Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 4.9
DOI: https://doi.org/10.24867/31BE23Krsmanovic

PREDIKCIJA TRAJANJA I CENE TAKSI VOŽNJI

TAXI TRIP TRAVEL TIME AND FARE PREDICTION

Natalija Krsmanović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – RAČUNARSTVO I AUTOMATIKA

Kratak sadržaj – *U radu je predstavljen postupak izrade sistema za analizu i obradu podataka o taksi vožnjama u Njujorku. Korišćena su dva skupa podataka - jedan koji obuhvata podatke o taksi vožnjama i drugi koji sadrži informacije o vremenskim uslovima. Nad ovim skupovima je sprovedeno pretprocesiranje, kako bi se formirao konačan skup podataka za obuku modela. Vršena je predikcija trajanja i cene upotrebom različitih algoritama mašinskog učenja. Sprovedeno je više eksperimenata, a dobijeni rezultati su upoređeni međusobno i sa rezultatima sličnih radova.*

Ključne reči: *analiza i obrada podataka, algoritmi mašinskog učenja, predikcija, taksi vožnje*

Abstract – *The paper presents the process of creating a system for analysing and processing data on taxi rides in New York. Two datasets were utilized – one containing data on taxi rides and the other containing weather data. Preprocessing was performed on these data sets to create the final dataset for model training. Different machine learning algorithms were employed to predict duration and price. Several experiments were conducted, and the results are compared to those from the literature.*

Keywords: *data analysis and processing, machine learning algorithms, prediction, taxi rides*

1. UVOD

Taksi prevoz predstavlja ključni segment urbanog saobraćaja, omogućavajući praktičan i pouzdan način transporta. Cilj ovog rada jeste rešiti problem predikcije trajanja i cene taksi vožnji u Njujorku, sa fokusom na žute taksije u oblasti Menhetn. Precizna predikcija ovih faktora bi omogućila mnogobrojne pogodnosti svim učesnicima u saobraćaju. Vozači bi imali mogućnost da izaberu najefikasnije rute i smanje vreme čekanja, putnici bi imali realističnija očekivanja na osnovu transparentnosti u cenama i vremenu, a taksi kompanijama bi bila zagarantovana bolja iskorišćenost resursa i zarada. U radu su izabrani sledeći modeli: linearna regresija, model nasumične šume (engl. *Random Forest*), *XGBoost* i višeslojni perceptron (engl. *Multi-Layer Perceptron - MLP*). Po uzoru na prethodna rešenja i slične radove, za mere evaluacije posmatrane su koren srednje kvadratne greške (engl. *Root Mean Square Error - RMSE*) i koeficijent determinacije (engl. *R-squared* - R^2). Najbolje rezultate sa *RMSE* od 2.45 za cenu i 3.98 za trajanje vožnje je imao *XGBoost* algoritam.

U narednom poglavlju su predstavljeni radovi koji se bave rešavanjem sličnog problema, a koji su u najvećoj meri poslužili kao inspiracija za ovaj rad. Treće poglavlje sadrži opis metodologije i upotrebljenih modela. Naredno poglavlje broj 4 predstavlja analizu dobijenih rezultata, dok poslednje, 5. poglavlje iznosi zaključak rada.

NAPOMENA:

Ovaj rad proistekao je iz master rada čiji mentor je bio dr Jelena Slivka, vanr.prof.

2. PREGLED STANJA U OBLASTI

Predikcija trajanja vožnje i cene taksi usluga postaje sve relevantnija tema istraživanja usled urbanizacije gradova i dinamičkih promena u saobraćaju. Shodno tome, poslednjih godina veliki broj radova se fokusira na rešavanje ovog problema predlažući različite metode i tehnike. U ranim fazama istraživanja, predikcija je bila zasnovana na klasičnim statističkim modelima, kao što su linearni regresioni modeli i modeli vremenskih serija. Vremenom postaju popularni modeli bazirani na stablima, potom neuronske mreže i u prethodnim godinama modeli dubokog učenja i hibridni modeli.

Zadatak autora u radu [1] jeste predikcija cene i trajanja taksi vožnje u Njujorku korišćenjem podataka iz skupa podataka *„New York City Taxi and Limousine Commission's trip data"*. Predloženi modeli su linearna regresija i model nasumične šume. Nakon izvršene analize uticaja obeležja, došli su do zaključka da su prosečna brzina, broj vožnji u satu, lokacija odredišta i dužina pređenog puta, obeležja od najvećeg značaja za predikcije modela. Ovi rezultati usmerili su fokus eksplorativne analize u ovom radu na pomenuta obeležja. Posmatrajući koren srednje kvadratne greške kao meru evaluacije, model nasumične šume je pokazao bolje rezultate u odnosu na linearnu regresiju i oni iznose 2.28 za cenu i 5.24 za trajanje.

Par godina kasnije, nad istim skupom podataka, vršena je takođe predikcija trajanja vožnje u radu [2]. Pored osnovnog skupa podataka sa taksi vožnjama, korišćen je skup podataka o vremenskim uslovima preuzet sa veb sajta *„National Wheater Service*". Izabrani su modeli bazirani na stablima, i to: *CART* (*engl. Classification And Regression Tree*), *Random Forest*, *Extra Trees*, *XGBoost* i *LightGBM*. Ono što ovaj rad dodatno uvodi jeste kratkoročna (*short-term)* predikcija, koja u obzir uzima podatke od poslednjih *i* vožnji, gde je uočeno da na trajanje vožnje najviše utiču informacije iz prethodnog sata. Rezultati za dugoročnu (*long-term)* predikciju pokazuju da najmanja greška iznosi 4.22 za *XGBoost*

model, dok se najlošijim pokazao *CART*. Iz tog razloga, jedan od korišćenih modela u ovom radu je *XGBoost*.

S obzirom na razvoj metoda mašinskog učenja, autori rada [3] pored stabala uvode i neuronske mreže za predikciju trajanja taksi vožnje. Konkretno koriste *XGBoost* i *MLP*. Ovaj rad se izdvaja po uvođenju novih obeležja, a to su *Manhattan*, *Haversin* i *Bearing* rastojanja, izračunata uz pomoć početnih i krajnjih koordinata. Dodatno je primenjen *K-means* algoritam radi podele vožnji u klastere u kojima je transport izvršen, što je rezultovalo sa dodatnih 200 obeležja i dodatnim povećanjem preciznosti modela. Tačnost modela merena je sa *RMSE*, koja iznosi 0.41 za *XGBoost*, a 0.44 za *MLP*.

3. METODOLOGIJA

U ovom poglavlju biće predstavljen proces implementacije sistema za predviđanje trajanja i cene taksi vožnji. S obzirom da su trajanje i cena numerička obeležja sa kontinualnim vrednostima, ovaj problem se definiše kao regresioni problem. U nastavku će biti opisani upotrebljeni skupovi podataka, način na koji je vršena analiza i potom način kreiranja i obučavanja modela.

3.1. Skupovi podataka

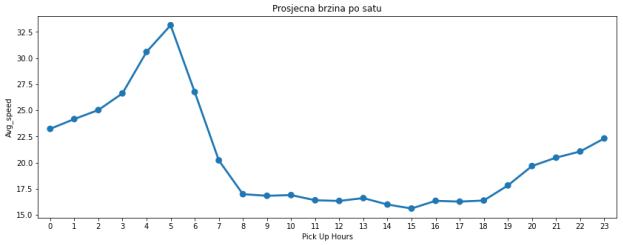
Po uzoru na pomenute radove iz literature, za glavni skup podataka izabran je „*NYC TLC (New York City Taxi and Limousine Commission) Trip Record Data*" [4]. Ovaj skup sadrži podatke o vožnjama žutih i zelenih taksija u Njujorku od 2009. godine do danas. Uključuje informacije o datumu i vremenu početka i kraja vožnje, broju putnika, ukupnom pređenom putu, početnoj i krajnjoj lokaciji, ceni, taksama i drugim obeležjima vezanih za dodatne naplate. Za potrebe ovog rada izabrani su podaci za žute taksije iz maja 2016. godine, koji sadrže 11 832 050 uzoraka i 19 obeležja. Na osnovu rezultata iz rada [2] primećeno je da se predikcija poboljšava uvođenjem dodatnih informacija o vremenskim uslovima. Kao rezultat toga, u analizu je uključen i skup podataka „*Historical Hourly Weather Data*" [5], koji sadrži podatke o temperaturi, vazdušnom pritisku, vlažnosti vazduha, brzini vetra i kratak opis vremena, za 30 američkih gradova, u periodu od 2012. do 2017. godine.

3.2. Pretprocesiranje i analiza podataka

Nad prethodno pomenutim skupovima podataka je vršen proces obrade i analize podataka, sa ciljem poboljšanja kvaliteta podataka i rezultata predikcije.

U prvom skupu podataka izbačene su nedostajuće vrednosti i nerelevantni podaci kao što su takse, putarine i drugi fiksni troškovi. Obeležja koja su sadržala datum i vreme su razdvojena u tri nova: datum, vreme i sat. Pored toga, dodato je novo obeležje koje ukazuje na to da li je u pitanju radni dan, vikend ili praznik. Takođe je dodata informacija o vremenskom periodu putovanja. Ciljno obeležje, trajanje vožnje, nije bilo direktno dostupno ali je izračunato uz pomoć početnog i krajnjeg vremenskog trenutka.

Nakon analize uočeno je da postoje vožnje kod kojih je dužina pređenog puta nula, a trajanje i cena negativne vrednosti. Ti uzorci su izbačeni, kako ne bi remetili obučavanje modela i predikciju. Prilikom posmatranja odnosa trajanja vožnje i pređenog puta, uočeno je da postoje uzorci kod kojih je pređeni put blizu nula kilometara, a vremena trajanja se kreću i preko 10 sati. Ti uzorci imaju nerealno male vrednosti srednje brzine, pa su zbog toga eliminisani oni sa prosečnom brzinom manjom od 5 km/h. Nakon ove transformacije odnos trajanja i distance postaje jasno definisan, pri čemu se uočava očekivana pozitivna korelacija između njih. Na slici 1 prikazana je raspodela prosečne brzine po satima tokom dana. Može se primetiti da su najveće vrednosti u noćnim časovima, što je i očekivano zbog manjeg intenziteta saobraćaja. Nasuprot tome, skoro dvostruko niže vrednosti za brzinu su između 08:00 h i 19:00 h. Iz ovoga se može zaključiti da gužve u saobraćaju imaju značajniji uticaj na tok vožnji i brzinu kretanja, u odnosu na ograničenja brzine. Pored toga, potvrđena je pretpostavka da su prosečne brzine kretanja znatno veće tokom vikenda naspram onih tokom radnih dana.



Slika 1. *Raspodela prosečne brzine po satima*

S obzirom na to da su podaci o vremenskim uslovima iz drugog skupa raspoređeni po odvojenim fajlovima, prvobitno je izvršeno spajanje podataka za maj 2016. godine. Nakon toga su izdvojena obeležja za datum i vreme, na osnovu kojih je proširen prvi skup podataka. U najvećem broju zabeleženih vožnji, vremenski uslovi su bili magloviti, oblačni ili sa blagim padavinama. Iako se po rezultatima iz sličnih radova očekuje da vremenske karakteristike utiču na ciljna obeležja, na osnovu matrice korelacije za sva obeležja proširenog skupa podataka, primećuje se veoma slaba korelacija pomenutih.

3.3. Priprema podataka za modele

Za upotrebu određenih modela za vršenje predikcije potrebno je obezbediti da sva obeležja budu numeričkog tipa. Međutim, u drugom skupu podataka postoji obeležje koje ukratko opisuje vremenske uslove za dati trenutak. Za njega je dostupno 15 različitih opisa odnosno kategorija. Slične kategorije su grupisane zajedno, što je dovelo do konačnih pet kategorija. Za njihovo konvertovanje u numeričke vrednosti iskorišćen je *One Hot Encoding* [6]. Rezultat ovog koraka jeste novih pet obeležja, odnosno skup podataka sa 21 obeležjem. Dati skup se za potrebe obučavanja i testiranja modela, prvobitno deli na trening i test skup u razmeri 80:20. Potom se isti metod primenjuje na dobijeni trening skup, gde se 10% podataka izdvaja u validacioni skup, a preostali deo označava konačan trening skup.

Poslednji korak pre obučavanja modela jeste standardizacija obeležja. Ovim postupkom se vrši skaliranje na takav način da svako obeležje ima srednju vrednost 0, a standardnu devijaciju 1. Na ovaj način je završen proces pripreme podataka.

3.4. Prediktivni modeli

U uvodnom delu ovog rada nabrojani su algoritmi mašinskog učenja koji će biti korišćeni za predviđanje cene i trajanja taksi vožnji. Pored izbora adekvatnih algoritama za konkretan problem, potrebno je konstruisati modele na način da vrše što preciznije predikcije. Ovaj proces uključuje treniranje modela, odnosno optimizaciju hiperparametara koji utiču na njegovo ponašanje i odluke.

Hiperparametri su optimizovani na validacionom skupu za svaki model posebno. Prva mera evaluacije jeste koren srednje kvadratne greške (engl. *Root Mean Squared Error- RMSE),* koja vrednosti prikazuje u jedinicama obeležja za koji se posmatra. Iako *RMSE* naglašava velike greške i otkriva autlajere, ne daje informaciju o tome koliko dobro model objašnjava ukupnu varijabilnost u podacima. Zbog toga se kao druga mera evaluacije uvodi koeficijent determinacije.

Za lineranu regresiju posmatrane su tri hipoteze. Prva predstavlja osnovni oblik linerane regresije, druga sadrži interakcije između nezavisnih obeležja, a treća pored interakcija sadrži i polinomijalne članove drugog reda. Kao što i očekivano od navedenih najbolje rezultate ima polinomijalna regresija reda 2.

Sledeći model koji je korišćen jeste model nasumične šume, za koga je vršena optimizacija broja stabala i dubine stabla. Prateći vrednosti *RMSE*, najveći pad se dešava za dubinu 15. Zbog toga su konačne vrednosti hiperparametara postavljene na 15 za dubinu stabla i 30 za broj stabala.

S obzirom na ograničene računarske resurse i veliki broj hiperparametara koji se podešavaju za *XGBoost*, vrednosti za njih nisu kombinovane i posmatrane zajedno. Prvobitno je posmatran odnos dubine stable (*max_depth*) i broja uzoraka po listu (*min_child_weights)*, jer oni najviše utiču na arhitekturu modela i oblik stabala. Njih je potrebno zajedno određivati kako bi se uspostavila balansiranost između varijanse i pristrasnosti modela. Nakon njih su podešavani procenat uzoraka (*subsample)* i procenat obeležja (*colsample_bytree)* koji se koriste pri izgradnji svakog stabla i na samom kraju brzina učenja (*eta)*. Konačne vrednosti za njih iznose: *max_depth* = 9, *min_child_weights* = 6, *subsample* = 1, *colsample_bytree* = 0.8 i *eta* = 0.2.

Na osnovu rezultata prethodnih modela, za *MLP* isprobane su različite kombinacije broja skrivenih slojeva i aktivacione funkcije. Zanimljivo je da je za svaku strukturu skrivenih slojeva aktivaciona funkcija *tanh* dala bolje rezultate. Konačna arhitektura jeste (20, 20, 20) – tri skrivena sloja od po 20 neurona.

4. REZULTATI I DISKUSIJA

U ovom poglavlju biće predstavljeni rezultati eksperimenata sprovedenih nad različitim skupovima podataka, primenjujući izabrane modele.

Tabela 1 prikazuje vrednosti *RMSE* svih modela za cenu taksi usluge, nad tri različita skupa podataka. Prvi predstavlja inicijalni skup podataka o taksi vožnjama koji je pročišćen od nedostajućih vrednosti i standardizovan. Naredni skup je isti kao prethodni uz dodatno transformisanje i obrađivanje uz pomoć određenih tehnika. Treći skup je nastao spajanjem prethodnog skupa sa vremenskim uslovima, kako bi se istražilo koliko dodatne informacije o vremenu utiču na tačnost i unapređenje modela. Kada se porede mere evaluacije među skupovima, primetno je da je *RMSE* za sve modele najveća i mnogo odskače za inicijalni nepročišćeni skup. Ovi rezultati ukazuju na to da modeli nisu uspeli adekvatno da uče iz podataka zbog prisustva šuma i autlajera koji su narušavali strukturu podataka u prvom slučaju. Najbolji rezultati za sve modele se vide u poslednjem redu, koji se odnosi na prošireni skup podataka.

Tabela 1. Prikaz *RMSE* za sve modele obučene i testirane nad tri različita skupa podataka za cenu taksi vožnje

Skupovi podataka	**Model**			
	Linearna regresija	**Model nasumične šume**	***XGBoost***	***MLP***
Skup 1	12.198	6.51	6.548	7.039
Skup 2	3.204	2.522	2.478	2.704
Skup 3	3.198	2.519	**2.458**	2.665

Rezultati pokazuju da iako *MLP* model predstavlja savremeniji pristup i kasnije je počeo da se koristi za rešavanje ovakvih problema, njegove performanse su slabije u poređenju sa algoritmima zasnovanim na stablima učenja. Pretpostavlja se da model nasumične šume i *XGBoost* prave manje greške jer bolje upravljaju interakcijama između obeležja, zbog načina na koji dele podatke i kreiraju stabla, dok MLP uči interakcije na globalnom nivou kroz optimizaciju težina. U predikciji trajanje taksi vožnje primetan je isti trend u vrednostima mera evaluacije kao i kod cene vožnje. Takođe i u ovom slučaju najmanju *RMSE* ima *XGBoost* algoritam sa vrednošću 3.98. Na osnovu pomenutih rezultata uočava se da modeli generalno vrše bolje predikcije za cenu u odnosu na trajanje vožnje. To se može objasniti činjenicom da je cena više korelirana sa distancom, koja predstavlja najvažnije obeležje prilikom obučavanja modela.

Pored međusobnog poređenja modela, izvršeno je poređenje najuspešnijeg modela u ovom radu sa odgovarajućim modelima iz sličnih radova. Ovakvo poređenje omogućava da se sagleda da li pristup iz ovog rada nadmašuje već postojeća rešenja i kako može dodatno da se unapredi.

Za model linearne regresije, vrednosti *RMSE* u ovom radu su veće za oba ciljna obeležja u odnosu na rad [1]. Iako postoji razlika, ona je jako mala, odnosno modeli predviđaju na sličan način. Ovakvi rezultati su i očekivani jer su u oba rada korišćeni isti podaci o taksi vožnjama za isti vremenski period.

Rezultati u tabeli 2 pokazuju da je model nasumične šume, prikazan u ovom radu, približan najboljima za oba izlazna obeležja. Zanimljivo je poređenje sa zadnjim radom jer je u njemu model nasumične šume podešen tako da koristi sve hiperparametre sa njihovim podrazumevanim vrednostima. Pošto je razlika u *RMSE* čak 1.5, sledi zaključak da je podešavanje vrednost

određenih hiperparametara uticalo na poboljšanje u odnosu na podrazumevane.

Tabela 2. Poređenje *RMSE* rezultata iz ovog rada sa sličnim rešenjima koristeći model nasumične šune

Model nasumične šume		
Model	**Cena ($)**	**Trajanje (min)**
Model prikazan u ovom radu	2.519	4.851
Model u radu [1]	**2.287**	5.240
Model u radu [2]	-	**4.232**
Model u radu [7]	3.953	-

Što se tiče rezultata za *XGBoost*, najbolji rezultati i veća razlika se zapaža u radu [3] gde *RMSE* iznosi samo 0.41 za predikciju trajanja. Pretpostavka je da je ovo posledica činjenice da se u tom radu koriste podaci u periodu od tri godine, pri čemu je izvršena i klasterizacija podataka. Takođe dodata su nova obeležja koja predstavljaju različite načine izračunavanja distance, pa je samim tim model bolje uspeo da se prilagodi podacima i izvrši približniju predikciju.

Na kraju su upoređeni rezultati *MLP* modela. Bitno je istaći da su arhitekture modela u ovom radu i radu [8] identične- mreža sadrži tri skrivena sloja sa po 20 neurona. Uprkos tome, *MLP* iz ovog rada je ostvario nešto bolje performanse, što može biti posledica različitih tehnika pretprocesiranja podataka.

5. ZAKLJUČAK

U ovom radu je rešavan problem predviđanja cene i trajanja taksi vožnje, sa fokusom na Njujork, grad poznat po velikoj upotrebi taksi usluga. Osnovni cilj sistema je pružanje tačnijih predikcija za ovu vrstu usluge, što bi značajno poboljšalo organizaciju svakodnevnih aktivnosti čoveka i omogućilo bolju optimizaciju vremena u okruženju urbanog života. Implementacija ovog sistema podrazumeva kombinovanje više modula sa definisanim ulogama. Započeto je sa prikupljanjem podataka, gde je prvi skup sadržao informacije o vožnjama za žute taksije u toku maja meseca 2016. godine. Drugi skup podataka je obuhvatao vremenske uslove za isti period, prikupljanje na svakih sat vremena. Isprobani su linearna regresija, model nasumične šume, *XGBoost* i *MLP*. Za svaki od modela izvršena je optimizacija hiperparametara Za potrebe treniranja ali i testiranja modela, skup podataka je podeljen na trening, validacioni i test skup. Budući da je u pitanju regresioni problem, za mere evaluacije su izabrane koren srednje kvadratne greške i koeficijent determinacije.

Poredeći performanse modela za više različitih kombinacija ulaznih obeležja i podataka, svi su se najviše obučili na proširenom skupu podataka. Najbolji rezultati za oba ciljna obeležja su postignuti upotrebom *XGBoost* algoritma, čija *RMSE* za trajanje iznosi 3.48, a za cenu vožnje 2.46. Analizom rezultata dolazi se do zaključka da su svi modeli osetljivi na anomalije u podacima, odnosno da ne uspevaju da nauče odnose između podataka i obeležja ukoliko se nad inicijalnim podacima ne odradi pretprocesiranje. Dodatno je uočeno da dodavanje vremenskih karakteristika nije doprinelo performansama modela u onoj meri koja se očekivala čitajući druge radove iz literature.

Na osnovu rezultata iz rada [3], jedan od načina za unapređenje ovog sistema jeste proširivanje skupa podataka. To znači da bi bilo potrebno posmatrati vožnje za više meseci ili čak za više godina ukoliko to računarski resursi dozvoljavaju. Pored navedenog, sistem bi se mogao poboljšati dodavanjem više različitih izvora podataka, poput podataka o gustini saobraćaja, radovima na putevima, događajima u gradu koji mogu uticati na saobraćajnu gužvu i drugim za određeni vremenski trenutak. Takođe dodatno proširenje bi moglo obuhvatiti bolji pregled geografskih podataka time što bi sistem uzeo u obzir različite karakteristike različitih delova grada.

6. LITERATURA

[1] C. Antoniades, Delara Fadavi, Antoine Foba Amon, "*Fare and Duration Prediction: A Study of New York City Taxi Ride",* Semantic Scholar, 43844792, 2016.

[2] Huang, H., Pouls, M., Meyer, A., Pauly, M, " *Travel Time Prediction Using Tree-Based Ensembles*", Lecture Notes in Computer Science, vol 12433, pp 412–427, 2020.

[3] Poongodi M, Malviya M, Kumar C, "*New York City taxi trip duration prediction using MLP and XGBoost"*, International Journal of System Assurance Engineering and Management, 13(Suppl 1), 16-27, 2022.

[4] https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page (pristupljeno u avgustu 2024.)

[5] https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data/code (pristupljeno u avgustu 2024.)

[6] https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/ (pristupljeno u avgustu 2024.)

[7] *https://github.com/jahnavi-chowdary/New-York-Taxi-Fare-Prediction/tree/master* (pristupljeno u septembru 2024.)

[8] https://github.com/raymonduchen/MLND-P6-New-York-City-Taxi-Fare-Prection/tree/master (pristupljeno u septembru 2024.)

Kratka biografija:

Natalija Krsmanović rođena je 1999.godine u Gradišci, BiH. Osnovne akademske studije završila je 2022. godine na Fakultetu tehničkih nauka, na kom brani master rad 2024. godine iz oblasti Računarstvo i automatika – Inteligentni sistemi.
kontakt:
krsmanovic.natalija99@gmail.com