Predikcija cijene i trajanja taksi vožnje u Njujorku

Srđan Topić
Fakultet tahničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
topicsrdjan99@gmail.com

Natalija Krsmanović
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
krsmanovic.natalija99@gmail.com

***Sažetak* — Kako je tehnologija napredovala tako su i načini kretanja i prevoza ljudi. Najuobičajniji i pouzdaniji način prevoza je automobilom, koji pokriva širok spektar slučajeva za upotrebu. Zbog same zastupljenosti i koristi automobila u prevozu, ali i zbog činjenice da ne može ili ne želi svaka osoba da priušti automobil, postoje uslužni prevozi automobilom od kojih je najistaknutiji prevoz taksijem. Cilj ovog rada je analiziranje taksi vožnji, kao i stvaranje modela koji će predviđati cijenu i trajanje vožnje tako što bi model na osnovu unijetih parametara dao što preciznije rezultate. Podaci na kojim je bazirana analiza i obuka modela su vezani za grad Njujork, koji je idealan za ovakav tip projekta, zato što je u velikoj mjeri zastupljen prevoz taksijem, a i takođe je jedno od najvećih poslovnih okruženja, gdje je ljudima svaki dolar i minuta bitna. Modeli koji se koriste u ovom radu su Linearna regresija, XGBoost, Random Forest i Multi-Layer-Perceptron. U okviru rada će se detaljnije objasniti izabrani modeli, komentarisaće se rezultati i eventualna poboljšanja.**

Ključne riječi — taksi vožnje; geografske lokacije; vremenski uslovi; stablo odlučivanja; regresija; neuronska mreža

I. UVOD

Kroz istoriju, ljudi su uvijek imali potrebu da se kreću od jedne do druge tačke iz raznih razloga. Kako je tehnologija napredovala, tako je i način na koji se može vršiti kretanje i transport osoba postao napredniji. Od samog kretanja kao pješak do leta avionom, čovjeku je na raspolaganju širok spektar izbora kretanja. Međutim, svaki od tih načina transporta su pogodniji i/ili određeni za specifične situacije i prilike. Jedan od najpogodnijih i najsvestranijih načina prevoza jeste prevoz automobilom koji omogućuje i obavljanje najosnovnijih obaveza, kao odlazak do prodavnice, ali i takođe omogućuje prevoz i transport na veće razdaljine u razumnom vremenskom periodu. Jedno od mjesta koje najviše koristi prevoz automobilom jeste Njujork. U velikom broju slučajeva u Njujorku ljudi koriste usluge prevoza taksijem, čak može biti isplativije nego korišćenje sopstvenog automobila za prevoz. Neki od razloga bi bili što može da se desi veliki zastoj u saobraćaju, a potrebno je da što prije stignete na neku lokaciju, te ste onda u mogućnosti da taksi napustite u bilo kom momentu, a i može takođe biti teško naći pogodno parking mjesto. Taksisti koji svaki dan voze po Njujorku i upoznati su sa saobraćajem, mogu sigurno da stignu u kraćem periodu do određene destinacije. Najbitnija prednost prevoza taksijem jeste pogodnost.

Pitanje koje je najbitnije i najčešće se postavlja je vezano za trajanje i cijenu usluge prevoza. Kada bi se moglo unaprijed znati vrijeme trajanja i cijena prevoza, to bi moglo da donese velike prednosti, kako za prevoznika tako i za putnika. Prevoznik bi imao mogućnost boljeg planiranja rute kojom bi se kretao do destinacije. Na osnovu generalnih informacija bi takođe mogao prilagoditi način naplaćivanja usluga. Takođe, zbog same transparentnosti koja se dobija poznavanjem cijene i trajanja vožnje, putnici bi mogli bolje organizovati svoje vrijeme i budžet u odnosu na te usluge, što bi rezultovalo većim zadovoljstvom mušterije.

U ovom radu će biti predstavljeno jedno od mnogih rješenja za određivanje trajanja i cijene vožnje. Rješenje će biti predstavljeno kao model za predikciju obučen nad podacima vožnji koji su javno dostupni. Bitno je pronaći koja obilježja najviše utiču na trajanje i cijenu vožnje.

U nastavku rada će biti detaljnije objašnjen problem i primjenjeni načini rješavanja problema. U narednom poglavlju će se navesti kratak opis radova koji su se bavili istim/sličnim problemom i koji su služili kao inspiracija za rješenja u ovom radu. Poglavlje III predstavlja kratak opis skupa podatka koji se koristi. U poglavlju IV se nalazi analiza i priprema podataka za naredno poglavlje, tj. obučavanje modela i predikcija. Na kraju će se diskutovati rješenja i zaključak ovog rada.

II. PREGLED POSTOJEĆE LITERATURE

U ovom poglavlju predstavljeni su radovi koji se bave rješavanjem istog problema, a koji su imali najveći uticaj na odabir metoda i načina dolaženja do našeg rješenja.

Prvi rad koji se izdvaja jeste “Fare and Duration Prediction: A Study of New York City Taxi Rides” [1]. U datom radu analiza i predikcija cijene i trajanja taksi vožnji vršena je nad NYC TLC (New York City Taxi and Limousine Commission) [2] skupom podataka. Ovaj skup podataka je iskorišten i u našem radu, pa se njegov detaljan opis nalazi u poglavlju 3. Kako postoji veliki broj uzoraka taksi vožnji, navedeni rad je suzio uzorke samo na one koji se odnose na žute taksije u toku mjeseca maja 2016. godine. Takvih vožnji ima oko 12 miliona, ali zbog ograničenih računarskih resursa prilikom razvoja modela, za trening skup na slučajan način je izabrano 8000, a za validaciju 2000 uzoraka. Kako početni skup u okviru jednog obilježja sadrži podatak o tačnom datumu i vremenu početka vožnje što nije pogodan oblik za dalju analizu, to obilježje je razdvojeno na obilježja koja

predstavljaju mjesec, dan, sat, minut, da li je u pitanju vikend ili radni dan za svaki uzorak. Ista podjela se odnosi i na obilježje za kraj vožnje. Po uzoru na ovo rješenje, u našem radu je takođe izvršena data transofrmacija obilježja za početak i kraj vožnje uz dodatne transformacije na osnovu sata tokom dana i drugih. Zbog činjenice da među dostupnim obilježjima u skupu podataka postoje i ona koja ne utiču na trajanje i cijenu vožnje, za predikciju je upotrebljen njihov podskup uz dodatna obilježja dobijena transformacijama nad postojećim. Pored transformacije vremenskih odrednica, izračunat je ukupan broj vožnji, kao i prosječna brzina u okviru jednog sata.

Predloženi modeli su linearna regresija i Random Forest. Kako bi se postigli što bolji rezultati upotrebom linearne regresije isprobana je Lasso regularizacija parametara. Ipak u radu navode kako to ne doprinosi rješenju, nego je izabran model sa selekcijom obilježja unaprijed. Za Random Forest je posmatrano koliko broj stabala utiče na tačnost nad validacionim skupom i na osnovu toga je biran konačan model. Dodatni pristup je transformacija koordinata odnosno njihova rotacija kako bi se model bolje prilagodio tom obilježju.

Mjera evaluacije modela je RMSE (Root Mean Square Error) koja je posmatrana za navedene modele zajedno sa *baseline* modelom koji za cijenu i trajanje vožnje predviđa srednju vrijednost. Na osnovu rezultata oba modela daju mnogo bolje predikcije u odnosu na *baseline* model. Ipak poredeći oba međusobno Random Forest predstavlja bolju opciju sa RMSE od 2.28$ za cijenu i 5.24min.

Sledeći rad na osovu koga je birana metoda rješavanja jeste "New York City taxi trip duration prediction using MLP and XGBoost" [3]. Navedeni rad vrši predikciju samo trajanja taksi vožnje. Skup podataka je preuzet sa *Kaggle* sajta. On sadrži podatke o taksi vožnjama od januara 2017.godine do januara 2020.godine. Kao i u prethodnom radu, izdvojena su najbitnija obilježja koja obuhvataju identifikator taksija, tačne koordinate kao i datum i vrijeme za početak i kraj vožnje, broj putnika i dužina trajanja. Na osnovu početnih i krajnjih koordinata izračunata su tri rastojanja: *manhattan*, *haversine* i *bearing*. Dodatno obilježje je i srednja vrijednost za prethodne tri razdaljine. Data rastojanja su dodata kako bi se uspostavila određena veza između početnih i krajnjih koordinata. Pored skupa podataka o vožnjama za predikciju dodat je i skup podataka o vremenkim prilikama u Njujork-u za isti vremenski period. S obzirom da vremenski uslovi utiču na kretanje i brzinu vozila, ideja da se dodatni skup podataka iskoristi za predikciju je usvojena i u našem radu.

Posmatrani modeli su XGBoost i MLP (Multi-layer Perceptron). Prije same upotrebe navedenih modela iskorišten je i metod klasterizacije odnosno *K-means* kako bi se grupisale vožnje koje su geografski bliske. Grupisanje je vršeno odvojeno po počenim i krajnjim koordinatama, što je dodatno proširilo skup podataka sa još po 100 obilježja za krajnje tačke. S obzirom da naš skup podataka ne sadrži konkretne koorinate nego idektifikatore po određenim zonama, upotreba klasterizacije nije bila potrebna.

Kao i u prethodnom radu evaluacija odnosno tačnost modela je mjerena putem RMSE. Za XGBoost model RMSE nad test skupom je 0.44 min, a za MLP 0.41 min. S obzirom da su rezultati skoro identični za oba modela, pri čemu se MLP pokazao boljom opcijom, u našem radu pored XGBoost biće primijenjen i MLP.

U trećem radu "Travel TIme Prediction using Tree-Based Ensembles"[4] za modele posmatrani su različiti algoritmi bazirani na stablima odluke. Za potrebe analize i predikcije trajanja vožnje upotrebljen je isti skup podataka kao u prvom pomenutom radu, ali od svih uzoraka izabrani su oni koji pripadaju vremenkom periodu između januara i juna 2016. godine. Broj uzoraka je dodatno sužen time što su birane vožnje kojima je početna i krajnja lokacija u Manhattan-u, čime se dolazi do brojke od prosječno 9 949 000 vožnji na mjesečnom nivou.Pored osnovnog skupa podataka sa taksi vožnjama, kao u prethodnom primjeru korišten je i skup podataka o vremenskim uslovima preuzet sa sajta "National Weather Service" [5].

Nad početnim skupom podataka vršene su transofrmacije tako das u izbačeni odreženi autlajeri sa netipičnim vrijednostima. Takođe su izračunate vrijednosti za dodata obilježja koja se odnose na distance puta, kao i ukupan broj skretanja, koraka itd. Data obilježja su dobijena korištenjem OSRM Express alata [6]. Za svako obilježje posmatran je njegov uticaj na performanse modela, pri ćemu je uočeno da su novodobivena obilježja od velikog značaja.

Za modele izabrani su: Random Forest, Extra Trees, XGBoos i LightGBM. Nad datim modelima izvršen je *fine-tuning* parametara kako bi se došlo do što boljeg rezultata. Za podešavanje parametara upotrebljena je kros validacija kroz trening skup koji sadrži 6 355 770 uzoraka, dok test skup sadrži 3 105 839 uzoraka.Evaluacija je takođe mjerena putem RMSE za sve modele. Ono što ovaj rad razlikuje od prethodnih je dodatan način upoređivanja modela na osnovu vremenskog trajanja treninga. Na osnovu rezultata najmanja RMSE je za XGBoost od 4.22min. Za razliku od XGBoost modela čije je treniranje trajalo oko 120min, Random Forest i LightGBM su trenirani manje od 10min. Iako je velika razlika u vremeneskom trajanju, tačnosti ovih modela odnosno RMSE se razlikuju za oko 3s. Zbog sličnih vrijednosti RMSE za sve modele, u našem radu će od navedenih biti posmatrani samo Random Forest i XGBoost.

Predstavljeni radovi su osnov za odabir skupa podataka i parametara modela koji će biti opisani u narednim poglavljima. Pored navedenih prilikom uporedbe rezultata modela, biće predstavljeno i referencirano još radova i rješenja.

III. Opis skupa podataka

Po uzoru na relevantnu literaturu posmatran je skup podataka "NYC TLC (New York City Taxi and Limousine Commission)". Sadrži podatke o vožnjama žutih i zelenih taksi vozila u period od januara 2009.godine do septembra 2022.godine. S obzirom na veliki broj uzoraka koji pripadaju datom skupu, odlučeno je da se daljna analiza i predikcija primijenjuju na njegovom podskupu. Na osnovu predstavljene literature odlučeno je da podksup sadrži samo vožnje za žute taksije tokom maja mjeseca 2016. godine.

Broj uzoraka po tim kriterijumima je 11 832 050 sa po 19 obilježja. Od dostupnih najznačajnija obilježja su:

- vendorID: identifikaciona oznaka vozila
- tpep_pickup_datetime: datum i vrijeme početka vožnje
- tpep_dropoff_datetime: datum i vijrme završetka vožnje
- PULocationID: identifikator TLC taksi zone iz koje je započeta vožnja
- DOLocationID: identifikator TLC taksi zone u kojoj je završena vožnja
- fare_amount: cijena taksi vožnje na taksimetru
- trip_distance: dužina vožnje tj. udaljenost od početne do krajnje tačke
- passenger_count: broj putnika u vozilu

Pored prikazanih obilježja, dostupne su kolone sa vrijednostima za dodatne naplate za takse, putarine, polazak sa aerodrome i druge, kao i tip plaćanja, bakšiš itd. S obzirom da je cilj ovog rada predikcija cijene i dužine trajanja taksi vožnje, potrebno je izračunati trajanje na osnovu početnog i krajnjeg vremena. Ono je izdvojeno u novo obilježje pod nazivom *duration* i konvertovano u minute. U poglavlju 4 će biti objašnjen način na koji su ova obilježja transformisana i koje su njihove osnovne karakteristike.

Na osnovu drugog i trećeg rada iz navedene literature, odlučeno je da se dati skup podataka proširi sa dodatnim podacima o vremenskim prilikama. Dati skup podataka je preuzet sa Kaggle sajta [7] i sadrži podatke o temperaturi, pristisku, vlažnosti vazduha, brzini vjetra i generalnom stanju vremena u Njujorku u periodu od početka 2012.godine do kraja 2016. godine. Vrijednosti za navedene parametre su dostupne na svakih sat vremena u toku dana. Sva obilježja su numeričkog tipa osim *weather_description* koji je kategoričkog i podrazumijeva 15 različitih kategorija za opis vremena. U pitanju su vrijednosti koje ukazuju na to da li je sunčano, oblačno, sa više ili manje padavina i druge. Takođe u poglavlju 4 su detaljnije opisane raspodjele ovih vrijednosti i statistike.

Kako bi se izvršila predikcija odnosno izgradili modeli, bilo je potrebno spojiti prethodno pomenute skupove podataka. Spajanje je učinjeno na način da je posmatran datum i sat u kome je započeta vožnja taksijem i za sve uzorke kojima su polasci u okviru tog sata, dodijeljene su odgovarajuće vrijednosti vremenskih parametara karakteristične za taj sat tog istog dana.

IV. Analiza i priprema podataka

Ovo poglavlje opisuje analizu skupa podataka kao i postupak dobijanja ciljnog skupa podataka koji će se koristitit u obučavanju modela. U nastavku će se uporedo navoditi analiza kao i priprema podataka koja obuhvata:

- generalnu analizu i pripremu
- analizu i pripremu sa ubačenim podacima o vremenskim uslovima
- analizu i pripremu sa ubačenim podacima o geografskim lokacijama

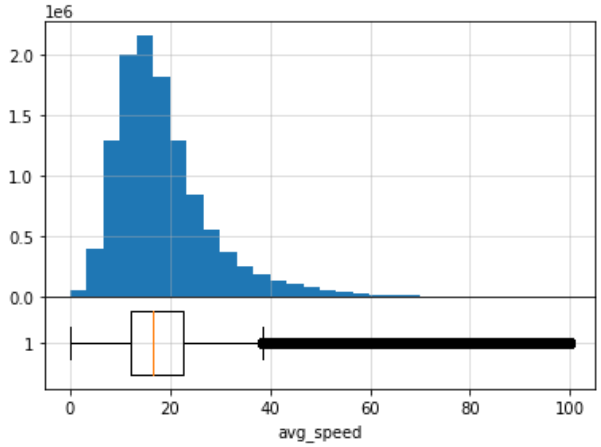
A. Generalna analiza i priprema

U ovom dijelu analiza i priprema podataka se vrši nad početnim skupom podataka koji je preuzet sa sajta. Priprema podataka obuhvatala sljedeće korake:

- izbacivanje nedostajućih podataka
- izbacivanje obilježja koja nisu relevantna, tj. nisu potrebna za analizu i ne utiču na dobijanje konačnog rezultata
- konverzija vrijednost (npr. milja u kilometre)
- transformacija postojećih obilježja
- izvođenje novih obilježja iz postojećih (prosječna brzina, period dana, trajanje vožnje itd.)

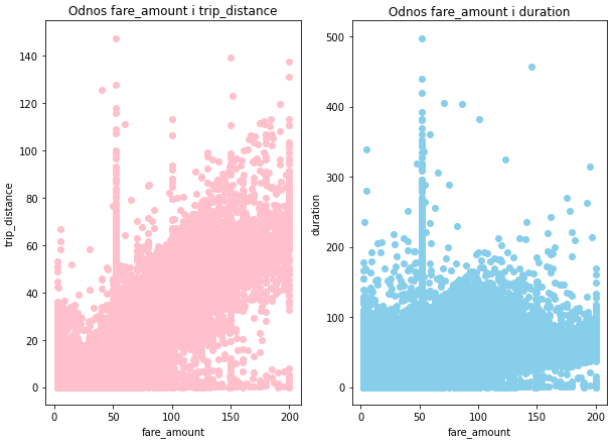
Gore navedeni koraci su obavljeni prvenstveno da bi se podaci bolje razumili, ali su i izvršena osnovna prečišćavanja podataka (kao što je npr. izbacivanje nepostojećih vrijednosti). Ovako transformisani podaci, koji se bolje razumiju, će poslužiti za dodatno filtriranje skupa podataka koji će se na kraju koristiti za obučavanje modela. U nastavku će se analizirati i filtrirati vrijednosti određenih obilježja i takođe će se prokomentarisati o nekim zanimljivim vrijednostima koje se pojavljuju.

Obilježje koje je najviše doprinijelo filtraciji skupa podataka jeste prosječna brzina vozila u datoj vožnji. Ovo obilježje je izvedeno korišćenjem trajanja i pređenog puta vožnje. Isfiltrirarn je skup podataka da sadrži vožnje čija je prosječna brzina u opsegu od 5km/h do 100km/h. Gornja granica je postavljena, jer se 90% vožnji odvija unutar grada, te bi bilo nerealno da se vozilo kreće većom brzinom, a i takođe, većina vožnji(oko 85%) ima brzinu u opsegu od 10km/h do 25k/h, kao što se može vidjeti na slici 1. Izbacivanjem vožnji van navedenog opsega se takođe izbacuju autlajeri, tj. vrijednosti koje su nemoguće da se dese ili su nerealne. Te vrijednosti obuhvataju ~19 000 000km pređenog puta u nekoj vožnji, ~10 000h trajanja jedne od vožnji ili 0km pređenog puta za 5h vožnje i mnoge druge vrijednosti koje se odnose trajanje i pređeni put vožnje.



Slika 1. Raspodjela vožnji po prosječnim brzinama

Sljedeće obilježje koje će filtrirati je cijena vožnje. Pored vožnji čija je cijena manja ili jednaka 0$ izbačene su i vožnje čija je cijena manja od 2,5$, jer je to početna cijena na taksimetrima. Postoje također i ovdje autlajeri gdje su cijene preko 600$, a čije je trajanje vožnje manje od 1h, te su i ti uzorci izbačeni. Može se primjetiti na slici 2 da za sve raspone vrijednosti trajanja i pređenog puta vožnje da je cijena 52$. Razlog za to je što sve vožnje do ili od JFK aerodroma imaju fiksnu cijenu od 52$.

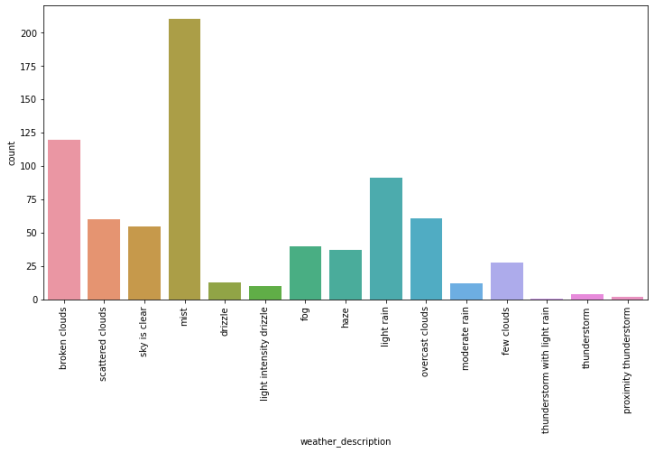


Slika 2. Cijena vožnje za pređeni put i vrijeme trajanja vožnje

Što se tiče broja putnika po vožnji, raspon je od 0 do 9 putnika, te su vožnje bez putnika (tj. 0) izbačene. Razmatralo se da se izbace vožnje sa više od 5 putnika, jer bi trebalo da su zakonski zabranjenje, međutim, postoje vozila kojim je moguće vršiti prevoz tog broja putnika.

B. *Analiza i priprema sa ubačenim podacima o vremenskim uslovima*

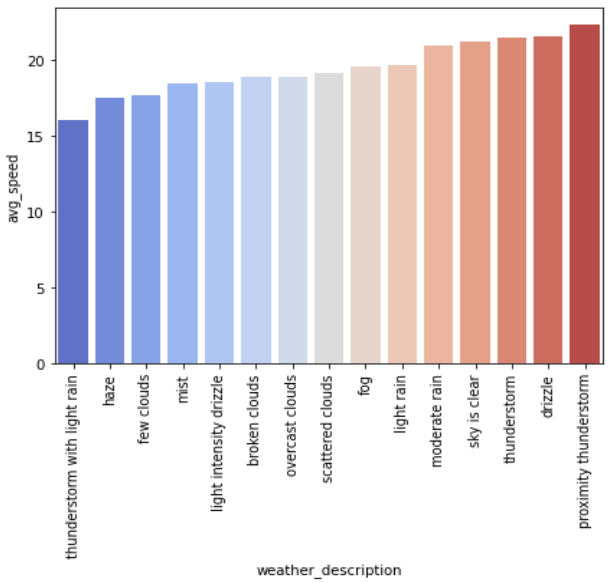
U ovom dijelu će se ukratko analizirati cijeli skup podataka sa ubačenim podacima o vremenskim uslovima. Podaci pokrivaju vremenske uslove za svaki sat određenog dana. Svi podaci u tom skupu su korektni, tj. nije bilo nikakvih autlajera niti pogrešnih podataka ili vrijednosti.



Slika 3. Vremenski uslovi tokom svakog sata u maju mjesecu

Kao što je prikazano na dijagramu može se uočiti da postoji najviše uzoraka tj. sati tokom kojih je bila magla, a poslije i nebo prekriveno oblacima koji obično ne donose padavine. Vidi se da je malo uzoraka kada je bilo nevrijeme i padala kiša što je i logično za očekivati da nema puno padavina tokom maja mjeseca.

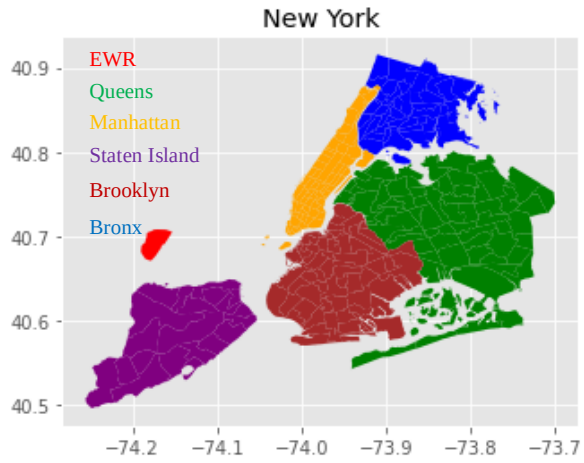
Ukoliko se posmatra prosječna brzina (slika 4) po vremenskim uslovima iako je očekivano da manje vrijednosti budu tokom kisnih sati, zapravo su veće kada su oluje i kise, dok se taksiji kreću sporije kada je zastupljena magla i sitna kisa. Što se tiče trajanja vožnje one su kraće po kišnom vremenu i olujama, dok su najduža trajanja za vremenske uslove bez kiše ali sa oblacima. To čak ima i smisla, jer je vjerovatno manja gužva u saobraćaju dok su oluje.



Slika 4. Prosječna brzina za određene vremenske uslove

C. *Analiza i priprema sa ubačenim podacima o geografskim lokacijama*

Ovaj dio će se baviti analizom cijelog skupa podataka sa ubačenim podacima o geografskim lokacijama. Njujork se sastoji od 5 opština i svaka od tih opština ima svoje taksi zone (oko 40 ili više zona po opštini) prikazanih na slici 5. Svaka vožnja ima navodi 2 identifikatora zona, prvi gdje je pokupljen putnik, a drugi gdje je ostavljen. Na osnovu tih podataka se mogu izvući zaključci gdje se vožnje obavljaju.



Slika 5. Prikaz svih opština u Njujorku

U ovom slučaju, skup podataka se odnosi na "Yellow cab" prevoznika. Ovaj prevoznik pruža usluge na nivou cijelog Njujorka, ali najzastupljenije su vožnje u Menhetnu(engl. Manhattan), tj. oko 90% svih vožnji započinju i završavaju u Menhetnu. Iz tog razloga dosta drugih taksi kompanija uopšte ne posluje u Menhetnu, zbog same zastupljenosti "Yellow cab"-a. Većina ostalih vožnji počinje ili završava na aerodromu, tj. većinom se ljudi prevoze od/do aerodroma. U nastavku će se komentarisati vožnje koje imaju najveću pređenu razdaljinu i koje su najduže trajale.

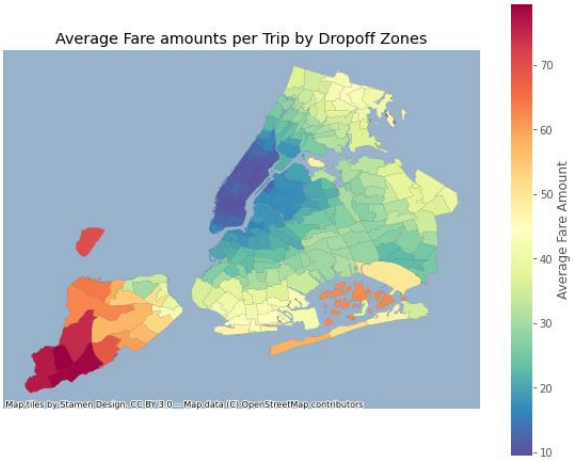
TABELA 1. POREĐENJE NAJDUŽE I NAJDALJE VOŽNJE

	Najduža vožnja	**Najdalja vožnja**
pređeni put	~ 49km	~ 147km
trajanje	~ 8h 17min	~ 3h
prosječna brzina	~ 6km/h	~ 50km/h
cijena	52$	52$
od (zona)	JFK Airport	JFK Airport
do (zona)	Battery Park City	Times Sq/Theatre District
doba dana	5:00 – 9:00	22:00 – 5:00

U tabeli 1 se porede najduža (vremenski) i najdalja (pređenim putem) vožnje. Obe vožnje započinju na aerodromu u Kvinsu(engl. Queens) a završavaju u Menhetnu i koštaju fiksno 52$, jer je to tarifa za aerodrom. Na Google mapama, najkraća ruta za obe vožnje ne bi trebalo da je više od 50km. Gledajući najdužu vožnju, pređeni put se poklapa, ali vožnja je trajala skoro 8 i po sati. Tu je moglo uticati više faktora. Neki od tih faktora je što se vožnja odvijala u jutarnjim časovima kada je veća gužva generalno, a i možda se desio neki zastoj u saobraćaju koji je uzrokovao toliko čekanje. Drugi razlog bi bio da je vožnja rezervisana u nekom trenutku, tj. da vozač može da upali taksimetar i čeka putnika. Sa druge strane, najdalja vožnja ima skoro 150km pređenog puta. Iz toga se može zaključiti da je putnik imao specifičnu rutu kroz grad, koju je vozač morao ispoštovati. Ta vožnja je trajala 3 sata, ali treba napomenuti da se odvija u kasnim časovima poslije ponoći, te je i velikom vjerovatnoćom manja gužva u saobraćaju.

Slični ako ne i isti zaključci se mogu izvući i za vožnje koje se obavljaju unutar jedne opštine ili čak unutar jedne zone u opštini. Gotovo sve te vožnje se odvijaju u Menhetnu, gdje je najdalja i najduža jedna te ista vožnja.

Pošto se većina vožnji odvija u Menhetnu ili iz drugih opština se putnici voze do/od Menhetna, logično će biti da i što su zone bliže menhetnu, tako su i cijene manje. To se takođe može uočiti na slici 6, gdje su najveće cijene u najdaljim zonama ili fizički odvojenim zonama, kao npr. Staten Island.



Slika 6. Prosječna cijena vožnje u zavisnosti od mjesta do kog se putnik vozi

V. METODOLOGIJA

U ovom poglavlju biće predstavljeni modeli koji su formirani tako da vrše predikciju cijene i dužine trajanja taksi vožnje. Za svaki od narednih modela cilj je bio podesiti parametre tako da se smanji greške predikcije ali ne dođe do overfitting-a. S obzirom na činjenicu da se parametri prilagođavaju u zavisnosti od rezultata potrebno je skup podataka podijeliti na trening, validacioni i test skup. Nad trening skupom se obučavaju modeli kako bi se uočile određene karakteristike i veze u podacima. Tako obučeni modeli se dalje pokreću nad validacionim skupom i vrše predikciju. U zavisnosti od dobijenih rezultata mijenjaju se vrijednosti parametara za sledeću iteraciju. Nakon što su konačni parametri određeni, model se upotrebljava nad test skupom i dobija se konačna mjera tačnosti modela.

Za evaluaciju modela korištena je RMSE, kako bi se greške uporedile sa greškama iz navedene literature. Pored toga računat je i R^2 score, radi jasnijeg razumijevanja performansi modela.

A. *Priprema podataka*

Za upotrebu određenih modela potrebno je da sva obilježja budu numeričkog tipa. Kako je prethodno spomenuto obilježje weather_description je kategoričko pa ga je potrebno ednkodovati na pravi način. S obzirom da kategorije nisu u određenom poretku, direktna konverzija u brojeve nije dobro rješenje. Iz tog razloga odlučeno je da se smanji broj kategorija sa 15 na 5 grupisanjem po sličnosti značenja. Na primjer kategorije scattered_clouds, broken_clouds, few_clouds su grupisane u jednu kategoriju "cloudly". Nad novim katgorijama primijenjen je One-hot encoding.

Radi poboljšanja performansi modela izvršena je standardizacija obilježja. Standardizacijom su sve vrijednosti za određeno obilježje modifikovane tako da je srednja vrijednost 0, a standardna devijacija 1. Ovaj postupak ubrzava trajanje obuke modela i smanjuje vjerovatnoću za favorizovanje određenih obilježja.

B. *Trostuka podjela podataka*

U ovom radu na osnovu istraživanja odlučeno je da 70% uzoraka iz skupa podataka pripada trening skupu, 10% validacionom,a 20% test skupu. Data podjela je izvršena primjenom funkcije train_test_split. Na ovaj način je izbjeguta pojava da se model ii obučavaju i testiraju nad istim uzorcima čime se smanjuje pristrasnost modela.

Nakon što je izvršena podjela skupa podataka, izvršena je analiza statističkih vrijednosti za izlazne varijable u okviru svakog od poskupova. Primjećuje se da su statistike odnosno raspon vrijednosti kao i srednja vrijednost približne. Iz toga slijedi pretpostavka da rezulatati u okviru validacionog i test skupa ne bi trebalo puno da se razlikuju i odstupaju.

C. *Modeli*

Na osnovu literature, izabrana su četiri modela koji su imali najmanje grešeke prilikom perdikcije. Nad njima će se vršiti eksperimenti u ovom radu, a to su:

1) *Regresija*

Po uzoru na prvi navedeni rad iz poglavlja 2, odlučeno je da se isproba linearna ali i polinomijalna regresija, zbog nelinearnih zavisnosti ciljnih atributa od atributa sa kojima su u pozitivnooj korelaciji. Dodatno je izvršen eksperiment sa selekcijom obilježja korišćenjem OLS (Ordinary Least Square) regresije i posmatranjem t-testa. Na osnovu rezultata ovog postupka zadržana su sva obilježja.

Nad validacionim skupom upoređivani su rezultati za linerarnu regresiju i polinomijalnu regresiju sa polinomima reda 2 i reda 3. Kao što je očekivano, posljednji model reda 3 ima najbolje rezultate kako za predikciju cijene tako i trajanja vozila. Zbog toga će on upotrebiti se nad test podacima kao konačan model regresije. Rezultati nad testnim skupom će biti analizirani u okviru narednog poglavlja za sve konačne modele uporedo.

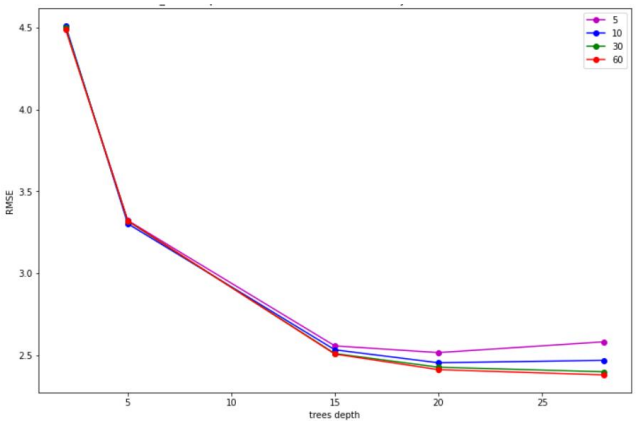
2) *Random Forest*

Za razliku od regresije, dati metod se zasniva na stablima odluke. Više je prilagođen kompleksnijim zadacima, pri čemu je model sposoban da uoči karakteristično ponašanje unutar podataka, što ga čini pogodnim rješenjem za naš problem.

S obzirom na veliki broj hiperparametara kojima se podešava dati model, bilo je potrebno izabrati one više relevantne i posmatrati njihove kombinacije. Po rezultatima iz prvog navedenog sličnog rada, jedan od najbitnijih parametara je broj stabala. Na osnovu rada "Travel TIme Prediction using Tree-Based Ensembles" dubina stable od 28 je najviše doprinjela smanjenju greške. Posljedično je izabrano da se posmatraju kombinacije sledećih vrijednosti:

- trees_num (broj stabala): 5,10,30,60
- depth (dubina stabla) : 2,5,15,20,28

Na slici 7 je prikazana RMSE za sve kombinacije prethodno navedenih parametara. Može se uočiti da broj stabala skoro ne utiče na grešku modela u našem slučaju. Zanimljivo je primijetiti da dubina stable pozitivno utiče na smanjenje greške ali do određene granice. Najveći skok u promjeni greške se dešava na dubini 15, dok je nakon toga blago poboljšanje performanski primjetno za dubinu 20, a na dubini 28 se dešava povećanje greške za određeni broj stabala. Iz navedenih razloga za konačne vrijednosti parametara izabrani su: trees_num =30 i depth=15.



Slika 7. RMSE za kombinaciju broja stabala u odnosu na dubinu

3) *XGBoost*

Kao i kod Random Forest-a, XGBoost model omogućava prilagođavanje velikog broja parametara. Vrijednosti koje će biti posmatrane su:

- max_depth (maksimalna dubina stabla): 6,7,8,9,10,11
- min_child_weights (minimalna zadovoljena težina za kreiranje novog čvora) : 3,4,5,6
- subsample (podskup posmatranih uzoraka): 0.7, 0.8, 0.9, 1
- colsample_bytree (podskup posmatranih obilježja po stablu) : 0.7, 0.8, 0.9, 1
- eta (stopa učenja): 0.01, 0.05, 0.1, 0.2

S obzirom da je u ovom slučaju izabran veći broj hiperparametara zbog računarskih resursa, nisu kombinovane sve vrijednosti u isto vrijeme.

Prvobitno je posmatrana kombinacija max_depth i min_child_weights jer oni najviše utiču na arhitekturu šume i oblik stabala. Potrebno ih je zajedno kombinovati kako bi se uspostavila balansiranost između varijanse i pristrasnosti modela. Prilikom eksperimenata sa vrijednostima za ove parametre ostali parametri su postavljeni na njihove podrazumijevane vrijednosti. Na osnovu rezultata uočeno je da je RMSE najmanja za dubinu od 9, dok se za veće vrijednosti drastično povećava.

Sledeći posmatrani parametri su subsample i colsample_bytree. Korištenjem ovih parametara smanjujemo vjerovatnoću preobučavanja modela. RMSE za različite vrijednosti subsample se linearno smanjuje sa povećanjem broja posmatranih uzoraka za oba ciljna obilježja. U slučaju colsample_bytree primjetna je razlika u performansama za cijenu i trajanje, pa je izabrana vrijednost od 0.8.

Posljednji podešavani parameter je eta tj. learning_rate koji za date vrijednosti skoro da i ne utiče na promjenu RMSE, ali je ipak prisutan blagi pad greške sa njegovim povećanjem .

Za konačan XGBoost model vrijednosti hiperparametara su sledeće: max_depth=9, min_child_weights=6, subsample=1, colsample_bytree=0.8 i eta=0.2.

4) MLP

MLP predstavlja drugačiji pristup rješavanje problema koristeći neuronsku mrežu tj. konkretno potpuno povezane neurone. Na osnovu rezultata prethodnih modela nad validacionim skupom uočeno da RMSE nije preko 4$ za cijenu i 8min za trajanje, pa je stoga isproban MLP sa manjim brojem slojeva i neurona.

Podešavani parametri su:

- hidden_layer_sizes (broj neurona po skrivenom sloju): (5,5), (10,10), (10,10,10), (15,15,15), (20,20,20)
- activation (aktivaciona funkcija): relu, tanh
- solver (metod optimizacije parametara): Adam

Prvobitno je za broj neurona po sloju maksimalno upotrebljeno 10, ali pošto su rezultati u tom slučaju bili lošiji u odnosu na Random Forest i XGBoost, MLP je naknadno proširen sa većim brojem neurona po sloju. To je doprinjelo smanjenju greške. Što se tiče aktivacione funkcije za svaku arhitekturu slojeva tanh funkcija je imala manju grešku, tako da su konačne vrijednosti za hidden_layer_sizes (20,20,20) i tanh za activation.

VI. REZULTATI I DISKUSIJA

U okviru ovog poglavlja biće analizirane performanse modela sa izabranim parametrima nad test skupom. Bitno je napomenuti da su modeli primijenjeni za predikciju cijene i vožnje nad skupom podataka koji sadži samo taksi vožnje i nad skupom podataka koji je proširen sa podacima o vremenskim uslovima.

Rezultati modela su upoređivani međusobno,u odnosu na radove predstavljene u poglavlju 2, kao i dodatne radove koji će biti spomenuti u nastavku.

Tabela 2 predstavlja rezultate našeg modela i prethodno pomenutog rada. Uočava se da je RMSE manja za oba ciljna obilježja, iako razlika nije velika. To se može objasniti time što se koriste isti podaci iz maja 2016.godine.

TABELA 2. REZULTATI PRIMJENOM REGRESIJE

RMSE	**Regresija**	
	Fare_amount ($)	***Duration (min)***
Naš model	3.198	6.161
Fare and Duration Prediction:AStudy of New York City Taxi Rides	3.522	6.513

U tabeli 3 se porede rezultati primjenom Random Forest-a. Posljednja kolona predstavlja rezultate iz rada "New_York_Taxi_Fare_Prediction"[8] u kojem su korišteni podrazumijevani parametri Naš model je tačnije izvršio predikciju za cijenu vožnje u odnosu na ostale radove. Primjetna je razlika od oko 1.5$ poredeći RMSE sa posljednjim radom, iz čega se dolazi do zaključka da je podešavanje vrijednosti određenih hiperparametara uticalo na poboljšanje u odnosu na podrazumijevane vrijednosti.

TABELA 3. REZULTATI PRIMJENOM RANDOM FOREST

RMSE	**Random Forest**	
	Fare_amount ($)	***Duration (min)***
Naš model	2.519	4.851
Fare and Duration Prediction:AStudy of New York City Taxi Rides	2.287	5.240
Travel Time Prediction using Tree-Based Ensembles		4.232
New_York_Taxi_Fare_Prediction	3.953	

Kao i u prethodnim slučajevima, tako su za XGBoost male razlike u RMSE. Jedina primjetna razlika jeste RMSE za trajanje vožnje u "New York City taxi trip duration prediction using MLP and XGBoost" radu od 0.44min primjenom XGBoost, a u našem 3.980min. Ovo se može objansiti time da se u tom radu koriste podaci u period od 3 godine, pri čemu je izvršena i klasterizacija podataka, pa je samim tip model bolje uspio da se prilagodi podacima i izvrši bolju predikciju.

U radu "New_York_Taxi_Fare_Prediction" [9] predložena je ista arhitektura slojeva MLP, ali je iskorišten manji broj

uzoraka za trening pa je R2 0.85 za cijenu vožnje, dok je primjenom našeg MLP 0.97.

Tabela 4 prikazuje performanse konačnih modela predloženih u ovom radu za osnovni skup podataka i skup proširen vremenskim uslovima. Može se uočiti da svi modeli imaju bolje rezultate za skup podataka sa vremenskim uslovima. Poredeći ih međusobno najveću grešku u predikciji cijene i trajanja vožnje daje model sa regresijom, dok se najbolje pokazao MLP, sa čak drastičnim skokom u poboljšanju RMSE nakon dodavanja obilježja za vrijeme.

Generalno modeli imaju tačniju predikciju za cijenu u odnosu na trajanje. To se objašnjava činjenicom da obilježje trip_distance ima najviše uticaja prilikom predikcije, a ono u skoro linearnom odnosu sa fare_amount, dok sa duration ima manju korelisanost.

TABELA 4. MEĐUSOBNO UPOREĐIVANJE MODELA

RMSE	**Table Column Head**			
	Taxi Data		***Taxi And Weather Data***	
	Fare_amount	Duration	Fare_amouont	Duration
Polinomijalna regresija	3.204	6.163	3.198	6.157
Random Forest	2.522	4.907	2.519	4.851
XGBoost	2.478	4.166	2.458	3.980
MLP	2.665	4.894	1.874	0.379

VII. ZAKLJUČAK

Rezultati svih modela nad početnim skupom podataka su neznatno lošiji u odnosu na rezultate sa uključenim podacima o vremenskim uslovima, izuzev MLP-a koji je postigao odlične rezultate ubacivanjem podataka o vremenskim uslovima. Takođe, rezultati ovog rada su približni rezultatima drugih radova, iako nisu u svim radovima korišteni isti podaci.

Same vrijednosti rezultata se svode na RMSE od 2 do 3$ za cijenu i 4 do 6 minuta za trajanje vožnje. Gledajući cijenu, to zapravo nisu tolike promjene, jer vožnje na kraće distance bi vjerovatno imale još manju grešku u procjeni cijene u odnosu na prosjek, ali bi suprotno važilo za vožnje na dalje distance. Bitnije je napomenuti trajanje vožnje, koje bi čak i značajnije bilo pogotovo u poslovnom okruženju u Njujorku, gdje je ljudima bitnije da znaju koliko je vremena potrebno da stignu do neke lokacije.

Kako bi se dobili bolji rezultati i proširio ovaj rad, predlaže se pokretanje najbolje pokazanih modela ali nad velikim skupom podataka, uz dodatno prilagođavanje parametara. Proširenje skupa podataka da obuhvata podatke iz jedne cijele godine bi znatno doprinijeli rezultatima, a i do većeg izražaja bi došla neka obilježja koja su u ovim modelima bila skoro zanemarena, kao npr. dan u sedmici ili vremenski uslovi. Znajući da se MLP najbolje pokazao među navedenim modelima, trebalo bi pogledati i isprobati različite arhitekture i kombinacije neuronskih mreža u svrhu dobijanja boljih rezultata.

Dodatno proširenje i poboljšanje bi moglo obuhvatati bolji pregled geografskih podataka. Trebalo bi detaljnije gledati zone od i do koje se vožnja obavlja. Na osnovu toga, znajući da li je u nekim zonama frekventniji saobraćaj, moglo bi se preciznije odrediti cijena i vrijeme trajanja vožnje.

Reference

[1] Christophoros Antoniades, Delara Fadavi, Antoine Foba Amon Jr. (2016). Fare and Duration Prediction: A Study of New York City Taxi

[2] [Online] Avaiable: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[3] M Poongodi , Mohit Malviya , Chahat Kumar , Mounir Hamdi , V Vijayakumar , Jamel Nebhen, Hasan Alyamani. (2021) New York City taxi trip duration prediction using MLP and XGBoost K. Elissa, “Title of paper if known,” unpublished.

[4] He Huang , Martin Pouls , Anne Meyer , Markus Pauly (2020). Travel TIme Prediction using Tree-Based Ensembles

[5] [Online] Avaiable: https://www.weather.gov/

[6] [Online] Avaiable: http://project-osrm.org/

[7] [Online] Avaiable: https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data?select=temperature.csv

[8] [Online] Avaiable: https://github.com/jahnavi-chowdary/New-York-Taxi-Fare-Prediction/blob/master/New_York_Taxi_Fare_Prediction.pdf

[9] [Online] Avaiable: https://github.com/raymonduchen/MLND-P6-New-York-City-Taxi-Fare-Prection/blob/master/New%20York%20City%20Taxi%20Fare%20Prediction.pdf