Izvješće s sredine putovanja

Faza 1: Formulacija problema

Definicija problema:

Skup podataka korišten u ovoj analizi sastoji se od zapisa o vožnji taksijem u New Yorku iz 2019. Podaci uključuju detalje o vožnji taksijem kao što su vremenske oznake ukrcaja i iskrcaja, broj putnika, udaljenost putovanja, iznos cijene i ID-ovi lokacija.

Ciljevi:

- Razumjeti obrasce vožnje tijekom vremena i geografije.
- Utvrditi korelacije i anomalije u cijeni prijevoza, udaljenosti i broju putnika.
- Očistiti i pripremiti podatke za izgradnju modela.
- Osmisliti nove značajke koje mogu poboljšati prediktivne performanse.
- Primijeniti redukciju dimenzionalnosti za prepoznavanje uzoraka.
- Odaberite odgovarajući model za prediktivne ili deskriptivne zadatke.

---

Faza 2: Analiza i čišćenje podataka

Izvor skupa podataka:

- Skup podataka: Putovanja taksijem u New Yorku 2019.
- Izvor: Kaggle (Skup podataka)

Koraci predobrade:

- Podaci su učitani pomoću Pandasa.
- Stupci s previše nedostajućih ili nebitnih vrijednosti su odbačeni.
- Primijenjeno je filtriranje za uklanjanje nevažećih podataka (npr. negativne cijene ili nula udaljenosti).
- Polja datuma i vremena su analizirana kako bi se izdvojile značajke dana, mjeseca i sata .
- Kategorični payment_type i RatecodeID kodirani su za modeliranje.
- Značajke cijene i udaljenosti skalirane su pomoću MinMaxScalera i StandardScaler.

Čišćenje i standardizacija podataka:

- Iznimke u udaljenostima putovanja i iznosima prijevoza uklonjene su na temelju određivanja praga.
- Broj putnika ograničen na 1–6 radi provjere mentalne ispravnosti.
- Nedostajuće vrijednosti su ili odbačene ili imputirane.

---

Analiza istraživačkih podataka (EDA)

Deskriptivna statistika i vizualizacije:

- Histogrami prikazani za udaljenost putovanja, cijenu prijevoza i broj putnika.
- Kutijasti dijagrami koji se koriste za otkrivanje odstupanja u cijeni prijevoza i udaljenosti.
- Toplinske karte i dijagrami raspršenja korišteni za korelacijsku analizu.

Ključni identificirani obrasci:

- Pozitivna korelacija između cijene prijevoza i udaljenosti putovanja.
- Vikendi i večeri pokazali su veću potražnju za vožnjom.
- Većina vožnji imala je 1-2 putnika, što ukazuje na tipično gradsko putovanje.

Smanjenje dimenzije:

- Za linearnu redukciju primijenjena je PCA (analiza glavnih komponenti) .
    - Pomoglo je u razumijevanju varijance značajki.
- UMAP (Uniformna mnogostruka aproksimacija i projekcija) koristi se za nelinearne projekcija.
    - Pokazalo se bolje razdvajanje klastera za vrste putovanja.

Početni uvidi:

- Izračun cijene ne ovisi samo o udaljenosti - vjerojatno uključuje vrijeme, lokaciju i vrstu plaćanja.
- Određeni parovi lokacija imali su visoku učestalost, što ukazuje na vruća mjesta putovanja na posao.

---

Testiranje hipoteza

Formulirane hipoteze:

- $H_0$ (Nulta hipoteza): Nema razlike u prosječnoj cijeni između plaćanja gotovinom i plaćanja karticama.

- $H_0$ (Alternativna hipoteza): Postoji statistički značajna razlika u prosječnoj cijeni vožnje na temelju vrste plaćanja.

Korišteni statistički test:

- Primijenjen je t-test s dva uzorka između iznosa prijevoza za gotovinu i karticu plaćanja.

Rezultati:

- Test je odbacio $H_0$, što ukazuje na značajnu razliku u cijenama prijevoza po načinu plaćanja metoda.

---

Faza 3: Odabir modela

Inženjering značajki:

Stvoreno je najmanje 10 novih značajki:

1. Sat u danu
2. Dan u tjednu
3. Mjesec
4. Je li vikend
5. Trajanje putovanja (procijenjeno)
6. Brzina putovanja (udaljenost/vrijeme)
7. Je li špica
8. Vrsta područja preuzimanja (gradsko/prigradsko ovisno o zoni)
9. Učini to jednu milju
10. Značajka razvrstana po udaljenosti (kratka, srednja, duga)

Kandidati za modele:

- Linearna regresija: Za početnu vrijednost predviđanja cijena.
- Slučajna šuma: Za rukovanje nelinearnim relacijama i mješovitim podacima.
- Pojačavanje gradijenta (XGBoost): Za visoku točnost i robusnost.
- K-srednje vrijednosti: Za nenadziranu analizu uzoraka na lokacijama.

Strategija validacije:

- Za jednostavne modele korištena je podjela vlaka/testa (80/20) .
- Za konačno podešavanje modela uzeta je u obzir unakrsna validacija K-Fold-a .

Opravdanje:

- Raspodjela značajki bila je iskrivljena; stoga su modeli temeljeni na stablima imali bolje rezultate nego linearni modeli.
- Smanjenje dimenzionalnosti pomoglo je smanjiti složenost modela.
- Skaliranje i kodiranje značajki osigurali su poštene usporedbe između algoritama.