🧠 Candidate Dropout Prediction — Final Report

📌 Executive Summary

In recruitment, final interview “no-shows” create delays and waste recruiter effort. This project predicts candidate dropouts using profile and behavioral features such as test scores, experience, and response times. A logistic regression model was trained, achieving 100% recall, ensuring all likely dropouts are flagged. The results are deployed in a user-friendly Streamlit dashboard.

📌 Business Context

- **Problem**: Final interview dropouts delay hiring timelines and increase recruiter workload.
- **Impact**: Wasted shortlisting effort, poor pipeline reliability, slower offer rollouts.
- **Goal**: Predict dropout risk early to enable proactive follow-up and prioritization.

📌 Data Science Solution

- **Problem Type**: Binary classification (`dropout = 1`)
- **Model Used**: Logistic Regression (interpretable + baseline)
- **Data**: Synthetic, but realistic candidate data with features like:
  - Experience
  - Test Score
  - Email Response Delay
  - Form Completion Time
  - Referral Status
  - Timezone Difference
  - CTC Expectation

📌 Exploratory Analysis

- Candidates referred internally are **79% less likely** to drop out.
- **Low test scores, long form times**, and **delayed email replies** are strong dropout signals.
- Higher experience sometimes correlates with higher dropout — possibly due to more options.

⚙️ Modeling Approach

- **Preprocessing**:
  - Skew detection & capping outliers
  - Feature scaling (StandardScaler)
  - Multicollinearity removal (dropped `resume_score`)
- **Assumption Checks**:
  - VIF < 5 for all features post-cleaning
  - Log-odds linearity confirmed
  - No perfect separation detected
- **Threshold Tuning**:
  - Default (0.5) missed all dropouts
  - Lowered to `0.05` to prioritize recall

📌 Model Performance

Metric	Value (Threshold = 0.05)
Accuracy	49% (not meaningful here)
Recall (1)	**100%**
Precision (1)	14%
AUC	0.70

Metric	Value (Threshold = 0.05)

- High recall means zero dropouts missed, which is valuable to HR teams.

Deployment & App Features

- Built with **Streamlit**
- Features:
  - Real-world slider inputs for single prediction
  - CSV batch upload for multiple candidates
  - Dashboard with odds ratios and ROC curve
- Ready for deployment on Streamlit Cloud or HuggingFace Spaces

---

Recommendations

1. Prioritize follow-ups for candidates with:
   - High email delays
   - Long form times
   - Low test scores
2. Shorten or streamline the application process
3. Invest in referral programs to reduce no-shows

Limitations & Next Steps

- Synthetic data may differ from real-world behavior
- Low precision: next step could be ensemble model or XGBoost
- Add SHAP or LIME for transparent candidate-level explanations

---

Project Assets

- `logistic_regression_dropout_model.pkl`
- `final_cleaned_candidate_data.csv`
- `streamlit_app_advanced.py`
- `example_batch.csv`

Made by a Data Scientist who understands both business and models.