FAQ

Popis pitanja pročišćen s GPT-4 tako da uključuje samo najrelevantnija pitanja, dedupliciranа, grupirana po sličnim temama i filtrirana po relevantnosti. Prompt history: ChatGPT.

Dataset:

- Koji dataset koristiti i opis varijabli za bolju dokumentaciju?
  - Koristiti drugi dataset i ignorirati prvi. Drugi dataset je dataset jednog hotela u Rijeci. Dobili ste Podatke za 2 godine poslovanja, a mi smo zadržali test dataset od iduće godine. Test dataset ćemo poslati kada se predaju rješenja. Za to vrijeme validirate i testirate na dobivenom datasetu, a završni model koji se predaje istrenirajte na cijelom datasetu.
- Prvi dataset možemo ignorirati?
  - Da.
- Podatci iz 2007 koji su prisutni što s njima?
  - Svi podaci koji su u datasetu možete iskoristiti, ili ne morate.

Rezolucija predviđanja:

- Predviđanje na koju rezoluciju? Dnevno, tjedno, mjesečno, godišnje?
  - Glavno predviđanje je na dnevnoj bazi, ali probajte konstruirati rješenje za tjedno, mjesečno i godišnje predviđanje.

Event tablica:

- Tablica događaja: Za koji period raditi Event tablicu?
  - Za period u kojem se nalazi dataset. Također za period u kojem se nalazi test dataset. Mi smo zadržali za testiranje cijelu 2010. godinu i pola 2011. godine
- Kada budemo integrirali našu EVENTS tabelu sa događajima kako možemo modelirati/prikazati uticaj datih događaja na popunjenost hotela?
  - Dodati kao stupce u početni dataset, napraviti feature engineering i modelirati.
- Vezano za Event dataset, da li ćemo ga mi ubaciti u zadanu dataset kao nove kolumne?
  - Da. Paziti na broj kolumni, možda vam ih bude previše. Probajte ih što više smanjiti. W Curse of dimensionality
- Pošto smo dobili novi dataset, da li je potrebno i za njega raditi analize podataka kao što smo za prvi jer se on ipak razlikuje od dosadašnjeg na kojem smo analizirali i počeli trenirati naše modele?
  - Da. Modelirate samo drugi dataset.
- Da li Event tablica mora biti kakva je na prezentaciji ili možemo proizvoljno dodavati stupce?
  - Na prezentaciji je bio prijedlog. Možete dodati što god mislite/saznate da poboljšava predviđanje. Npr. vremenska prognoza, broj utakmica u mjesecu, itd. Mašti na volju.
- Prilikom Exploratory Data Analysis dijela zadatka, kako trebamo ovo istraživanje pretočiti u dokumentaciji?
  - Grafovi koji pokazuju distribuciju podataka, zanimljivosti dataseta i druge karakteristike. Svaki crtež ili izračun koji se napravi objasniti, napisati zaključke i otkuda su došli. Svaki feature engineering opisati i obrazložiti zašto se radi. Ako se neke ideje odbace napisati zašto su se odbacile. Zašto su se odabrale neke druge?
  - Kako se odražava na rezultate? Svo znanje i ideje koje su vam pale na pamet tijekom EDA zapisati. Ne zaboravite da s analizom morate znati ispričati priču. Koju priču pričaju vaši podaci?

Predviđanja i modeli:

- Da li možemo koristiti Neuralne mreže na predviđanje i ako predvidimo da li možemo konformno predviđanje naknadno ili moramo koristiti u sklopu predviđanja?
  - Neuralne mreže su loše za time series prediction. Lako se overfitaju, nisu explainable, spori su i dugo se treniraju. Izbjegavati u širokom luku. Bolje vam je napraviti 3 reda veličine eksperimenata više. Više ćete naučiti.
  - Koristite klasične ML algoritme. Neka vrsta će vam biti dovoljna. XGBoost, Catboost, Random forest.

- Konformno predviđanje je post hoc metoda za kvantifikaciju neodređenosti, tako da jedino i možete naknadno.

- Što je metrika predviđanja? Okupiranost, Canceled ili zarada? Predviđanje za cijelu godinu?
  - Broj zauzetih soba u danu.
  - Druge stvari vam možda mogu pomoći u predviđanju.
- Ima li smisla ciljati i na neke druge predikcije osim popunjenosti, npr. zarada ili otkazivanje?
  - Možda mogu pomoći, ali glavni target je broj zauzetih soba u danu.
- Prilikom objave dali sve u jednom notebooku šaljemo, ili možemo odvojeno što smo analizirali podatke, a u drugi treniranje i testiranje modela?
  - Šaljete github repozitorij s lijepo napisanim README, dokumentacijom u PDF-u, jupyter bilježnicama, skriptama i konfiguracijskim file-ovima za docker za podizanje aplikacije.

Varijable i čišćenje podataka:

- Kako handlati šum i greške u podacima? Izbaciti ili mijenjati? Da li to radimo matematički ili na gut feeling?
  - I jedno i drugo. Također i jedno i drugo.
  - Trebate pronaći outliere i izbaciti ih ako smetaju predviđanju. Također ćete možda morati neke podatke ispuniti ako nedostaju. Ili možda izbaciti jer ispunjavanje može narušiti distribuciju podataka pa time i predviđanje. Eksperimentirajte, izmjerite, dokažite što funkcionira bolje ili lošije.
- Nepravilnosti u podacima (Nemogući datumi itd.), kako odrediti outliere i što s njima?
  - Nepravilnosti nažalost uvijek postoje u podacima. Na vama je da pronađete adekvatno rješenje. Outliere možete odrediti pomoću statistike. Pazite kako birate thresholde, mjerite koji vam je najpovoljniji. Neke podatke možda treba izbaciti, neke ispraviti. Provjerite.
- Je li datum otkazivanja rezervacije prije datuma kreiranja rezervacije greška ili nismo dobro shvatili podatke?
  - Postoje takve greške u datasetu. Odlučite kako ćete to riješiti.

Druge teme:

- Način predaje: Docker, GitHub ili Jetbrains Space?
  - Molim vas predajte na github-u s lijepo napisanim README, dokumentacijom u PDF-u, jupyter bilježnicama, skriptama i konfiguracijskim file-ovima za docker za podizanje aplikacije.
  - Napravite lijepu i čitku strukturu foldera. Dodajte requirements da možemo pokrenuti rješenja s vašim verzijama biblioteka.
  - Napišite upute o pokretanju, korištenju repozitorija i testiranju modela.
  - Mora nam biti jasno što čitamo i što tražimo kroz prvih 5 minuta. Ako će nam trebati predugo da pokrenemo vaša rješenja to nije dobro.
- Bit će samo hotel iz Rijeke nećete dodati još neki?
  - Samo hotel iz Rijeke.