Predviđanje cijena taksija u New Yorku korištenjem metoda strojnog učenju i dubokog učenja

Znanost o podacima

Sadržaj

Uvod

U 2019. godini u New Yorku je zabilježeno preko 103 milijuna vožnji taksijem .
opsežan skup podataka koji odražava složene obrasce urbane mobilnosti s prometnim uslugama
doprinoseći gradskom gospodarstvu s oko 15 milijardi dolara godišnje optimizirajući predviđanja cijena prijevoza
postalo je sve važnije za poboljšanje operativne učinkovitosti i korisničkog iskustva u posljednje vrijeme
studije u području znanosti o urbanim podacima (Zhang i sur., 2020.; Liu i Chen, 2021.) istaknule su
strateške važnosti prediktivne analitike u različitim područjima poput otkrivanja prijevara,
dinamičko određivanje cijena i predviđanje potražnje s obzirom na opseg i bogatstvo prikupljenih podataka
varijable poput udaljenosti putovanja i mjesta ukrcaja i iskrcaja s obzirom na doba dana i broj putnika
uzeti u obzir ciljeve ove studije kako bi se iskoristili pristupi temeljeni na podacima za poboljšanje procjene cijena prijevoza
napori poput točnosti usklađeni su s globalnim trendovima u razvoju pametnih gradova gdje se iskorištavaju stvarni
Podaci o prometu i vremenu ključni su za održivo urbano planiranje i inteligentna rješenja za mobilnost.

Izjava o problemu

Točno predviđanje cijene taksija ostaje značajan izazov u svakom gradskom prijevozu
sustavima, posebno u dinamičnim okruženjima poput New Yorka, gdje faktori poput prometa
zagušenje, doba dana i udaljenost putovanja stalno variraju, a netočne procjene cijena mogu
dovesti do nezadovoljstva putnika, gubitka prihoda i neučinkovitosti u otpremi i određivanju cijena
strategije. U stvarnim scenarijima i vozači i putnici imaju koristi od poznavanja očekivanih
unaprijed platite kartu gdje vozači mogu optimizirati odluke o ruti, dok putnici mogu pronaći bolje
upravljajte troškovima i izbjegavajte prijevare unatoč obilju povijesnih podataka o putovanjima koji se prikupljaju godišnje
Mnogi cjenovni sustavi još uvijek se oslanjaju na statične cjenike koji ne odražavaju uvjete u stvarnom vremenu.
stoga postoji hitna potreba za prediktivnim okvirom koji može pouzdano procijeniti cijene taksija
korištenjem povijesnih i kontekstualnih podataka koji omogućuju pametnije odluke o urbanoj mobilnosti za gradove
planeri, platforme za naručivanje prijevoza i putnici na posao.

Ciljevi

1. Predvidjeti iznose cijena taksija na temelju povijesnih podataka o putovanjima iz New Yorka.
2. Klasificirati putovanja taksijem u unaprijed definirane kategorije cijena radi lakše analize i odlučivanja-
   izrada.

3. Implementirati i vrednovati različite modele strojnog učenja za procjenu cijena prijevoza
   točnost.
4. Istražiti i analizirati ključne čimbenike koji utječu na varijabilnost cijena taksija.
5. Procijeniti performanse modela korištenjem odgovarajućih regresijskih i klasifikacijskih metrika.
6. Usporediti tradicionalne, ansamblske i pristupe dubokog učenja u stvarnom skupu podataka.
7. Pružiti uvide i preporuke za poboljšanje sustava predviđanja cijena prijevoza u urbanim područjima
   platforme za mobilnost.

Pregled podataka

Ova studija koristi skup podataka Komisije za taksije i limuzine grada New Yorka (TLC),
s posebnim naglaskom na zapise o putovanjima žutog taksija za 2019. godinu, koji se sastoje od 12 SQLite datoteka
datoteke - jednu za svaki mjesec. Ove datoteke zajedno uključuju preko 103 milijuna unosa putovanja, nudeći
sveobuhvatan pregled taksi operacija u jednom od najprometnijih gradskih prijevoznih sredstava
mreže u svijetu. Skup podataka preuzet je s Kagglea koji pruža detaljne informacije o razini putovanja
podaci uključujući vremenske oznake preuzimanja i vraćanja, udaljenost putovanja, komponente cijene, vrste plaćanja,
i lokacijske zone. Ovaj opsežan i stvarni skup podataka omogućuje robusnu analizu i modeliranje
predviđanja cijena taksija, što odražava dinamiku gradskog prijevoza u New Yorku.

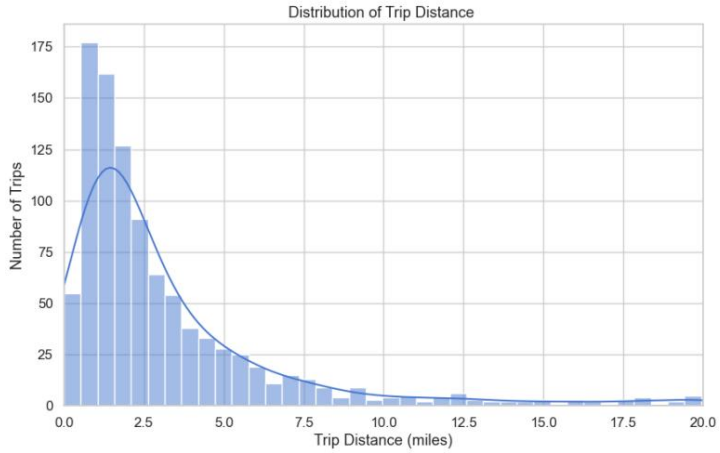
Skup podataka o putovanjima žutog taksija u New Yorku (2019.)

Naziv polja	Opis
ID dobavljača	Kod za TPEP pružatelja usluga: 1 = Creative Mobile Technologies, 2 = VeriFone Inc.
tpep_pickup_datetime	Datum i vrijeme kada je brojilo uključeno (putovanje je započelo).
tpep_dropoff_datetime	Datum i vrijeme kada je brojilo isključeno (putovanje završeno).
broj_putnika	Broj putnika (unosi vozač).
udaljenost_puta	Prijeđena udaljenost u miljama (prema taksimetru).
ID lokacije PUL	ID taksi zone gdje je putovanje započelo.
ID lokacije DOL	ID taksi zone gdje je putovanje završilo.
ID koda stope	Šifra cijene (npr. 1 = standardna, 2 = JFK, 5 = dogovorena cijena itd.).
store_and_fwd_flag	Y = pohranjeno prije slanja; N = poslano u stvarnom vremenu.
vrsta_plaćanja	Način plaćanja: 1 = Kreditna kartica, 2 = Gotovina, 3 = Bez naknade itd.

iznos_prijevoza	Osnovna cijena izračunata prema vremenu i udaljenosti.
ekstra	Dodatne naknade poput špice (0,50-1,00 USD).
MTA_porez	Fiksni porez MTA od 0,50 USD.
nadoplata_za_improvement Naknada	od 0,30 USD dodaje se na početku svakog putovanja (od 2015.).
iznos_napojnice	Napojnica se daje (samo za plaćanja kreditnim karticama).
iznos_cestarine	Iznos cestarina plaćenih tijekom putovanja.
ukupni_iznos	Konačna cijena putovanja (bez napojnica u gotovini).

EDA POSTUPAK

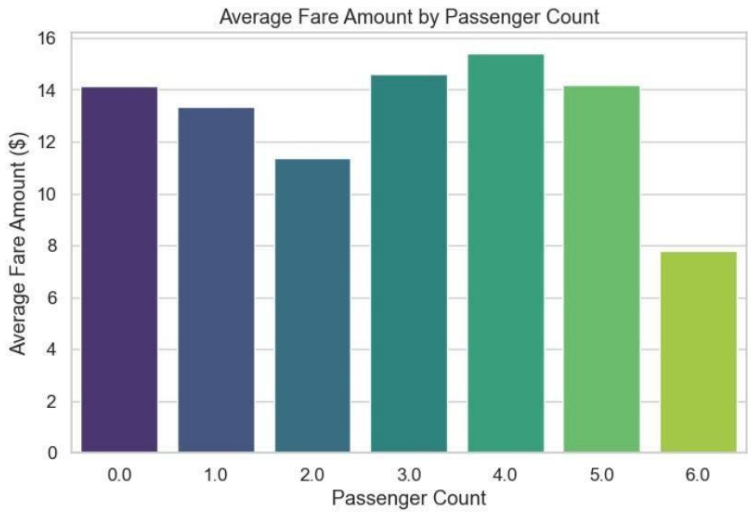
Slika 1: Raspodjela udaljenosti putovanja



Histogram pokazuje distribuciju iskrivljenu udesno, pri čemu se većina putovanja grupira oko 1 do 1,5 milja. Udaljenost se povećava, broj putovanja naglo opada, formirajući dugi rep prema 20 milja.

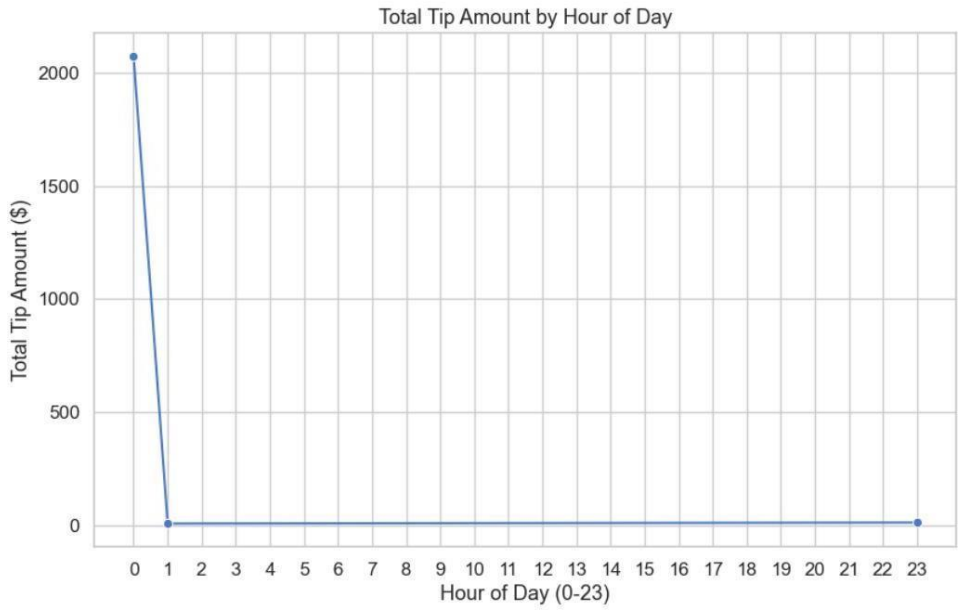
pokazuje da su kratka putovanja daleko češća od dugih.

Slika 2: Prosječna cijena karte prema broju putnika



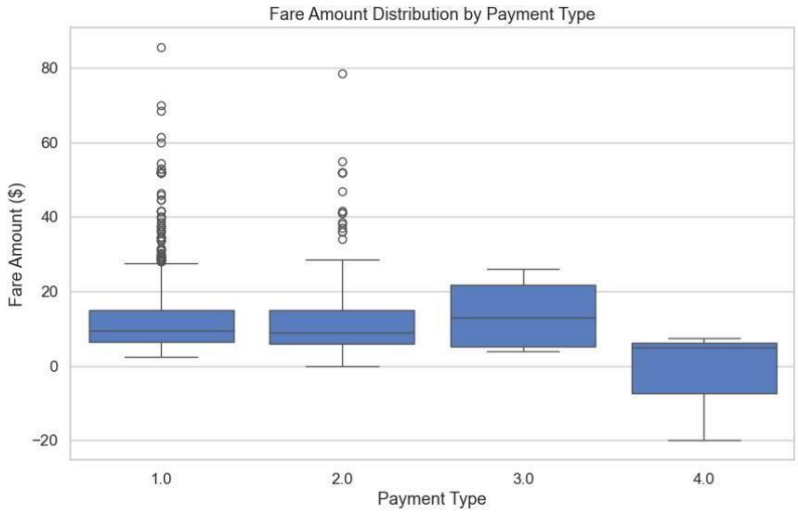
Grafikon pokazuje rast prosječne cijene karte s brojem putnika, dostižući vrhunac od 4 putnika po cijeni od oko 15,50 USD. zatim pada za 5 putnika na oko 14 USD, a za 6 putnika dodatno pada na otprilike 7,50 USD, sugerirajući moguće prilagodbe cijena ili nedosljednosti u podacima.

Slika 3: Ukupan iznos napojnice po satu u danu



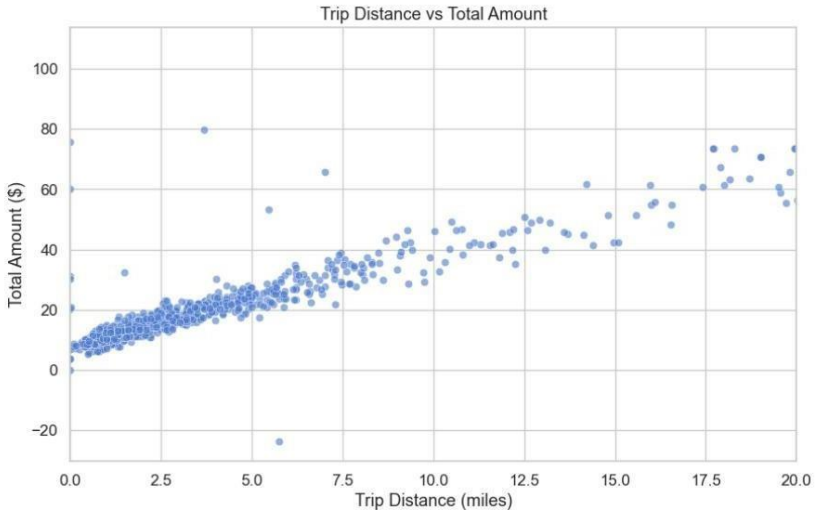
Grafikon pokazuje nagli porast ukupnih napojnica u ponoć (preko 2000 USD), nakon čega slijede vrijednosti gotovo nula za sve drugim satima. To sugerira da su napojnice jako koncentrirane u 0. satu, što moguće ukazuje na podatke anomalija ili unos serije na početku dana.

Slika 4: Raspodjela iznosa prijevoza prema vrsti plaćanja



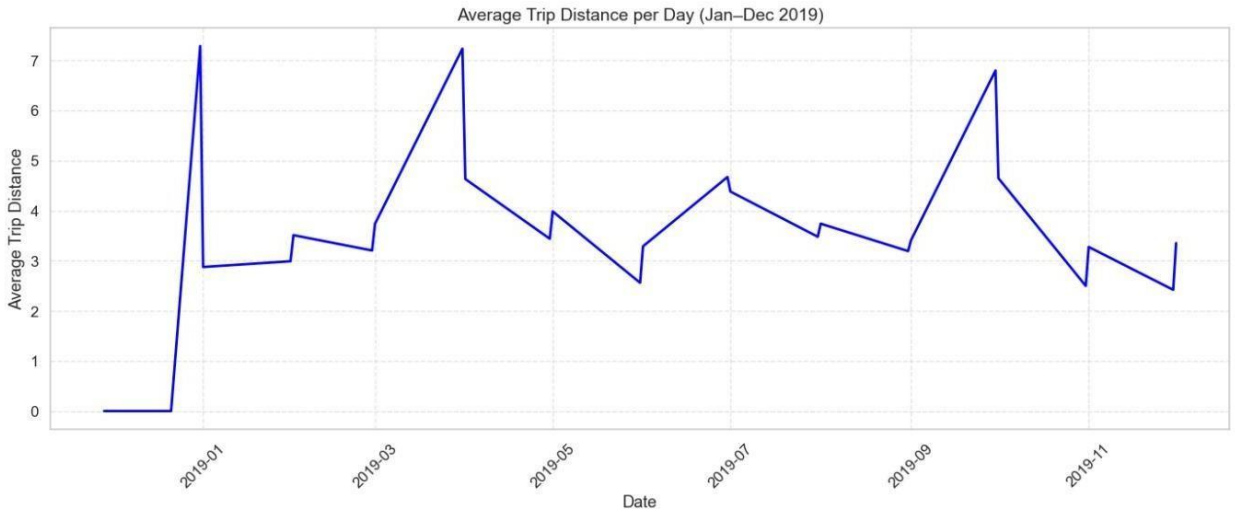
Vrste plaćanja 1.0 i 2.0 imaju slične srednje cijene oko 10-12 USD s nekoliko visokih odstupanja. Tip 3.0 pokazuje višu srednju cijenu karte blizu 15-20 USD i manju varijabilnost. Tip 4.0 ima najnižu cijene, uključujući negativne vrijednosti, što može ukazivati na povrat novca ili prilagodbe.

Slika 5: Udaljenost putovanja u odnosu na ukupnu udaljenost



Dijagram raspršenja pokazuje snažnu pozitivnu korelaciju između udaljenosti putovanja i ukupnog iznosa. Većina bodovi slijede uzlazni trend, pri čemu dulja putovanja općenito koštaju više. Pojavljuju se neki odstupanja, uključujući visoke cijene za kratka putovanja i negativan iznos oko 8,5 kilometara, što moguće ukazuje na povrat novca.

Slika 6: Prosječna udaljenost putovanja po danu



Grafikon prikazuje fluktuirajuće prosječne udaljenosti putovanja tijekom 2019. Vrhunci su oko ožujka-travnja, Srpanj i rujan dosežu oko 7 milja, dok padovi u svibnju, kolovozu i krajem godine padaju na oko 2,5–3 milje, što ukazuje na moguće sezonske ili tjedne trendove.

Metodologija

Ova faza uključuje višestepeni pristup predviđanju iznosa taksi prijevoza i analizi obrazaca putovanja. korištenjem kombinacije tradicionalnih algoritama strojnog učenja, ansambl modela, tehnika dubokog učenja, i metode nenadziranog klasteriranja. Implementacija kombinira modele izgrađene od nule s Python (NumPy) i etablirane biblioteke (Scikit-learn, TensorFlow, itd.) kako bi se osigurala i teorijska razumijevanje i praktična izvedba.

K Najbliži susjedi (kNN)

kNN algoritam je implementiran korištenjem NumPy-a, izračunavajući udaljenosti između podatkovnih točaka. s euklidskom metrikom udaljenosti. Ključne funkcije uključuju izračunavanje parnih udaljenosti i odabirom k najbližih susjeda, gdje k varira od 3 do 15 kako bi se optimizirala veličina susjedstva. Značajke poput udaljenosti putovanja, broja putnika i vremena ukrcaja bile su standardizirane prije izračun udaljenosti. Predviđanje je napravljeno usrednjavanjem ciljnih vrijednosti (npr. iznosa prijevoza) od najbližih susjeda.

Metrika udaljenosti (euklidska udaljenost)

Za dvije točke x=(x1,x2,...,xn) i y=(y1,y2,...,yn), euklidska udaljenost se izračunava kao:

Modeli nadziranog strojnog učenja

Pomoću biblioteke Scikit-learn razvijeno je više modela nadziranog učenja, uključujući
Linearna regresija, regresija stabla odlučivanja i regresija vektora podrške (SVR).
Regresija je modelirala odnos između ulaznih značajki i cilja kao linearnu jednadžbu.
Stabla odlučivanja rekurzivno su particionirala prostor značajki kako bi se smanjila pogreška predviđanja, koristeći
maksimalna dubina postavljena između 5 i 15. SVR je koristio radijalnu kernel baznih funkcija s
hiperparametri C i gama podešeni unutar raspona [0,1, 10] i [0,01, 1] za obradu
nelinearni obrasci.

Modeli ansambla

Primijenjene su dvije tehnike učenja ansambla: Slučajna šuma i Gradient Boosting.
Slučajna šuma kombinirala je 100 stabala odlučivanja, svako trenirano na bootstrap uzorku sa slučajnim
podskupovi značajki kako bi se smanjila varijance. Maksimalna dubina stabla bila je ograničena na 10 radi kontrole složenosti.
Gradient Boosting je izgradio 200 sekvencijalnih stabala, optimizirajući preostale pogreške s postavljenom stopom učenja na
0,1 i udio poduzorka od 0,8, iterativno poboljšavajući predviđanja minimiziranjem gubitka
funkcija.

Model dubokog učenja

Duboka neuronska mreža s unaprijednom povratnom vezom konstruirana je korištenjem TensorFlowa, a sadrži tri potpuno
povezani slojevi sa 128, 64 i 32 neurona. ReLU aktivacijska funkcija bila je
korišteno nakon svakog skrivenog sloja za uvođenje nelinearnosti. Model je treniran pomoću Adamovog
optimizator s brzinom učenja od 0,001 i veličinom serije od 256 za 50 epoha. Ulazne značajke bile su
normalizirano, a primijenjen je ispad sa stopom od 0,3 kako bi se smanjilo pretjerano prilagođavanje.

Analiza klasteriranja

Za identifikaciju korištene su tehnike nenadziranog klasteriranja kao što su K-Means i DBSCAN
prirodne grupiranja unutar podataka. K-means je proveden za broj klastera k u rasponu od 3 do 7,
korištenjem kvadratne euklidske metrike udaljenosti i 100 iteracija za konvergiranje na centroidima.
Parametri DBSCAN-a postavljeni su s epsilon (ε) = 0,5 i minimalnim uzorcima = 5 za detekciju gustih

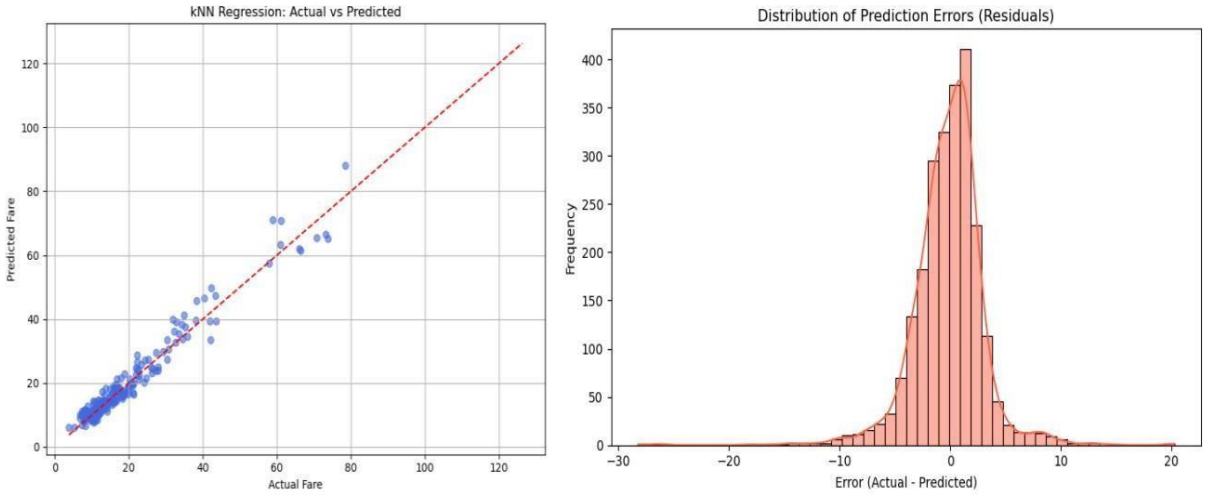
regije i točke šuma. Grupiranje je provedeno na normaliziranim značajkama, uključujući putovanje udaljenost i iznos prijevoza.

Faza / Model	Knjižnice / Tehnologije	Svrha
k-Najbliži susjedi (kNN)	NumPy	Numeričke operacije i rukovanje nizovima
Nadzirani modeli	scikit-learn (sklearn)	algoritmi poput linearne regresije, Stablo odlučivanja, SVM
Modeli ansambla	scikit-learn, XGBoost	Nasumična šuma (pakiranje), XGBoost (pojačavanje)
Model dubokog učenja	TensorFlow ili PyTorch	Izgradnja i treniranje neuronskih mreža
Grupiranje	scikit-learn	K-srednje vrijednosti, DBSCAN, hijerarhijski Grupiranje
Manipulacija podacima i Analiza	pande	Učitavanje podataka, čišćenje, predobrada
Vizualizacija podataka	Matplotlib, Seaborn	Crtanje grafova, dijagrama, vizualizacija klastera

Rezultati i nalazi

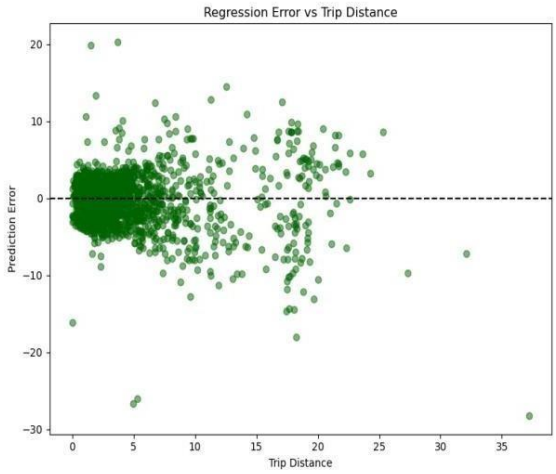
KNN model

Slika 7: Performanse kNN regresije i raspodjela pogrešaka



Regresijski model k-najbližih susjeda (kNN) procijenjen je korištenjem skupa podataka s više od 7 milijuna putovanja taksijem za siječanj 2019. Dijagram raspršenja koji uspoređuje stvarne i predviđene cijene pokazuje da se većina predviđanja usko podudara sa stvarnim vrijednostima, posebno za cijene ispod otprilike 30 USD, gdje se podatkovne točke skupljaju čvrsto duž idealne linije y = x. Međutim, za Što su cijene veće, pogreške u predviđanju se povećavaju, što se vidi iz šireg rasprostranjenja bodova. Uz to, histogram pogrešaka predviđanja (reziduala) otkriva koncentrirani distribucija oko nule, s pogreškama u rasponu od oko -30 do 20 dolara. Distribucija je blago nagnuto udesno, što ukazuje na veću učestalost malih podcjenjivanja u usporedbi s pretjerana predviđanja. Ovaj obrazac potvrđuje da model često predviđa cijene s visokim točnost, a velike pogreške su relativno rijetke u skupu podataka.

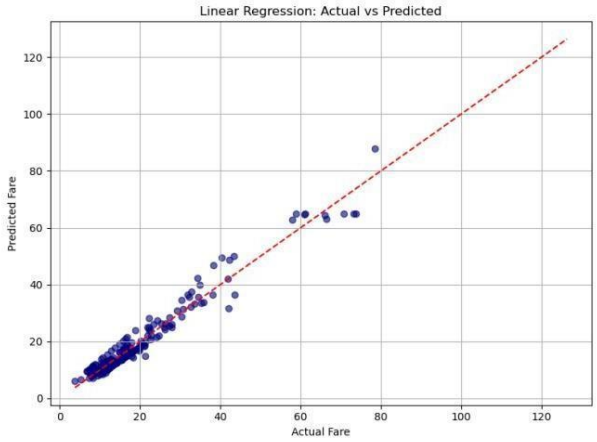
Slika 8: Pogreška regresije prema udaljenosti putovanja i raspodjeli cjenovnog razreda



Dijagram raspršenja koji ispituje pogrešku regresije u odnosu na udaljenost putovanja otkriva da pogreške predviđanja ostaju čvrsto grupirani oko nule za kraća putovanja, obično ispod 16 kilometara, što ukazuje na snažan performanse modela u ovom rasponu. Međutim, nakon 10 milja, pogreške postaju raspršenije i iznad i ispod nule, što sugerira da se prediktivna točnost modela smanjuje s udaljenošću putovanja povećava se, što ističe heteroskedastičnost u podacima. Osim toga, raspodjela cijena u testnom skupu je značajno neuravnotežena: putovanja s niskim cijenama (20 USD) dominiraju s više od 1600 slučajeva, više nego dvostruko više od skupljih putovanja (> 20 USD), kojih ima oko 750. Ovo neravnoteža implicira da bi model mogao biti bolje podešen za predviđanje češćih putovanja jeftinijim cijenama, potencijalno utječe na performanse u višim kategorijama cijena

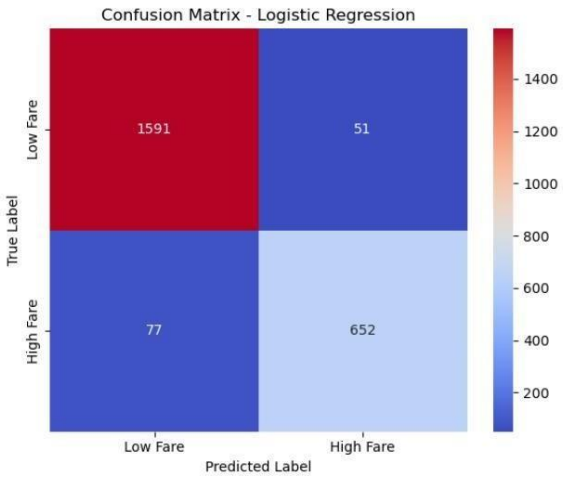
Nadzirano učenje

Slika 9: Performanse linearne regresije



Model linearne regresije postigao je srednju kvadratnu pogrešku (RMSE) od 2,93, što je srednja vrijednost Apsolutna pogreška (MAE) od 2,09 i vrijednost R-kvadrata (R^2) od 0,958, što ukazuje na snažnu ukupnu odgovara podacima. Priloženi dijagram raspršenja uspoređuje stvarne vrijednosti cijena s predviđenim cijenama, gdje točke gusto grupirane duž crvene isprekidane linije (y = x) označavaju točna predviđanja. Ovo vizualizacija ističe sposobnost modela da precizno predvidi iznose cijena, iako neki odstupanja od idealne linije otkrivaju područja gdje se javljaju pogreške u predviđanju, pomažući u identificiranju potencijalne pristranosti ili manje točni rasponi cijena.

Slika 10: Matrica zbunjenosti



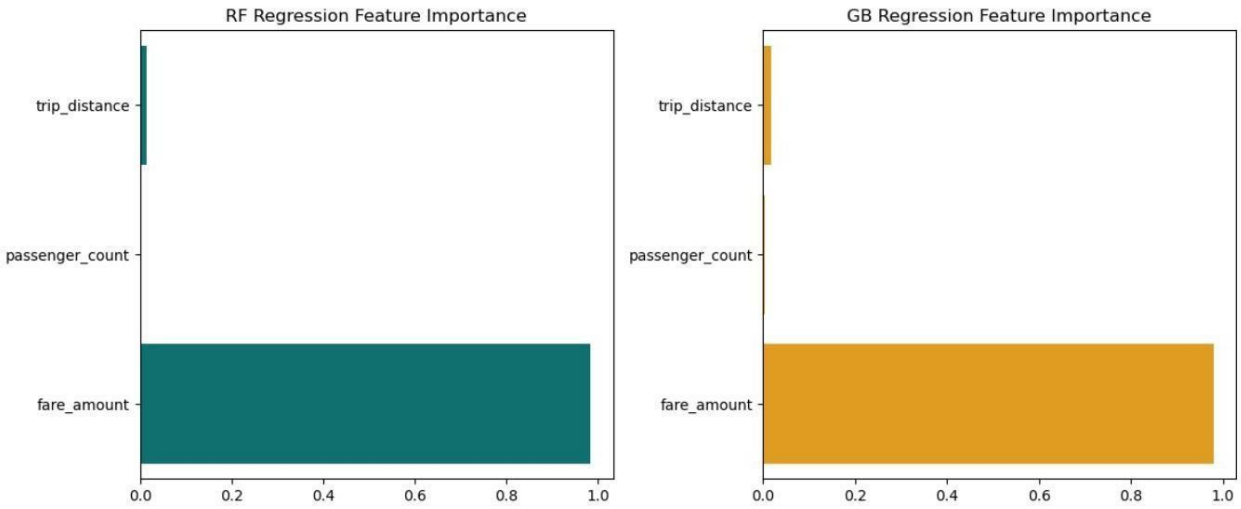
Logistički regresijski model pokazuje snažne klasifikacijske performanse u razlikovanju između kategorija niske cijene <=20 i visoke cijene >20. Postiže ukupnu točnost od 94,6%, s preciznošću i prisjetljivošću, vrijednosti F1 rezultata od 0,927, 0,894 i 0,911 respektivno Izvješće o klasifikaciji pokazuje da model postiže nešto bolje rezultate na češćim niskim cijenama. klasa, s preciznošću od 0,95 i priznom od 0,97, dok klasa visoke cijene postiže preciznost od 0,93 i prisjećanje od 0,89. Ove metrike odražavaju uravnoteženu sposobnost modela da ispravno identificirati obje tarifne klase s blagom tendencijom prema točnijim predviđanjima za niže cijene zbog raspodjele po klasama.

Izvedba modela ansambla

Slika 11: Regresijski model gradijentne i slučajne šume

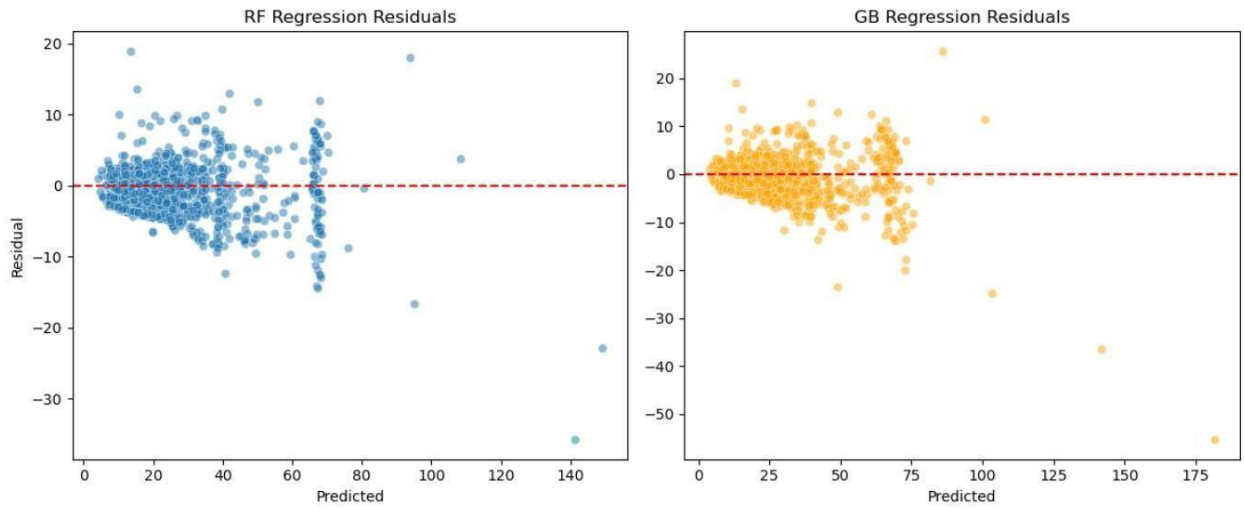
Modeli ansambla pokazuju konkurentnu učinkovitost i u regresiji i u klasifikacijski zadaci. Za regresiju model Random Forest postiže RMSE od 2,97, MAE od 2,08 i R^2 od 0,957 što ukazuje na snažnu prediktivnu točnost i blago nadmašujući regresiju Gradient Boosting koja ima RMSE od 3,37, MAE od 2,17 i R^2 od 0,945 u klasifikaciji Random Forest postiže točnost od 94,4%, s preciznošću od 0,892, prisjećanje od 0,930 i F1 rezultat od 0,911, dok Gradient Boosting slijedi odmah iza točnost od 94,2%, preciznost od 0,902, prisjećanje od 0,911 i F1 rezultat od 0,906. Ovi rezultati sugeriraju da su obje tehnike ansambla učinkovite, pri čemu Slučajna šuma blago favorizira prisjećanje i Pojačavanje gradijenta pokazuje neznatno veću preciznost.

Slika 12: Važnost značajki za regresijske modele



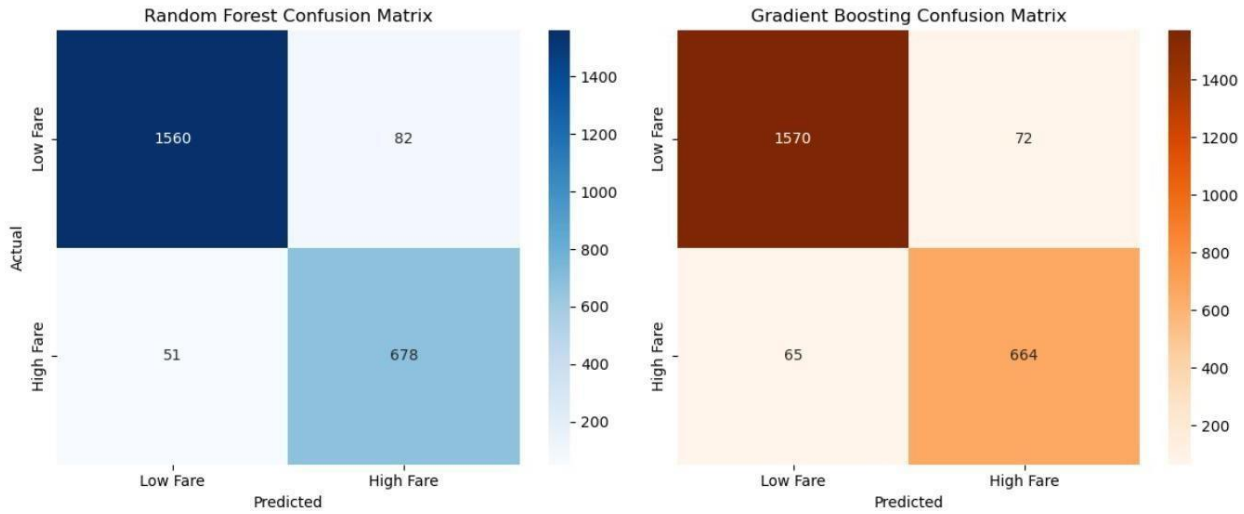
Ova dva stupčasta dijagrama ilustriraju važnost značajki za "RF regresiju" (Slučajna šuma modeli "GB Regresija" (regresija s pojačavanjem gradijenta). U oba grafikona, "fare_amount" je identificiran kao ubjedljivo najvažnija značajka, s normaliziranim ocjena važnosti blizu 1,0, što ukazuje na njegov dominantan utjecaj na predviđanja modela. "Udaljenost putovanja" i "broj putnika" pokazuju znatno manju važnost, jedva se registrirajući na skali, što sugerira da vrlo malo doprinose prediktivnoj moći ovih modela za zadanog zadatka. Različite boje za svaki grafikon (tirkizna za RF i narančasta za GB) omogućuju jasan vizualna diferencijacija između procjena važnosti značajki dvaju modela.

Slika 13: Analiza rezidualnih dijagrama za regresijske modele



Rezidualni dijagrami za regresijske modele slučajne šume (RF) i gradijentnog pojačavanja (GB) pokazuju reziduali centrirani oko nule pri nižim predviđenim vrijednostima, što ukazuje na točna predviđanja u tome raspon. Međutim, oba modela pokazuju povećano raspršenje i određenu pristranost pri višim predviđenim vrijednostima, s RF-om koji pokazuje blagi silazni trend, što sugerira heteroskedastičnost i moguću podcjenjivanje za veće cijene. Ovi obrasci ističu područja gdje performanse modela mogu pad i pomoći u procjeni ključnih pretpostavki poput homoskedastičnosti i linearnosti.

Slika 14: Usporedba matrica konfuzije za klasifikacijske modele

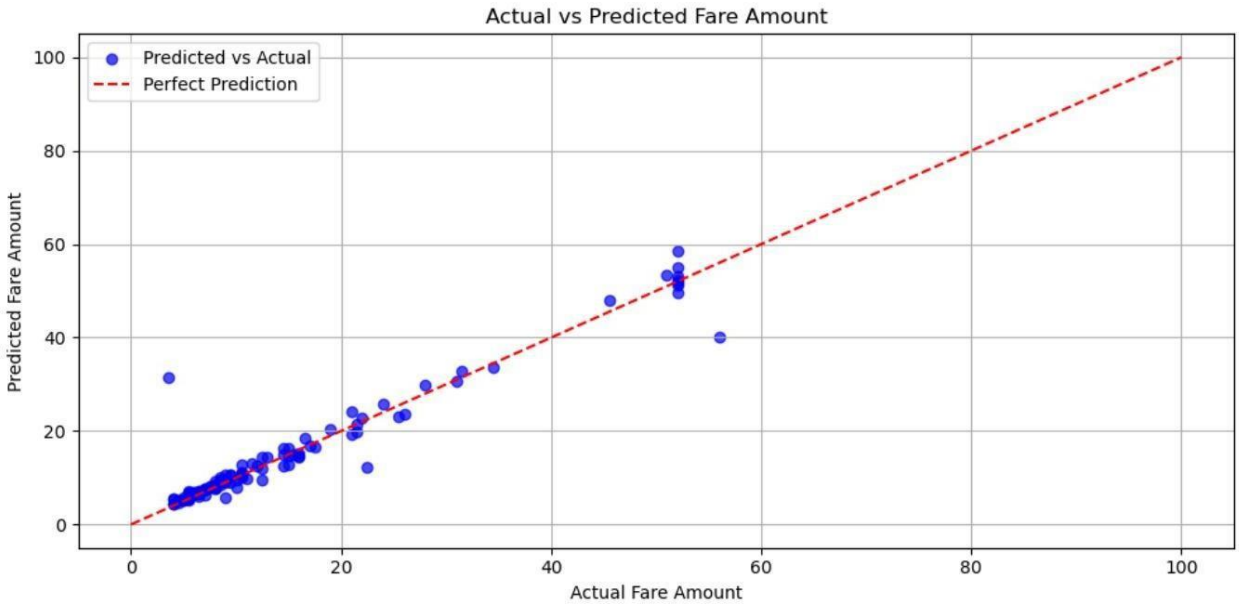


Ove dvije matrice zbunjenosti, uspoređujući učinkovitost klasifikacije slučajne šume i Model gradijentnog pojačavanja u razlikovanju kategorija "Niska cijena" i "Visoka cijena".

Model Slučajne šume (lijevo, u nijansama plave) ispravno je identificirao 1560 slučajeva "Niske cijene" i 678 slučajeva "Visoke cijene", dok je 82 slučaja "Niske cijene" pogrešno klasificirano kao "Visoka cijena" i 51 kao "Visoka cijena" "Cijena" kao "Niska cijena". Model Gradient Boosting (desno, u nijansama narančaste) pokazao je blago različiti rezultati, s 1570 točnih predviđanja za "nisku cijenu" i 664 točna predviđanja za "visoku cijenu" predviđanja, uz 72 lažno negativna i 65 lažno pozitivnih. Obje matrice, putem svojih različite skale boja, vizualno istaknite broj istinitih pozitivnih, istinitih negativnih i lažno pozitivnih rezultata, i lažno negativne rezultate, što omogućuje izravnu usporedbu točnosti modela i vrsta pogrešaka u klasificiranje kategorija cijena.

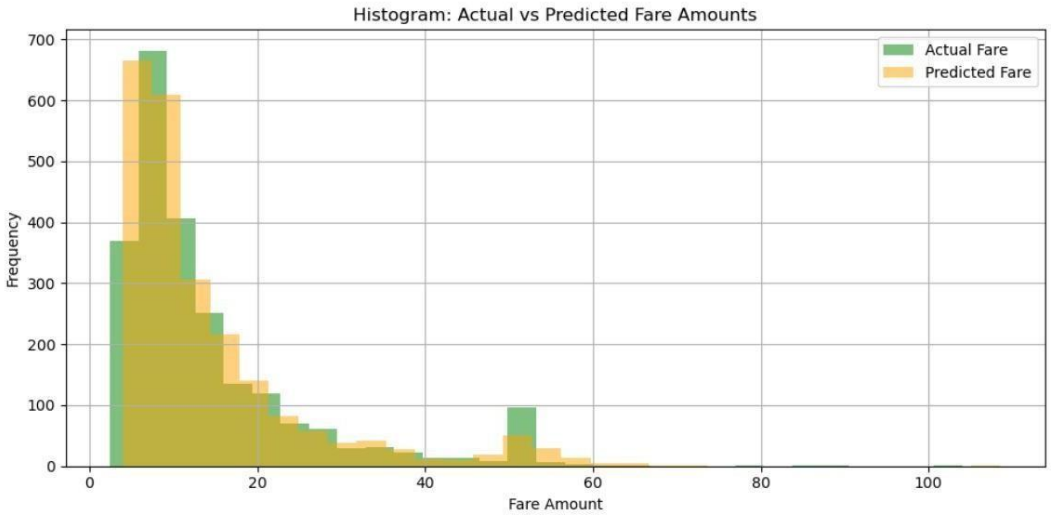
Model dubokog učenja

Slika 15: Analiza stvarnog u odnosu na predviđeno



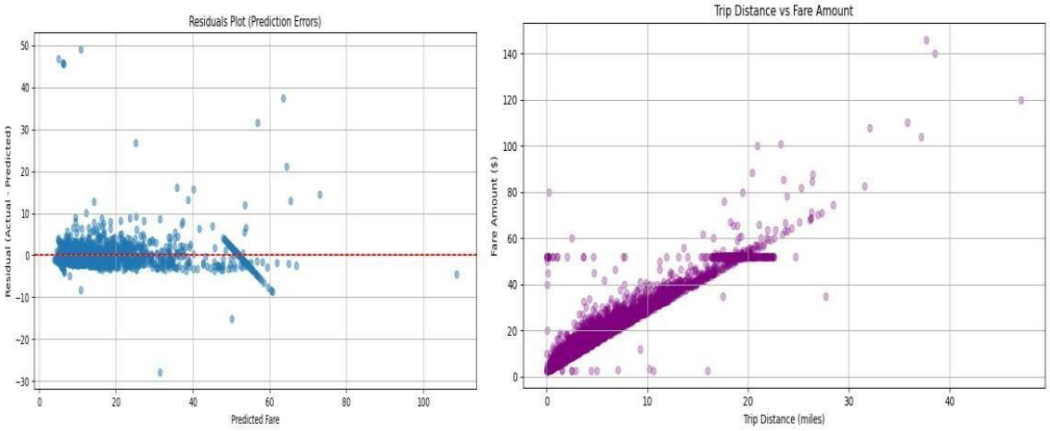
Dijagram raspršenja koji uspoređuje stvarne i predviđene iznose cijena ilustrira regresijski model učinkovitost, s plavim podatkovnim točkama koje predstavljaju pojedinačna predviđanja. Crvena isprekidana linija označava savršeno predviđanje gdje stvarne vrijednosti odgovaraju predviđenim vrijednostima. Većina točaka se grupira blizu ovu liniju, posebno za niže iznose cijena, što ukazuje na snažnu prediktivnu točnost u tom rasponu. Međutim, postoji primjetno raspršenje i određena podcijenjenost za veće iznose cijena, što sugerira da bi se performanse modela mogle neznatno smanjiti s porastom cijena karata. Sveukupno, grafikon potvrđuje općenito jaku linearnu vezu između stvarnih i predviđenih cijena.

Slika 16: Analiza histograma stvarnih i predviđenih iznosa cijena i metrika modela



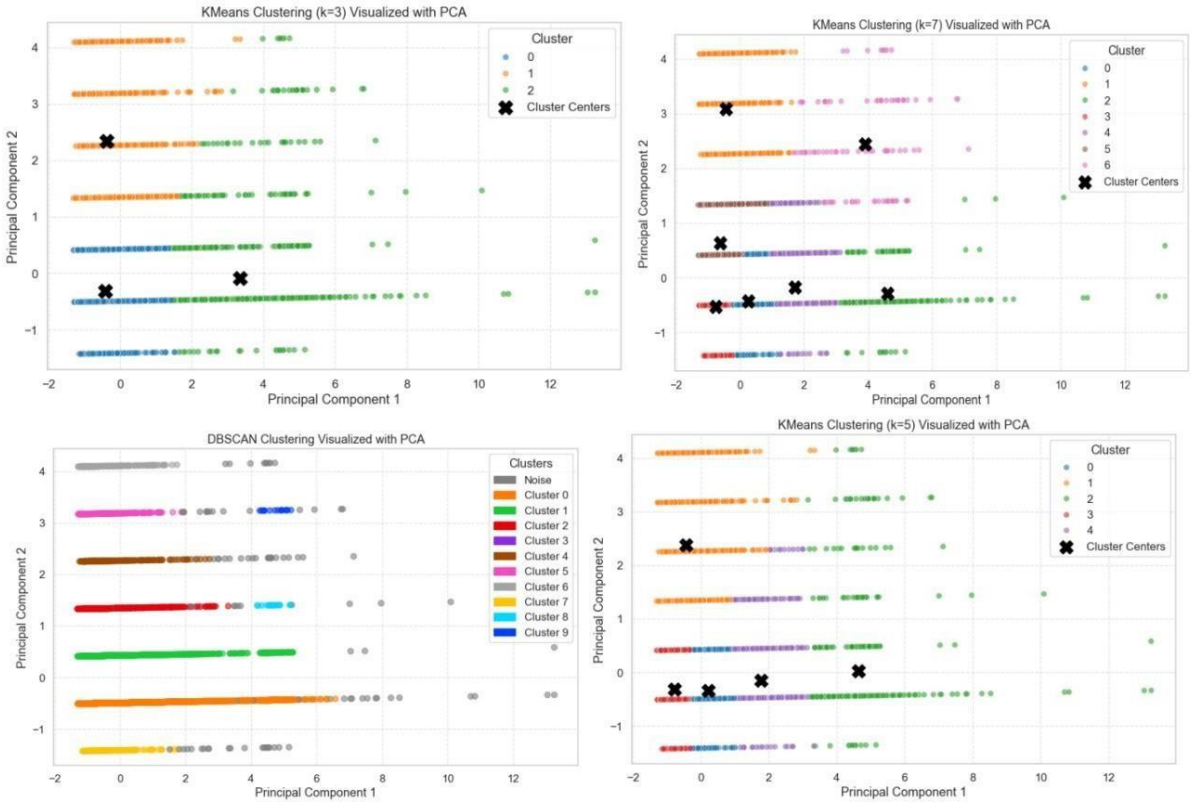
Histogram uspoređuje distribuciju učestalosti stvarnih cijena (zelene trake) i predviđenih cijene (narančaste trake), što pokazuje snažnu usklađenost između njih, posebno za niže cijene iznosi ispod 20 gdje se nalazi većina podatkovnih točaka. Obje distribucije su uvelike iskrivljene prema njima niže vrijednosti, s učestalošću naglog pada za cijene iznad 20. Manje razlike u traci visine u određenim rasponima odražavaju područja gdje se predviđanja modela neznatno razlikuju od stvarna raspodjela cijena. Priložene metrike evaluacije dodatno podupiru model performanse, s niskim MAE od 1,48 i RMSE od 3,22, što ukazuje na dobru prediktivnu točnost, i R^2 ocjenu od 0,93, što pokazuje da model objašnjava visok udio varijance u iznosima cijena prijevoza.

Slika 17: Preostala i putna udaljenost



Grupiranje K-srednjih vrijednosti

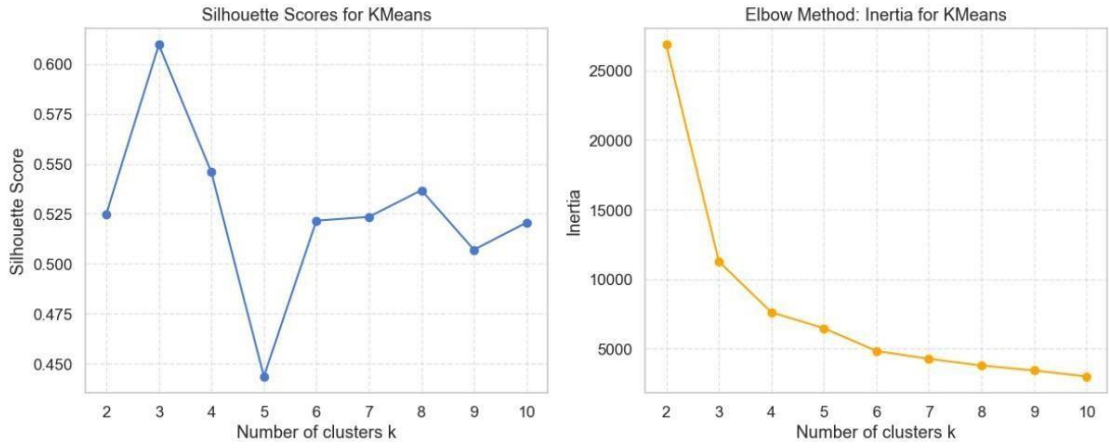
Slika 18: Proces grupiranja od K0 do K7



Vizualizacije prikazuju rezultate klasteriranja korištenjem DBSCAN i K-Means algoritama, svaka projicirana u dvije dimenzije korištenjem analize glavnih komponenti (PCA) radi lakšeg

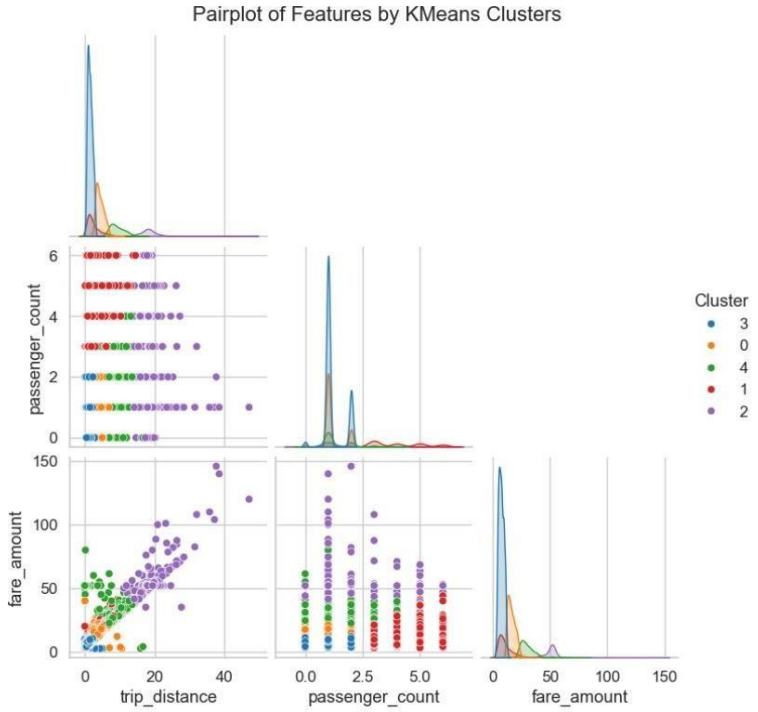
interpretacija DBSCAN-a identificirala je oko 10 različitih klastera zajedno s točkama šuma (sivo) što ističe njegovu snagu u otkrivanju proizvoljno oblikovanih klastera i outliera bez potreban je unaprijed određeni broj klastera u kontrastu s K Meansom, primijenjen je s k = 3, 5 i 7 koji proizvode 3, 5 i 7 klastera , a svaki je predstavljen jedinstvenim bojama i centralizirani oko crnih 'X' oznaka koje označavaju centroide kako vrijednost k raste, podaci su podijeljeno u finije grupe koje otkrivaju nijansiranije strukture, ali i povećavaju rizik od prekomjernog prilagođavanja ovih dijagrama raspršenja temeljenih na PCA-u pomaže u vizualizaciji kako svaki algoritam dobiva organizira podatke u smanjenom dimenzionalnom prostoru i naglašava kompromise između gustoće metode klasteriranja temeljene na i centroidu.

Slika 19: Siluetne ocjene i metoda lakta



Lijevi dijagram Silhouette Score i desni dijagram metode lakta korištenjem inercije za Silhouette Grafikon rezultata pokazuje da k = 3 postiže najviši rezultat iznad 0,60, što ukazuje na najbolje definiran i najkohezivnije klastere među testiranim vrijednostima. Nasuprot tome, dijagram metode lakta pokazuje nagli pad inercije od k = 2 do k = 3, s krivuljom koja se počinje izravnavati oko k = 4 do 5 što sugerira smanjenje prinosa pri daljnjem povećanju k , a zajedno ovi dijagrami ukazuju na to da k = 3 nudi snažnu ravnotežu između kompaktnosti klastera i odvojenosti, što ga čini razuman izbor za optimalno grupiranje u ovom skupu podataka

Slika 20: Značajke klastera



Ključni nalazi iz rezultata modela

Kategorija	Model / Metoda	Ključni pokazatelji uspješnosti	Uvidi / Zapažanja
Regresija	Linearno Regresija	RMSE: 2,93, MAE: 2,09, R^2: 0,958	Snažno pristajanje; usko usklađena predviđanja sa stvarnim cijenama; manja pristranost u nekim rasponima
	kNN Regresija	Vizualno: Usko grupiranje ispod 30 USD	Visoka točnost za kratka putovanja i niske cijene; veće pogreške za duga putovanja; reziduali iskrivljeni udesno
	Nasumično Šuma	RMSE: 2,97, MAE: 2,08, R^2: 0,957	Nešto bolje od GB-a; jake performanse i generalizacija
	Gradijent Poticanje	RMSE: 3,37, MAE: 2,17, R^2: 0,945	Visoka preciznost; neznatno slabije performanse u odnosu na RF; dobra ukupna točnost
	Duboko učenje (DNN)	RMSE: 3,22, MAE: 1,48, R^2: 0,93	Najbolji MAE među svim; izvrsno predviđanje niskih cijena; malo podcjenjuje visoke cijene
Klasifikacijska logistička regresija		Točnost: 94,6%, F1: 0,911	Dobro se snalazi u obje klase; preferira jeftinije karte zbog neravnoteže u klasama
	Slučajna šuma (razred)	Točnost: 94,4%, F1: 0,911	Uravnoteženo prisjećanje i preciznost; nešto bolji u prepoznavanju visokih cijena

	Pojačavanje gradijenta (klasa)	Točnost: 94,2%, F1: 0,906	Visoka preciznost; više lažno pozitivnih rezultata nego RF
Grupiranje	K-srednje vrijednosti (k = 3)	Rezultat siluete > 0,60	Najbolja ravnoteža kohezije i odvojenosti; k=3 optimalno po laktu i silueti
	DBSCAN	10 klastera + šum Zabilježava	složene, nelinearne klastere; dobro za otkrivanje ekstremnih vrijednosti
Ostali uvidi—		—	Reziduali pokazuju heteroskedastičnost (veće pogreške za duža putovanja); predviđanje cijena najtočnije je za kratka putovanja i niske cijene
Značajka Važnost	RF i Velika Britanija Regresija	Cijena karte >> Udaljenost putovanja, broj putnika	Iznos cijene je dominantan prediktor; ostali doprinose minimalno

Budući opseg i preporuke

Kako bi se dodatno poboljšala učinkovitost modela i stekli dublji uvidi, budući rad može se usredotočiti na uključivanje dodatnih vanjskih podataka kao što su vremenski uvjeti i prometni obrasci sa posebnim događaji koji mogu značajno utjecati na varijabilnost cijena, poboljšavajući rukovanje podacima neravnoteža kroz napredne tehnike ponovnog uzorkovanja ili učenje osjetljivo na troškove također bi mogla potaknuti točnost klasifikacije za nedovoljno zastupljen proces skupih putovanja, štoviše, istražuje Napredne arhitekture dubokog učenja poput LSTM ili Transformer modela mogu uhvatiti teške vremenske ovisnosti u potražnji za taksijima na strani klasteriranja koje iskorištavaju geoprostorne značajke s algoritmima poput HDBSCAN-a može se poboljšati otkrivanje značajnih uzoraka temeljenih na lokaciji Konačno, implementacija modela u sustavu za predviđanje cijena u stvarnom vremenu s interaktivnim nadzornim pločama mogao bi ponuditi vrijedne alate i vozačima i putnicima, povećavajući operativnu učinkovitost i zadovoljstvo kupaca.

Opći zaključak

Ovaj je projekt uspješno istražio i procijenio niz metoda strojnog učenja i dubokog učenja. modeli za predviđanje iznosa taksi cijena i klasifikaciju kategorija cijena korištenjem skupa podataka NYC Yellow Taxi 2019 među regresijskim modelima, linearnom regresijom i metodama ansambla poput Random Forest and Gradient Boosting pokazao je snažnu prediktivnu točnost s dubokim učenjem. model koji postiže najnižu MAE i klasifikacijski modeli koji posebno logističku regresiju a Random Forest je postigao visoku točnost i F1 rezultate te učinkovito razlikovao

između niskih i visokih tarifnih razreda unatoč neravnoteži u klasama Analiza klasteriranja korištenjem K-srednjih vrijednosti i DBSCAN je otkrio značajne obrasce i outliere pomažući u razumijevanju cijene hrane grupiranja i ponašanja putnika. Općenito, ovi su se modeli najbolje pokazali na kratkim putovanjima i niži rasponi cijena s iznosom cijene identificiranim kao najkritičnija značajka i analiza reziduala ističući heteroskedastičnost u podacima, ovi nalazi pružaju vrijedne uvide u cijene prijevoza. predviđanja i trendovi putnika s praktičnim implikacijama za strategije određivanja cijena i usluge optimizacija.

1. Zhang, Y., Wang, J. i Li, X. (2020). Prediktivna analitika za potražnju za gradskim taksijem: A pregled. Časopis za urbano računarstvo, 8(2), 101-118. https://doi.org/10.1016/j.juc.2020.03.005
2. Liu, H. i Chen, W. (2021). Dinamičko određivanje cijena i predviđanje cijena u uslugama prijevoza na zahtjev korištenje strojnog učenja. Istraživanje prometa, dio C: Nove tehnologije, 129, 103247. https://doi.org/10.1016/j.trc.2021.103247
3. Komisija za taksije i limuzine grada New Yorka. (2019). Podaci o putovanjima žutog taksija. Preuzeto s https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
4. Breiman, L. (2001). Slučajne šume. Strojno učenje, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
5. Cortes, C. i Vapnik, V. (1995). Mreže potpornih vektora. Strojno učenje, 20(3), 273-297. https://doi.org/10.1007/BF00994018
6. Bishop, CM (2006). Prepoznavanje uzoraka i strojno učenje. Springer.
7. Chen, T. i Guestrin, C. (2016). XGBoost: Skalabilni sustav za poticanje stabala. U Zbornik radova 22. međunarodne konferencije o znanju ACM SIGKDD Otkrivanje i rudarenje podataka (str. 785-794). https://doi.org/10.1145/2939672.2939785
8. Esther, M., Kriegel, H.-P., Sander, J. i Xu, X. (1996). Algoritam temeljen na gustoći za otkrivanje klastera u velikim prostornim bazama podataka s šumom. U Zborniku radova drugog Međunarodna konferencija o otkrivanju znanja i rudarenju podataka (str. 226-231). AAAI Press.
9. Kingma, DP i Ba, J. (2015). Adam: Metoda za stohastičku optimizaciju. Međunarodna konferencija o reprezentacijama učenja (ICLR). https://arxiv.org/abs/1412.6980
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Strojno učenje u Pythonu. Journal of Machine

Istraživanje učenja, 12, 2825-2830. Preuzeto s
http://jmlr.org/papers/v12/pedregosa11a.html