals-via-linear-regression-41a5d11f5220, Pristupljeno 17.6.2018.

3.2. Statistika

Statistika i strojno učenje su dva srodna područja. Stoga, poznavanje statističkih metoda je prijeko potrebno za izradu modela strojnog učenja. Ovaj odjeljak će prikazati nekoliko primjera primjene statistike u strojnom učenju.[9]

[9] *Statistical Methods in an Applied Machine Learning Project*, Machine Learning Mastery, https://machinelearningmastery.com/statistical-methods-in-an-applied-machine-learning-project/, pristupano 18.06.2018.

Primjer 1 – razumijevanje podataka

Razumijevanje podataka označava poznavanje povezanosti između varijabli. Dio poznavanja te povezanosti može doći od poznavanja posla kojim se pojedinac bavi. Dvije najveće grane statističkih metoda koje se koriste u razumijevanju podataka su:

1. tzv. summary statistics – metode koje se koriste za sažimanje distribucije i veze između varijabli
2. Vizualizacija podataka – metode koje se koriste za sažimanje distribucije i veze između varijabli korištenjem dijagrama i grafikona

Primjer 2 – odabir podataka

Svi podaci i sve varijable ne moraju biti relevantne pri izradi modela. Odabir podataka je naziv za smanjenje količine podataka na one koji su najviše korisni kod predviđanja. Dva tipa statističkih metoda koje se koriste za odabir podataka su:

1. Uzorak podataka – metode koje kreiraju manje, reprezentativne uzorke iz velikih skupova podataka
2. Odabir značajki – metode za automatsko prepoznavanje relevantnih varijabli

3.3. Vjerojatnost

Vjerojatnost je matematičko istraživanje nesigurnosti te ima vrlo važnu ulogu u strojnom učenju zato što se sama izrada algoritama oslanja na pretpostavku vjerojatnosti u podacima.[10] Može se reći da je vjerojatnost da će se neki događaj dogoditi omjer broja načina na koji se može dogoditi i ukupnog broja ishoda. To se može iskazati sljedećom formulom:

$$Vjerojatnost\ događaja\ = \frac{Broj\ načina\ na\ koji\ se\ može\ dogoditi}{Ukupan\ broj\ ishoda}$$

Primjer – vjerojatnost da će se bačena kocka okrenuti na broj 4[11]

Broj načina na koji se može dogoditi: 1 – samo jedna strana kocke ima broj 4. Ukupan broj ishoda: 6 – kocka ima 6 strana. Stoga, vjerojatnost da će se kocka okrenuti na broj 4 je 1/6.

[10] *Probability Theory Review for Machine Learning*, Stanford University, https://see.stanford.edu/materials/aimlcs229/cs229-prob.pdf, pristupano 17.06.2018.
[11] *Probability*, Math is Fun, https://www.mathsisfun.com/data/probability.html, pristupano 17.06.2017.

4. STROJNO UČENJE NA PRIMJERU

Ovaj dio rada praktičnim će primjerom pokazati kako već ranije navedena izmišljena tvrtka koja se bavi recenziranjem filmova i TV serija može unaprijediti svoje poslovanja primjenom strojnog učenja. Ali najprije je potrebno reći nešto uvodno o sustavima preporuka te pojasniti kako oni rade. Dva najčešća tipa sustava preporuka su:

1. Sustav baziran na sadržaju
2. Sustav kolaborativnog filtriranja

O sustavu baziranom na sadržaju ne treba ništa posebno govoriti, njegov naziv ga prilično dobro opisuje. Valja samo reći da se on fokusira na značajke predmeta kojeg analizira, u ovom slučaju filma ili TV serije, te daje preporuke bazirane na sličnosti određenih značajki. Sustav kolaborativnog filtriranja je malo drugačiji, on izbacuje preporuke koje se baziraju na sviđanjima i ne sviđanjima drugih ljudi. U praksi se upravo on češće koristi jer ga je lakše razumjeti, i, najvažnije, daje bolje rezultate. Njegov algoritam ima sposobnost samostalnog učenja što znači da sam može naučiti koje značajke uspoređivati i na temelju njih preporučiti određeni predmet.

4.1. Istraživanje i priprema podataka

Sada će biti prikazana praktična primjena jednostavnog sustava preporuka zasnovanim na MovieLens[12] skupu podataka. Taj skup podataka sastoji se od 1682 filma te je za svaki film naveden njegov identifikacijski broj, identifikacijski broj korisnika koji je dao ocjenu te ocjena koju je korisnik dao. Sustav je izrađen u programskom jeziku *Python* uz pomoć biblioteka poput: *Numpy, Pandas, Matplotlib i Seaborn*.

Prva stvar koju je potrebno napraviti je dati ime stupcima u tablici. Imena su: *user_id, item_id, rating, timestamp*. Sve navedene nazive stupaca potrebno je povezati s podacima. To je najlakše napraviti s bibliotekom *Pandas*, pomoću nje je lako pročitati skup podataka o filmovima te mu dodijeliti imena stupaca kako su gore navedena. Prethodno objašnjeno implementirano u kodu daje slijedeći rezultat.

[12] https://grouplens.org/datasets/movielens/

Tablica 2. MovieLens skup podataka

	user_id	item_id	rating	timestamp
0	0	50	5	881250949
1	0	172	5	881250949
2	0	133	1	881250949
3	196	242	3	881250949
4	186	302	3	891717742

Izvor: Sistematizacija autora

Gore prikazana tablica prikazuje samo pet filmova zbog toga što bi bilo vrlo nepraktično prikazati cijelu listu. Može se vidjeti da je korisnik s korisničkom oznakom 0 ocijenio film s oznakom 50 s ocjenom 5 u vremenskoj oznaci 881250949. Pojednostavljeno, vremenska oznaka označava broj sekundi proteklih od 1. siječnja 1970. godine.

Nadalje, potrebno je stvoriti novu tablicu iz posebnog dokumenta koji sadrži samo identifikacijski broj filma i naziv filma te ju spojiti s prethodnom tablicom. Tablice se spajaju po stupcu s naslovom *item_id*, tako da se može vidjeti kako su različiti korisnici ocijenili isti film. Implementacija prethodne dvije rečenice u kodu rezultira slijedećim:

Tablica 3. Modificirani MovieLens skup podataka

	user_id	item_id	rating	timestamp	title
0	0	50	5	881250949	Star Wars (1977)
1	290	50	5	880473582	Star Wars (1977)
2	79	50	4	891271545	Star Wars (1977)
3	2	50	5	888552084	Star Wars (1977)
4	8	50	5	879362124	Star Wars (1977)

Izvor: Sistematizacija autora

Iz prethodne tablice vidimo kako je film *Star Wars* dobio odlične ocjene. Idući zadatak je malo bolje istražiti podatke. Za početak će biti prikazani filmovi s najboljom ocjenom. Implementacijom prethodno navedenog u kodu dobivaju se sljedeći rezultati:

Tablica 4. Filmovi s najboljom ocjenom

title	
Marlene Dietrich: Shadow and Light (1996)	5.0
Prefontaine (1997)	5.0
Santa with Muscles (1996)	5.0
Star Kid (1997)	5.0
Someone Else's America (1995)	5.0
Name: rating, dtype: float64	

Izvor: Sistematizacija autora

Pogled na prethodnu tablicu ukazuje na očit i ozbiljan problem – prikazani su podaci o navodno najboljim filmovima, ali za te filmove je rijetko koja osoba čula. To ukazuje da je film pogledala možda jedna osoba i dala joj ocjenu 5. Da je više osoba pogledalo i ocijenilo film, rezultati ne bi bili ni približno takvi.

Tablica 5. Filmovi s najviše ocjena

title	
Star Wars (1977)	584
Contact (1997)	509
Fargo (1996)	508
Return of the Jedi (1983)	507
Liar Liar (1997)	485
Name: rating, dtype: int64	

Izvor: Sistematizacija autora

Sada bi bilo zanimljivo istražiti koji filmovi imaju najviše ocjena, možda to otkrije jesu li zaključci da su filmovi iz prethodne tablice krivi. Implementacija u kodu daje sljedeći rezultat. Navedena tablica dokazuje da zaključci prethodnog odlomka nisu bili krivi jer najboljih pet filmova iz dvije prethodne tablice nisu isti. Idući zadatak kojeg je potrebno napraviti je izraditi novu tablicu u kojoj će imati sve filmove iz glavnog skupa podataka te prosječnu ocjenu za svaki film. Implementacijom u kodu dobiva se slijedeći rezultat:

Tablica 6. Filmovi i prosječna ocjena

	rating
title	
'Til There Was You (1997)	2.333333
1-900 (1994)	2.600000
101 Dalmatians (1996)	2.908257
12 Angry Men (1957)	4.344000
187 (1997)	3.024390

Izvor: Sistematizacija autora

Ranije je donesen zaključak da ocjena filma ovisi o tome koliko je ljudi dalo ocjenu. Filmovi koji su ocjenjeni s ocjenom 5 apsolutno ništa ne znače ukoliko ih je ocijenila samo jedna osoba. Sada je potrebno dodati važan stupac prethodnoj tablici, a to je broj ocjena kojeg svaki film ima. Taj podatak će biti koristan kasnije, kod izrade modela. Nakon dodavanja novog stupca u kodu dobivaju se sljedeći rezultati.

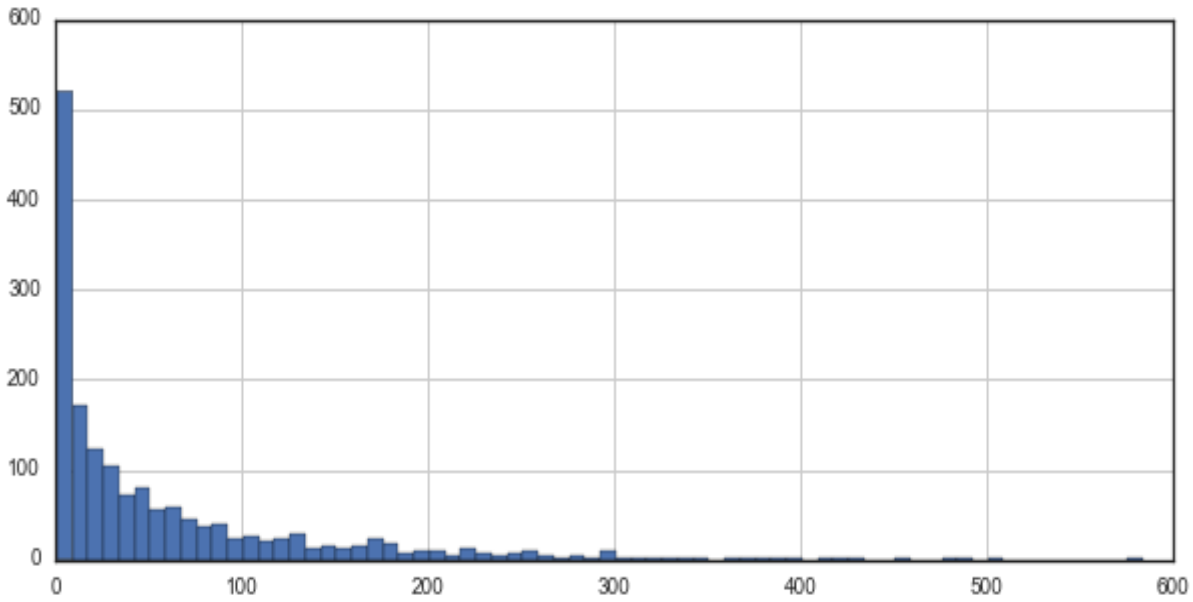
Sljedeća tablica pokazuje koliko je ljudi ocijenilo koji film te koja je prosječna ocjena. Onome tko izrađuje sustav preporuka vrlo je važna vizualna reprezentacija kako bi mogao vidjeti na koji način funkcioniraju osobe koje te filmove ocjenjuju. To je sljedeći zadatak te će rezultat biti prikazan histogramom.

Tablica 7. Filmovi, prosječna ocjena i broj ocjena

	rating	num of ratings
title		
'Til There Was You (1997)	2.333333	9
1-900 (1994)	2.600000	5
101 Dalmatians (1996)	2.908257	109
12 Angry Men (1957)	4.344000	125
187 (1997)	3.024390	41

Izvor: Sistematizacija autora

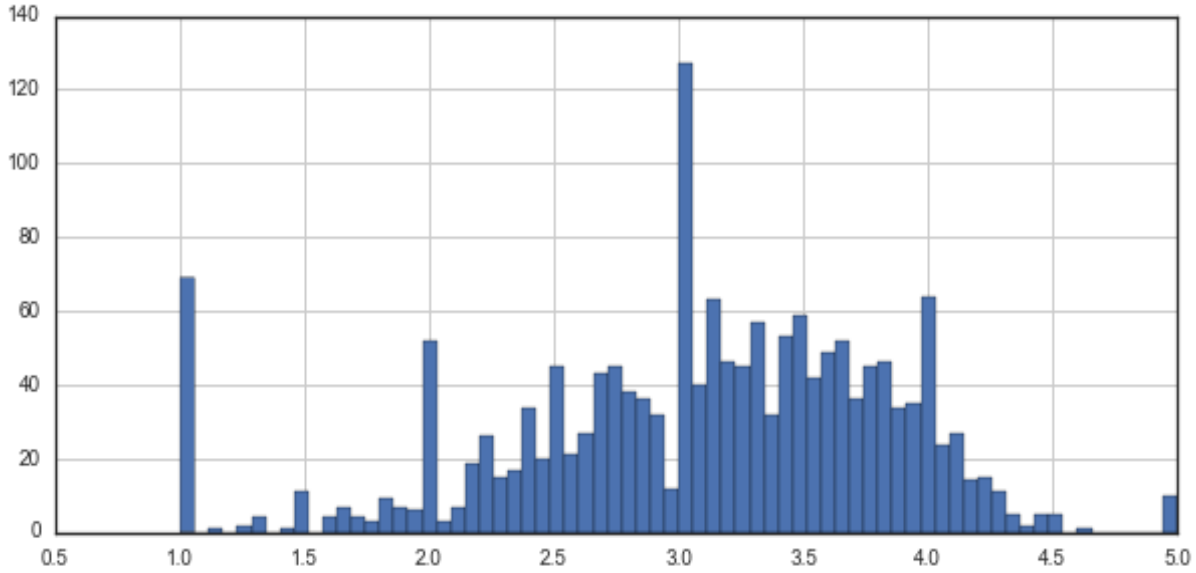
Grafikon 2. Odnos filmova i ocjena



Izvor: Sistematizacija autora

Histogram ukazuje na vrlo očitu stvar – većina filmova nije nikad ocjenjena. Razlog tomu je što većina ljudi gleda samo najpoznatije filmove tako da će ti filmovi biti upravo oni s najviše ocjena. Sljedeći zadatak će biti prikazati stvarne ocjene. One će također biti prikazane histogramom.

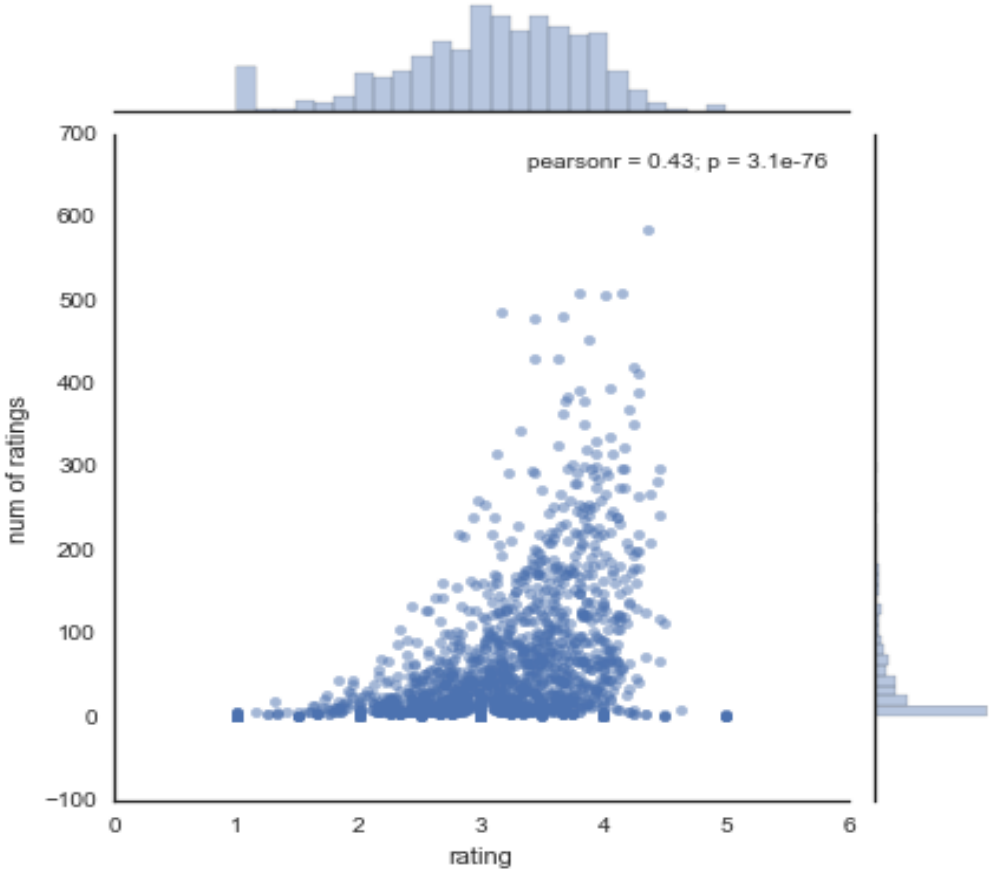
Grafikon 3. Prosječna ocjena filma



Izvor: Sistematizacija autora

Prethodni histogram ukazuje na više stvari. Najveći broj ocjena dodijeljen je cijelom ocjenom, odnosno ocjenom 1, 2, 3, 4 ili 5. Također, izgleda da je distribucija ocjena najvećeg broja filmova normalna – između ocjene 3 i 4. Još jedan važan zaključak je da ima puno filmova ocijenjenih s ocjenom 1 što također ima smisla, jer, naravno, postoji puno loših filmova. Postoji i nagli prijelaz na ocjeni 5 što može biti uzrokovano jednim od dva razloga: ili su to najpopularniji filmovi, ili su filmovi koje je pogledala samo jedna osoba i dala im ocjenu 5. Nadalje, potrebno je prikazati graf s vezom između prosječne ocjene i broja ocjena. To je prikazano sljedećim grafom.

Grafikon 4. Korelacija broja ocjena i ocjene



Izvor: Sistematizacija autora

Iako je Perasonov koeficijent samo 0.43, što ukazuje na relativno slabu korelaciju između broja ocjena i ocjene, na grafu se vidi da su u pravilu bolje ocijenjeni filmovi oni koji su ocijenjeni od strane većeg broja ljudi. To ima smisla jer što je bolji film to će ga više ljudi pogledati, a što ga više ljudi gleda to će ga više ljudi ocijeniti. Nadalje, graf dokazuje dokaz objašnjen u prethodnim odlomcima, a to je da filmovi s ocjenom 5.0 imaju mali broj ocjena. P vrijednost u grafu govori u značajnosti povezanosti između varijabli. Ukoliko je vrijednost manja od 0.05 postoji značajna povezanost, a to se na prethodnom grafu može vidjeti.

4.2. Izrada sustava preporuka

Sada kada su podaci istraženi može se pristupiti stvarnoj izradi sustava preporuka baziranog na sličnosti filmova. Prvi zadatak će biti izrada matrice koja ima identifikacijsku oznaku korisnika na jednoj strani i naziv filma na drugoj strani. Kako bi se dobio oblik matrice koristi će se pivot tablica. Svaka ćelija će se sastojati od ocjene koju je svaka osoba dala za svaki

film. Naravno, biti će puno *NaN* vrijednosti, odnosno vrijednosti neće posojati jer svaka osoba nije ocijenila svaki film.

Tablica 8. Matrica ocjena svakog filma od strane svakog korisnika

title	**'Til There Was You (1997)**	**1-900 (1994)**	**101 Dalmatians (1996)**	**12 Angry Men (1957)**	**187 (1997)**	**…**	**unknown**	**Á köldum klaka (Cold Fever) (1994)**
user_id								
0	NaN	NaN	NaN	NaN	NaN	…	NaN	NaN
1	NaN	NaN	2.0	5.0	NaN	…	4.0	NaN
2	NaN	NaN	NaN	NaN	NaN	…	NaN	NaN
3	NaN	NaN	NaN	NaN	2.0	…	NaN	NaN
4	NaN	NaN	NaN	NaN	NaN	…	NaN	NaN
5 rows x 1664 columns								

Izvor: Sistematizacija autora

Pivot tablica dokazuje prethodne tvrdnje, postoji mnogo *NaN* vrijednosti upravo iz navedenog razloga, većina ljudi nije pogledala i ocijenila većinu filmova. Sada je potrebno izraditi tablicu koja prikazuje filmove s najvećim brojem ocjena. Nakon implementacije u kodu dolazi se do sljedećeg:

Tablica 9. Filmovi s prosječnom ocjenom i najvećim brojem ocjena

	rating	**num of ratings**
title		
Star Wars (1977)	4.359589	584
Contact (1997)	3.803536	509

Fargo (1996)	4.155512	508
Return of the Jedi (1983)	4.007890	507
Liar Liar (1997)	3.156701	485
English Patient, The (1996)	3.656965	481
Scream (1996)	3.441432	478
Toy Story (1995)	3.878319	452
Air Force One (1997)	3.631090	431
Independence Day (ID4) (1996)	3.439228	429

Izvor: Sistematizacija autora

Sada je potrebno izabrati neki film i uzeti njegovu ocjenu. Za primjer će se uzeti dva filma – *Star Wars (1977)* i *Liar Liar (1996)* te će ih se povezati s prethodno izrađenom matricom i spremiti u zasebne varijable. Nakon implementacije u kodu varijabla za film *Star Wars* ima sljedeću vrijednost:

Tablica 10. Vvrijednosti za Star Wars

user_id	
0	5.0
1	5.0
2	5.0
3	NaN
4	5.0
Name: Star Wars (1977), dtyle: float64	

Izvor: Sistematizacija autora

Sada je potrebno maksimalno iskoristiti *Pandas* biblioteku, odnosno pomoću nje pogledati kakva je korelacija između ocjena za film *Star Wars* i svih ostalih filmova. Implementacija u kodu rezultira sljedećim:

Slika 2. Faktori korelacija ocjena filma Star Wars i ostalih filmova

```
title
'Til There Was You (1997)                                    0.872872
1-900 (1994)                                                -0.645497
101 Dalmatians (1996)                                        0.211132
12 Angry Men (1957)                                          0.184289
187 (1997)                                                   0.027398
2 Days in the Valley (1996)                                  0.066654
20,000 Leagues Under the Sea (1954)                          0.289768
2001: A Space Odyssey (1968)                                 0.230884
3 Ninjas: High Noon At Mega Mountain (1998)                       NaN
39 Steps, The (1935)                                         0.106453
8 1/2 (1963)                                                -0.142977
8 Heads in a Duffel Bag (1997)                              -0.577350
8 Seconds (1994)                                            -0.755929
A Chef in Love (1996)                                        0.868599
Above the Rim (1994)                                        -0.645497
Absolute Power (1997)                                        0.085440
Abyss, The (1989)                                            0.203709
Ace Ventura: Pet Detective (1994)                            0.062689
Ace Ventura: When Nature Calls (1995)                        0.094514
Across the Sea of Time (1995)                               -0.132453
Addams Family Values (1993)                                  0.132264
Addicted to Love (1997)                                      0.028173
Addiction, The (1995)                                        0.507093
Adventures of Pinocchio, The (1996)                          0.111616
Adventures of Priscilla, Queen of the Desert, The (1994)     0.054740
Adventures of Robin Hood, The (1938)                         0.144587
Affair to Remember, An (1957)                                0.225451
African Queen, The (1951)                                    0.230540
Afterglow (1997)                                             0.420084
Age of Innocence, The (1993)                                -0.037176
...                                                              ...
```

Izvor - Sistematizacija autora

Nadalje, potrebno je pretvoriti prethodni prikaz u tablicu radi lakšeg prikaza te ukloniti sve *NaN* vrijednosti. Ništa se ne mijenja, implementacijom u kodu dobiva se sljedeće:

Tablica 11. Tablični prikaz korelacije ocjena filma Star Wars i ostalih filmova

	Correlation
title	
'Til There Was You (1997)	0.872872
1-900 (1994)	-0.645497

101 Dalmatians (1996)	0.211132
12 Angry Men (1957)	0.184289
187 (1997)	0.027398

Izvor: Sistematizacija autora

Vidljivo je da su podaci iz sljedeće tablice i prikaza prije identični, samo je jednostavnije percipirati podatke prikazane u tablici. Stupac *Correlation* prikazuje korelaciju između ocjena korisnika za svaki film i ocjena korisnika za film *Star Wars*. Prema tome, ukoliko se tablica sortira po korelaciji trebali bi se prikazati najsličniji filmovi. Nakon sortiranja tablice dobiven je sljedeći rezultat:

Tablica 12. Najsličniji filmovi filmu Star Wars

	Correlation
title	
Commandments (1997)	1.0
Cosi (1996)	1.0
No Escape (1994)	1.0
Stripes (1981)	1.0
Man of the Year (1995)	1.0
Hollow Reed (1996)	1.0
Beans of Egypt, Maine, The (1994)	1.0
Good Man in Africa, A (1994)	1.0
Old Lady Who Walked in the Sea, The (Vieille qui marchait dans la mer, La) (1991)	1.0
Outlaw, The (1943)	1.0

Izvor: Sistematizacija autora

Prema prethodnoj tablici, svi filmovi navedeni u njoj imaju savršenu korelaciju s filmom *Star Wars.* Kratak pogled na tablicu ukazuje da nešto nije u redu. *Star Wars* je vrlo popularan film s mnogo korisničkih ocjena, dok su filmovi iz prethodne tablice vrlo vjerojatno pogledani od strane samo jedne osobe koja je film iz tablice i film *Star Wars* ocijenila ocjenom 5. Ono što se može napraviti kako bi se dobili točni rezultati je ponovno filtiriranje tablice, ovog puta na način da filmovi koji imaju manje od određenog broja ocjena ne mogu biti preporučeni osobi koja je pogledala *Star Wars*. Prvi korak je dodavanje novog stupca tablici s korelacijom. Stupac će prikazivati broj ocjena kojeg svaki film ima. Nakon implementacije u kodu dobiven je sljedeći rezultat:

Tablica 13. Faktori korelacija ocjene s filmom Star Wars i ukupni broj ocjena

	Correlation	**num of ratings**
title		
'Til There Was You (1997)	0.872872	9
1-900 (1994)	-0.645497	5
101 Dalmatians (1996)	0.211132	109
12 Angry Men (1957)	0.184289	125
187 (1997)	0.027398	41

Izvor: Sistematizacija autora

Sljedeći korak je filtriranje tablice tako da se u njoj ne prikazuju filmovi koji nemaju bar 100 ocjena korisnika. Nakon implementacije u kodu dobiven je sljedeći rezultat:

Tablica 14. Konačne preporuke za Star Wars

	Correlation	**num of ratings**
title		
Star Wars (1977)	1.000000	584
Empire Strikes Back, The (1980)	0.748353	368

Return of the Jedi (1983)	0.672556	507
Raiders of the Lost Ark (1981)	0.536117	420
Austin Powers: International Man of Mystery (1997)	0.377433	130

Izvor: Sistematizacija autora

Napokon, dobiveni su očekivani rezultati. Prva dva preporučena filma su također iz franšize *Star Wars* tako da njih ne treba dodatno objašnjavati. Može se vidjeti značajan pad u korelaciji na zadnjem filmu. Razlog tomu je što sustavi preporuka često preporučuju vrlo popularan film nekome tko je pogledao i ocijenio vrlo popularan film. Sada se mogu napraviti isti koraci za film *Liar Liar* kako bi se vidjelo koji će filmovi biti preporučeni osobi koja ga je pogledala. Nakon implementacije u kodu dobiven je sljedeći rezultat:

Tablica 15. Konačne preporuke za Liar Liar

	Correlation	**num of ratings**
title		
Liar Liar (1997)	1.000000	485
Batman Forever (1995)	0.516968	114
Mask, The (1994)	0.484650	129
Down Periscope (1996)	0.472681	101
Con Air (1997)	0.469828	137

Izvor: Sistematizacija autora

Iz prethodne tablice također možemo iščitati odlične rezultate. Na primjer, *Liar Liar* je komedija iz 1990-ih godina u kojoj je glavni glumac Jim Carrey, kao i film *Mask* koji je također komedija iz 1990-ih u kojoj je glavni glumac Jim Carrey.

Naravno, cijeli primjer je samo pojednostavljena verzija sustava preporuka. Napredniji sustav bi zahtijevao veći skup podataka, više vremena i, naravno, daleko kompleksniju matematičku analizu. Sljedeće pitanje koje se postavlja je gdje se i na koji način može unaprijediti poslovanje neke tvrtke sa sustavom preporuka i strojnim učenjem općenito. Na primjer, ukoliko pojedinac ima tvrtku i prodaje robu putem interneta (na vlastitoj web stranici) sustav preporuka može njegovom kupcu preporučiti proizvod kupljen od strane drugih ljudi, a da on uopće nije bio svjestan o njegovom postojanju. Na taj način dolazi se do više prodaja, a više prodaja znači veću zaradu. Sve velike kompanije koriste sustave preporuka – *Amazon* i *eBay* za prodaju proizvoda, *Netflix* za preporuku filmova, *Google* za prodaju aplikacija i igara putem servisa *Google Play*. Naravno, njihovi sustavi su puno kompleksniji od gore objašnjenog, ali može se zaključiti kako su upravo te kompanije došle do novih klijenata i/ili povećale potrošnju postojećih zahvaljujući strojnom učenju i mogućnostima koje im ono pruža.

4.3. Verifikacija istraživačkog pitanja

Istraživačko pitanje IP1: *Što su sustavi preporuka?* Sustavi preporuka su podklasa sustava filtriranja informacija i žele predvidjeti ocjenu ili sviđanje koju bi im korisnik dao. Kao što je u radu već rečeno, postoje sustavi bazirani na sadržaju i sustavi kolaborativnog filtriranja.

Istraživačko pitanje IP2: *Kako tvrtke mogu koristiti sustave preporuka kako bi unaprijedile poslovanje?* Tvrtke mogu na svoju online trgovinu uključiti sustav preporuka i tako korisniku preporučiti proizvode koji su se sviđali drugim kupcima ili preporučiti proizvode kupcu koji su kupljeni od strane drugih kupaca koji imaju slične preferencije prvotnom kupcu.

5. ZAKLJUČAK

Ovim radom istražena je mogućnost primjene strojnog učenja u poslovanju tvrtke s ciljem pružanja bolje, individualne usluge kupcima. Rad je objasnio jednostavan primjer primjene sustava preporuka, ali u praksi su mogućnosti neograničene.

Strojno učenje će definitivno dovesti do revolucije u svijetu, svake godine sve veće zbog napretka tehnologije. Svakom idućom godinom važnost strojnog učenja će sve više dobivati na važnosti, najviše u unaprjeđenju poslovanja, ali uvelike i u automatizaciji poslova. Već danas svjedočimo o automatizaciji poslova u mnogobrojnim poslovnicama, trgovinama, kafićima i restoranima, ali i u poslovima poput transporta. Valja istaknuti da mnoge vodeće automobilske kompanije razvijaju ili su razvile tehnologiju samovozećih automobila koja bi, nakon nekog vremena, mogla potpuno automatizirati zanimanja u prijevozu, bilo da se radi o prijevozu ljudi ili roba.

Naravno, sumnjam da će takve promjene biti lako prihvaćene te bi razdoblje u kojem sada živimo usporedio s industrijskom revolucijom. Naime, čak i tih davnih dana ljudi su shvatili koju "opasnost" za njih nosi automatizacija poslova, ponajviše zbog smanjenja broja radnih mjesta koje dovodi do otkaza koje na kraju rezultira nedostatnim ili nepostojećim primanjima. Također, smatram da će napredna primjena strojnog učenja rezultirati još većim kaosom nego industrijske revolucije primarno zbog toga što dobro napravljen stroj ili program može određeni posao obavljati brže i bolje od najboljeg zaposlenika, ali najvažnije, može i "razmišljati" te se konstantno poboljšavati. Za takve programe i strojeve koristi se pojačano učenje, koje uči iz svojih pogrešaka te konstantnim obavljanjem istog zadatka na drugačiji način dolazi do optimalnog načina. Nedavno je Google-ov *AlphaZero AI* pobijedio najboljeg svjetskog igrača šaha, nakon što je sam naučio igrati igru.

Zbog svega gore navedenog kroz idućih nekoliko desetljeća očekujem dramatičnu promjenu na tržištu rada. Kako će se škole i fakulteti prilagoditi brzorastućem i promjenjivom svijetu tehnologije ostaje nepoznanica.

LITERATURA

Knjige

1. Mitchell, T. : *Machine Learning*, McGaw-Hill 1997
2. O'Reilly, M. Conway, D., Myles White, J.:*Machine Learning for Hackers*, 2012

Internetske stranice

1. Coursera, *Machine Learning online course*, URL: https://www.coursera.org/learn/machine-learning (2018-06-14)
2. Machine Learning Mastery, URL: *https://machinelearningmastery.com/examples-of-linear-algebra-in-machine-learning/* (2018-06-16)
3. Machine Learning Mastery, URL: *https://machinelearningmastery.com/gradient-descent-for-machine-learning/* (2018-06-16)
4. Machine Learning Mastery, URL: *https://machinelearningmastery.com/statistical-methods-in-an-applied-machine-learning-project/* (2018-06-16)
5. Math is Fun, URL: https://www.mathsisfun.com/data/probability.html (2018-06-17)
6. MovieLens Dataset, URL: https://grouplens.org/datasets/movielens/ (2018-06-01)
7. Towards Data Science, URL: *https://towardsdatascience.com/machine-learning-fundamentals-via-linear-regression-41a5d11f5220* (2018-06-17)
8. Stanford University, URL: https://see.stanford.edu/materials/aimlcs229/cs229-prob.pdf (2018-06-17)

POPIS SLIKA

Slika 1. Primjer algoritma spuštanja po gradijentu 11
Slika 2. Faktori korelacija ocjena filma Star Wars i ostalih filmova 22

POPIS TABLICA

Tablica 1. Iris flowers dataset 9
Tablica 2. MovieLens skup podataka 14
Tablica 3. Modificirani MovieLens skup podataka 14
Tablica 4. Filmovi s najboljom ocjenom 15
Tablica 5. Filmovi s najviše ocjena 15
Tablica 6. Filmovi i prosječna ocjena 16
Tablica 7. Filmovi, prosječna ocjena i broj ocjena 17
Tablica 8. Matrica ocjena svakog filma od strane svakog korisnika 20
Tablica 9. Filmovi s prosječnom ocjenom i najvećim brojem ocjena 20
Tablica 11. Tablični prikaz korelacije ocjena filma Star Wars i ostalih filmova 22
Tablica 12. Najsličniji filmovi filmu Star Wars 23
Tablica 13. Faktori korelacija ocjene s filmom Star Wars i ukupni broj ocjena 24
Tablica 14. Konačne preporuke za Star Wars 24
Tablica 15. Konačne preporuke za Liar Liar 25

POPIS GRAFIKONA

Grafikon 1. Zahtjevi strojnog učenja 8
Grafikon 2. Odnos filmova i ocjena 17
Grafikon 3. Prosječna ocjena filma 18
Grafikon 4. Korelacija broja ocjena i ocjene 20