Advanced Analytics Engineer: TakeHome Assessment

Background

Dana Farber's pharmacy team must identify bottlenecks in their drug-preparation Turnaround Time (TAT) – the time between when a physician order's a medication to when a patient receives the medication – to improve throughput and patient care. Bottlenecks throughout the drug prep process can delay care and lead to poor patient satisfaction. As we currently understand it, the acceptable threshold for TAT is 60 minutes. While this is not real patient data, it does reflect similar healthcare data that you might encounter in this role.

Objective

You've been tasked with analyzing a large dataset to help us understand better:
(1) the time it takes between steps in the medication preparation process,
(2) possible drivers (e.g., reasons or correlations) for these delays, and
(3) how we can predict a drug's TAT (being above/below recommended threshold)

Also, having a strong background in MLOps, you must also provide recommendations for setting up and deploying a production-grade model that will allow us to deploy, monitor, and maintain your solution at scale (think cloud platform solutions/architecture).

Dataset Description

- **Rows:** 100,000 orders
- **Missingness:** ~10% of intermediate timestamps are missing at random
- **Data Span:** Jan 1 2025 – Jul 25 2025
- **Columns:**

  Timestamps (in sequential order, some may be missing)
  - doctor_order_time
  - nurse_validation_time
  - prep_complete_time
  - second_validation_time
  - floor_dispatch_time
  - patient_infusion_time

  Target: TAT_minutes (continuous)

  Patient and Clinical Context
  - **age** *(years)*
  - **sex** *(F/M)*
  - **race_ethnicity** *(White, Black, Hispanic/Latino, Asian, Other/Unknown)*
  - **insurance_type** *(Commercial, Medicare, Medicaid, Self-pay)*
  - **diagnosis_type** *(SolidTumor, Hematologic, Autoimmune, Other)*

- **severity** *(Low, Medium, High)*
- **treatment_type** *(Chemotherapy, Immunotherapy, TargetedTherapy, SupportiveCare)*
- **patient_readiness_score** *(1=not ready, 2=prep underway, 3=in chair/waiting)*
- **premed_required** *(0/1)* - whether pre-meds are required
- **stat_order** *(0/1)* - order flagged as STAT

Location, Staffing & Operations

- **floor** *(1, 2, 3)*
- **shift** *(Day, Evening, Night)*
- **floor_occupancy_pct** *(0–100)* — load on the treatment floor at order time
- **queue_length_at_order** *(0+)* — active queue count at order time
- **nurse_credential** *(RN, BSN, MSN, NP)*
  - **RN** — Registered Nurse (associate/bachelor's pathway; core bedside care).
  - **BSN** — RN with Bachelor of Science in Nursing (broader clinical/leadership prep than associate RN).
  - **MSN** — Master of Science in Nursing (advanced training; often leadership/education/clinical specialist).
  - **NP** — Nurse Practitioner (advanced practice clinician; can assess/manage/treat; often MSN/DNP prepared).
- **nurse_employment_years** *(0–40, numeric)*
- **pharmacist_credential** *(RPh, PharmD, BCOP)*
  - **RPh** — Registered Pharmacist (legacy bachelor's route; fully licensed; foundational clinical practice).
  - **PharmD** — Doctor of Pharmacy (current standard clinical pharmacist; direct patient care training).
  - **BCOP** — Board Certified Oncology Pharmacist (advanced oncology certification; expert in chemo/biologics).
- **pharmacist_employment_years** *(0–40, numeric)*
- **pharmacists_on_duty** *(count at order time)*
- **ordering_physician** *(synthetic names)*
- **ordering_department** *(MedicalOncology, Hematology, StemCellTransplant, RadiationOncology, ImmunotherapyClinic)*

Labs (with normal ranges)

- **lab_WBC_k_per_uL**: **4.0–11.0** ($\times 10^3/\mu L$)

- **lab_HGB_g_dL**: **12.0–16.0 g/dL** (adult reference used in data; real ranges vary by sex/age)
- **lab_Platelets_k_per_uL**: **150–400** ($\times 10^3/\mu L$)
- **lab_Creatinine_mg_dL**: **0.6–1.3 mg/dL**
- **lab_ALT_U_L**: **7–56 U/L**

Specific Tasks

1. **Data Exploration & Preprocessing**
   - Summarize basic statistics, distributions, and missingness patterns.
   - Engineer features to capture intervals and delays between steps.
   - Visualize step-to-step delays.
   - Handle missing timestamps in a sensible way.
2. **Modeling Approach**
   - Train and fit at least one model (e.g. XGBoost, Random Forest, or linear regression) to predict TAT that would exceed the acceptable threshold.
   - Evaluate performance using RMSE and MAE.
   - Identify top drivers of TAT.
3. **Key Findings & Drivers**
   - Describe any TAT bottlenecks and their potential impact.
   - Discuss pros/cons of your modeling approach.
   - Suggest potential interventions to optimize flow and reduce overall TAT.
4. **MLOps Scheduling & Orchestration:**
   - Share a diagram that shows how you might automate pipeline for data ingestion, data preprocessing, model prediction-making and model retraining
   - How often would you recommend a model like this run? Why?
   - Propose any key telemetry (data schema drift, model performance over time, inference latency) and the tools you would use to capture it (e.g., MLflow metrics, Databricks monitoring).
   - Briefly describe strategies to detect when the model has degraded (e.g. rolling window evaluation, PSI/KL divergence).
5. **MLOps Deployment Strategy**
   - Outline your plan for version control of this project, and how would you structure this project in Git?
   - How would you integrate automated testing into your CI?
   - More specifically, what unit tests, assertions or schema validation would you create to ensure proper CI/CD of the model from development to production?
   - How would you automate data retraining and deployment?

As a suggestion, you perform Python analysis in your IDE of choice, but present in your findings in a slide deck format (~10 slides) that covers these 8 topics:

1. **Personal Introduction:** About you and your experience

2. **Problem Statement & Impact**
3. **Data Exploration & Preprocessing**
4. **Modeling Approach**
5. **Key Findings & Drivers**
6. **MLOps Scheduling & Orchestration**
7. **MLOps Deployment Strategy**
8. **Next Steps & Q&A**

What we're looking for:

- Clear framing of problem and description of findings.
- Thoughtful MLOps Design that shows you've considered reliability, reproducibility, and scalability—especially around version control, CI/CD, scheduling, and monitoring.
- Communication: clean, succinct slides that would land with both technical and operational stakeholders.

Submission

Please email your **Jupyter notebook** (or script), the **CSV of any processed data**, and the **PowerPoint deck** to gabe_verzino@dfci.harvard.edu.

Good luck! We look forward to your insights on reducing pharmacy turnaround times.