Izvješće o napretku projekta (Mid-Journey Report)

Faza 1: Formulacija problema

Definicija problema:
Skup podataka korišten u ovoj analizi sastoji se od **zapisa o vožnjama taksijem u New Yorku (New York City Taxi Trip Records) iz 2019. godine**. Podaci uključuju detalje o vožnjama poput vremena polaska i dolaska, broja putnika, udaljenosti putovanja, iznosa vožnje i identifikacijskih oznaka lokacija (location IDs).

Ciljevi:

- Razumjeti obrasce vožnji tijekom vremena i po geografskim područjima.
- Identificirati korelacije i anomalije u iznosu vožnje, udaljenosti i broju putnika.
- Očistiti i pripremiti podatke za izradu modela.
- Kreirati nove značajke (features) koje mogu poboljšati prediktivnu izvedbu.
- Primijeniti smanjenje dimenzionalnosti za prepoznavanje obrazaca.
- Odabrati odgovarajući model za prediktivne ili deskriptivne zadatke.

Faza 2: Analiza i čišćenje podataka

Izvor skupa podataka:

- Skup podataka: NYC Taxi Trips 2019
- Izvor: Kaggle (Skup podataka)

Koraci predobrade:

- Podaci su učitani pomoću biblioteke Pandas.
- Stupci s previše nedostajućih ili nevažnih vrijednosti su uklonjeni.
- Primijenjeno je filtriranje kako bi se uklonili neispravni podaci (npr. negativni iznosi vožnji ili nulte udaljenosti).
- Polja s datumom i vremenom (datetime) su raščlanjena (parsed) kako bi se izdvojile značajke za **dan, mjesec i sat**.
- Kategorijske značajke payment_type i RatecodeID su enkodirane za modeliranje.
- Značajke iznosa vožnje i udaljenosti skalirane su pomoću MinMaxScaler i StandardScaler.

Čišćenje i standardizacija podataka:

- Atipične vrijednosti (outliers) u udaljenostima i iznosima vožnji uklonjene su na temelju pragova.
- Broj putnika ograničen je na 1–6 radi provjere ispravnosti podataka (sanity checks).

- Nedostajuće vrijednosti su ili uklonjene ili imputirane (popunjene).

Eksploratorna analiza podataka (EDA)

Deskriptivna statistika i vizualizacije:

- Iscrtani su histogrami za **udaljenost putovanja, iznos vožnje i broj putnika**.
- Korišteni su dijagrami s pravokutnicima i crtama (box plots) za otkrivanje **atipičnih vrijednosti (outliers)** u iznosu vožnje i udaljenosti.
- Korištene su toplinske mape (heatmaps) i dijagrami raspršenja (scatter plots) za analizu korelacije.

Ključni identificirani obrasci:

- **Pozitivna korelacija** između iznosa vožnje i udaljenosti putovanja.
- Vikendi i večernji sati pokazali su **veću potražnju za vožnjama**.
- Većina vožnji imala je **1–2 putnika**, što ukazuje na tipičan gradski prijevoz.

Smanjenje dimenzionalnosti:

- Primijenjena je **PCA (Analiza glavnih komponenti)** za linearno smanjenje.
    - Pomogla je u razumijevanju varijance značajki.
-
- Korišten je **UMAP (Uniform Manifold Approximation and Projection)** za nelinearnu projekciju.
    - Pokazao je bolje razdvajanje klastera za različite tipove vožnji.
-

Početni uvidi:

- Izračun cijene vožnje **ne ovisi samo o udaljenosti**—vjerojatno uključuje i vrijeme, lokaciju i vrstu plaćanja.
- Određeni parovi lokacija imali su **visoku učestalost**, što ukazuje na prometna čvorišta (commute hotspots).

Testiranje hipoteza

Formulirane hipoteze:

- **H_0 (Nulta hipoteza):** Ne postoji razlika u prosječnom iznosu vožnje između plaćanja gotovinom i karticom.
- **H_a (Alternativna hipoteza):** Postoji statistički značajna razlika u prosječnom iznosu vožnje ovisno o vrsti plaćanja.

Korišteni statistički test:

- Primijenjen je **t-test za dva nezavisna uzorka** na iznose vožnji za plaćanja gotovinom i karticom.

Rezultati:

- Test je **odbacio H_0**, što ukazuje na značajnu razliku u iznosima vožnji ovisno o načinu plaćanja.

Faza 3: Odabir modela

Inženjering značajki (Feature Engineering):
Kreirano 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 prometna špica**
8. **Tip područja polaska (gradsko/prigradsko na temelju zone)**
9. **Cijena po milji**
10. **Grupirana udaljenost (kratka, srednja, duga)**

Kandidati za modele:

- **Linearna regresija:** Kao osnovni (baseline) model za predviđanje cijene vožnje.
- **Slučajna šuma (Random Forest):** Za obradu nelinearnih odnosa i mješovitih podataka.
- **Gradijentno pojačavanje (XGBoost):** Za visoku točnost i robusnost.
- **K-Means:** Za nenadziranu analizu obrazaca na lokacijama.

Strategija validacije:

- **Podjela na skup za učenje i testiranje (80/20)** korištena je za jednostavne modele.
- **Unakrsna validacija s K-preklopa (K-Fold Cross-Validation)** razmatrana je za konačno podešavanje modela.

Obrazloženje:

- Distribucije značajki bile su asimetrične (skewed); stoga su modeli temeljeni na stablima (tree-based) imali bolje performanse od linearnih modela.
- Smanjenje dimenzionalnosti pomoglo je u smanjenju složenosti modela.

- Skaliranje i enkodiranje značajki osigurali su pravednu usporedbu između algoritama.