TruBridge Patient No-Show Analysis (FINAL)

Project: patient no-show analysis (EDA + stats + baseline model + dashboard)

Stack: Python (Pandas, NumPy), Seaborn/Matplotlib, SciPy, scikit-learn, React dashboard

Links

- Dashboard: https://patient-noshow-dashboard.vercel.app/
- GitHub: https://github.com/ivan-711/Externship-TruBridge

What I built

I cleaned the raw appointment dataset, engineered features (WaitingDays), ran statistical tests, trained a baseline logistic regression model, and published results in a React dashboard to highlight patterns that can help reduce missed appointments.

Deliverables: final Colab notebook + cleaned_appointments.csv + publicsafefinalnoshowdataset.csv + React dashboard

Data

- Source file: Kaggle "Medical Appointment No Shows" (May 2016)
- Target: NoShow (0 = show, 1 = no-show)
- Key fields used: Age, SMS_received, ScheduledDay, AppointmentDay, derived WaitingDays
- Engineered WaitingDays from timestamps and filtered unrealistic values to keep the analysis consistent (kept 0–365 days; removed negative values).
- Rows: 110,527 → 71,959 after cleaning.
- **Public-safe Dataset**: Removed PatientID and AppointmentID for the dashboard.

Methods

Feature engineering

- Converted NoShow to 0/1 and computed WaitingDays = AppointmentDay − ScheduledDay (days).
- Normalized Age and WaitingDays for modeling.

Stats + EDA

- Descriptive rates: no-show rate ≈ 20.19%, show-up rate ≈ 79.81%, SMS_received mean ≈ 0.321.(~32% of appointments received an SMS)
- Chi-square test for SMS_received vs NoShow using a contingency table.
- Welch t-test for WaitingDays (Show vs NoShow).
- Visuals: WaitingDays distribution (full + capped), WaitingDays by NoShow, SMS/no-show counts, weekly trend, Age visuals, Age vs WaitingDays scatter.

Key results (what matters)

- No-shows are common (~20%) in this dataset.
- WaitingDays is higher for no-shows (mean show ≈ 8.75 days vs mean no-show ≈ 15.84 days; p-value < 0.001).
- SMS_received is associated with different no-show rates (chi-square statistic ≈ 1767.98, p-value < 0.001). This is association, not cause.

Modeling (baseline logistic regression)

- Trained a logistic regression using Age_normalized, WaitingDays_normalized, and SMS_received to predict NoShow.
- Baseline accuracy (predict all "show"): ~0.7981
- Model accuracy: ~0.7960 (about the same as baseline)
- The model has very low recall for the no-show class, which shows class imbalance + limited feature power for true prediction (good for learning, not enough for deployment).
- **Interpretation:** The dataset is imbalanced and these few features don't capture enough signal to predict no shows reliably.

What I would do next

- Add stronger features (appointment type, lead-time buckets, prior no-show history, day-of-week/time-of-day).
- Handle class imbalance (class weights, threshold tuning, precision/recall focus).
- Validate with cross-validation + calibration, then target high-risk appointments for outreach.