Project Overview: Clinical Trial Participant Dropout Prediction

End-to-end machine learning + deployment (Streamlit) | DTSC 691

Student: Gbolahan (Abdul) Oladosu **Date:** December 5, 2025

Project in one sentence: Predict dropout risk after Visit 2 using early-visit demographic, clinical, and engagement signals, and deploy the model in a user-friendly web app for clinical operations.

Contents

1. Objective and scope

A major issue in clinical trials is the frequent loss of participants before reaching the primary endpoint in the trial.When participants drop out, it increases cost, slows timelines, and can bias results if the people who leave are systematically different from those who stay. The goal of this project is to build an end-to-end machine learning system that will estimate a participant's probability of dropping out with the use of information collected up to participant's second visit (Visit2) . The output is designed as a decision support for clinical research associates (CRAs) and study managers; it will be an effective way of prioritizing outreach and retention efforts, not a substitute for clinical judgment.

Scope: (1) Exploratory Data Analysis (EDA) to understand patterns related to dropout; (2) build a reproducible preprocessing and modeling pipeline; (3) Evaluate and tune a classification threshold for operational use; (4) Explain predictions using SHAP; (5) Deploy the model in a Streamlit web app inside a personal website with required pages.

Out of scope: Real patient data, Causal claims, or predicting outcomes beyond the Visit 2 horizon.

2. Data acquisition and dataset description

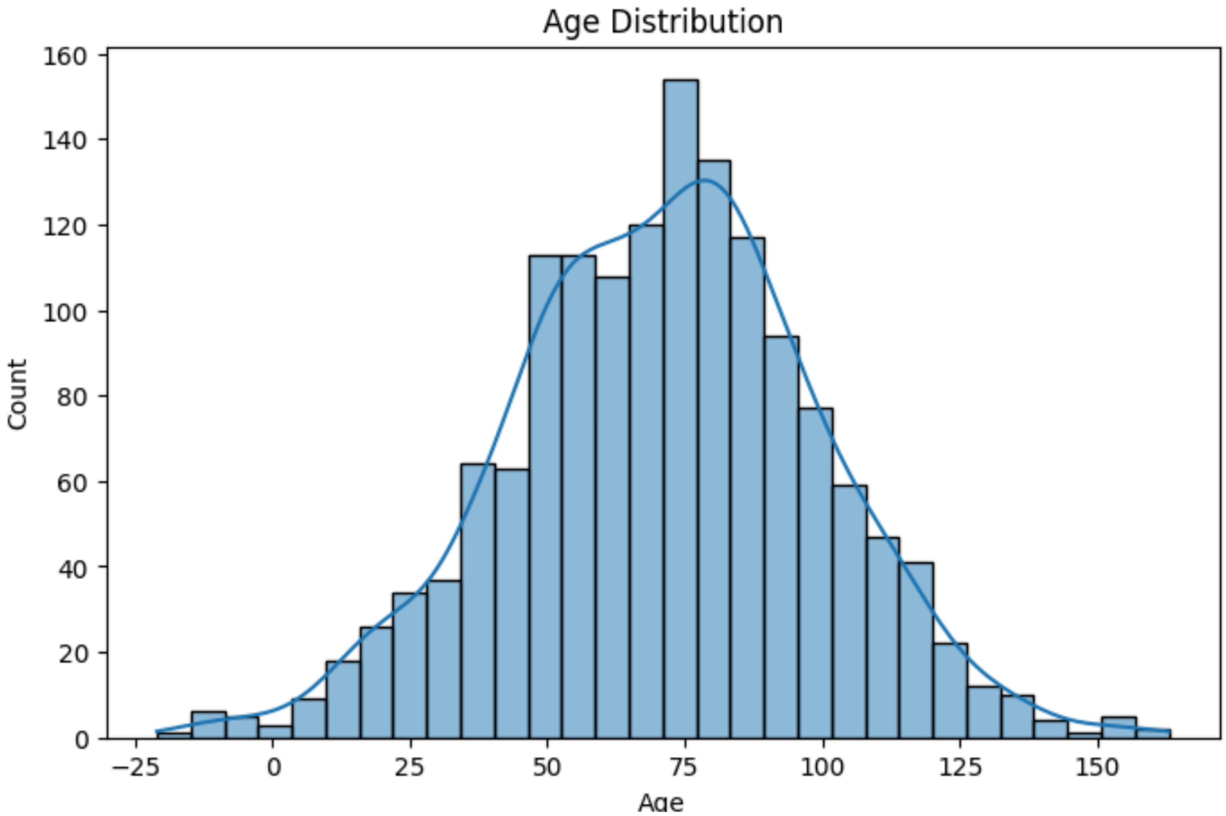
Because real clinical trial datasets are often restricted, this project uses a synthetic dataset designed to mimic real-world trial structure and data issues. The goal initially was to use datasets from Kaggle; upon finding out those dataset were also synthetic datasets I decided on creating my own while making sure it captures real-world data issues. The dataset includes 1,500 participants with demographics (age, sex, race), baseline clinical variables (BMI, baseline lab score, disease severity, prior treatments), Visit 1 and Visit 2 measures (symptoms, adherence, adverse events), engagement signals (missed appointments, communication score), and a binary target (**dropout**: 1=dropout, 0=complete).

Synthetic data still creates realistic modeling challenges: imperfect ranges, noisy variables, and missing visit information. Instead of deleting large chunks of data, the preprocessing pipeline applies clinically motivated constraints (for example, adult-only ages) and uses missingness indicators to preserve information.

3. Exploratory data analysis

EDA was used on the original uncleaned dataset to (1) verify variable ranges and distributions, (2) check class balance, and (3) identify relationships between features and dropout. Figures below are taken directly from the notebook outputs and interpreted in the context of clinical trial operations.

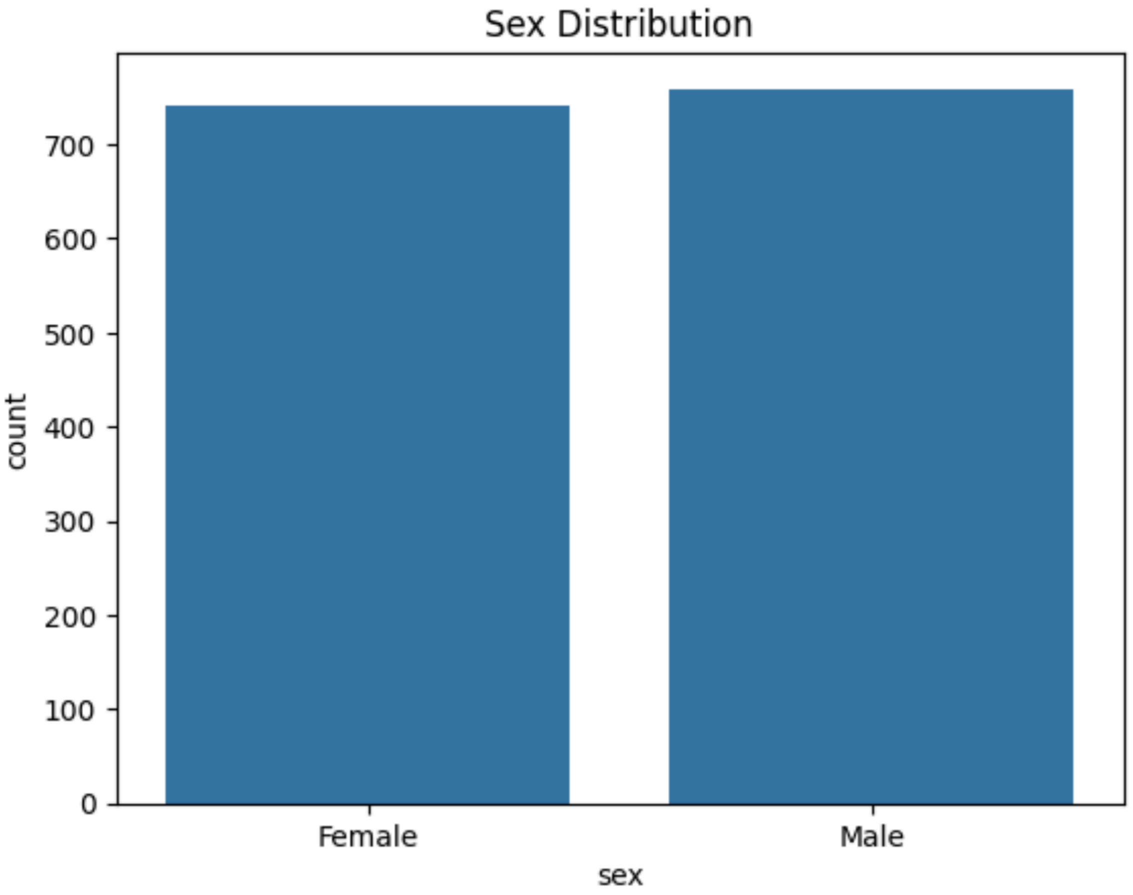
Figure 1. Age distribution



The age distribution is centered in older adulthood, which indicates that this clinical trial is based on disease or conditions affecting majorly older people. This is also important to keep in mind that, it is logically expected that older participants may have more missed visits, higher dropout rates and so on. Also, as shown in the graph it includes implausible values (below 18 and above 90). Upon analysis the total count of below 18 and above 90 years of age is 408.

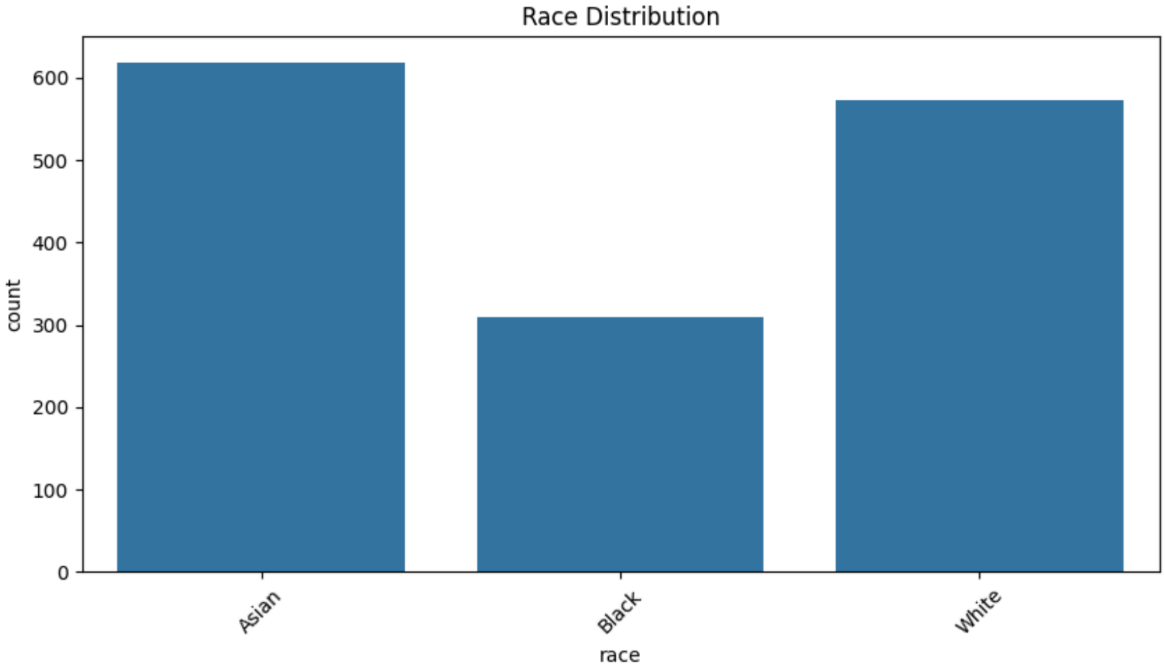
Rather than removing records, real- world solution was applied by assuming participant in the clinical trial will require a adult with no need to parental consent (>18) and a realistic older age.The preprocessing pipeline caps ages to an adult trial range (18-90) so the model sees clinically plausible inputs while preserving sample size.

Figure 2. Sex distribution



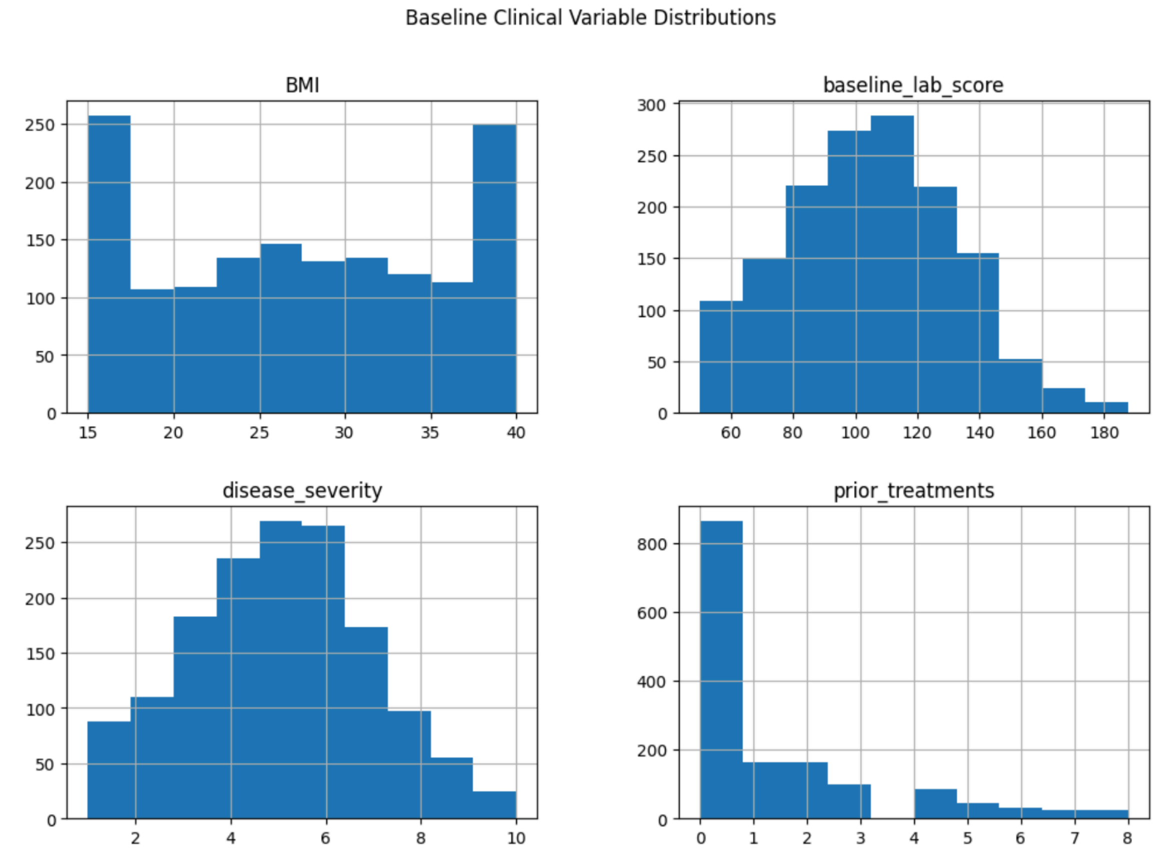
Sex is reasonably balanced, which reduces the risk that model performance is driven by a heavily skewed demographic subgroup and will be generally good for prediction outputs, avoiding bias.

Figure 3. Race distribution



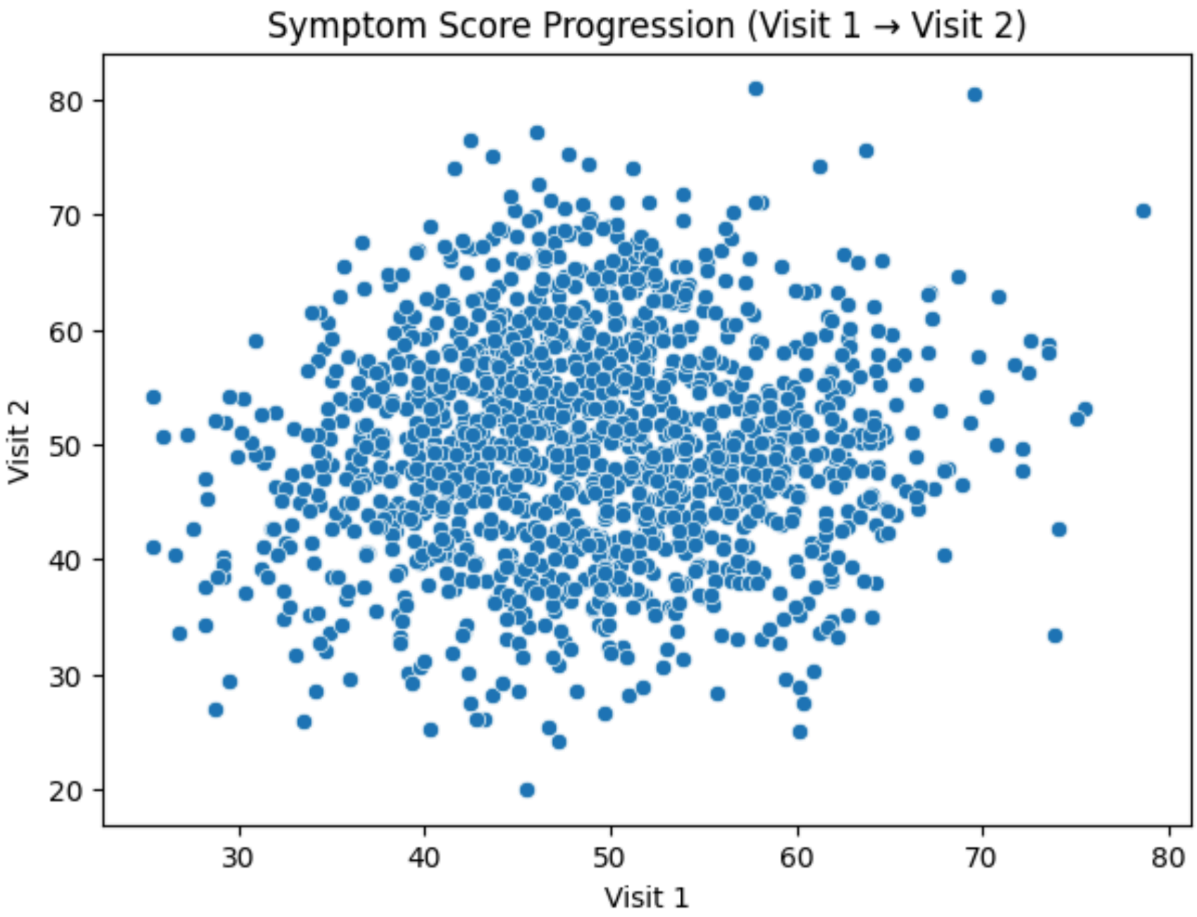
Multiple race categories are represented. This supports subgroup checking during evaluation and encourages fairness-aware interpretation when discussing drivers of dropout.

Figure 4. Baseline clinical variable distributions



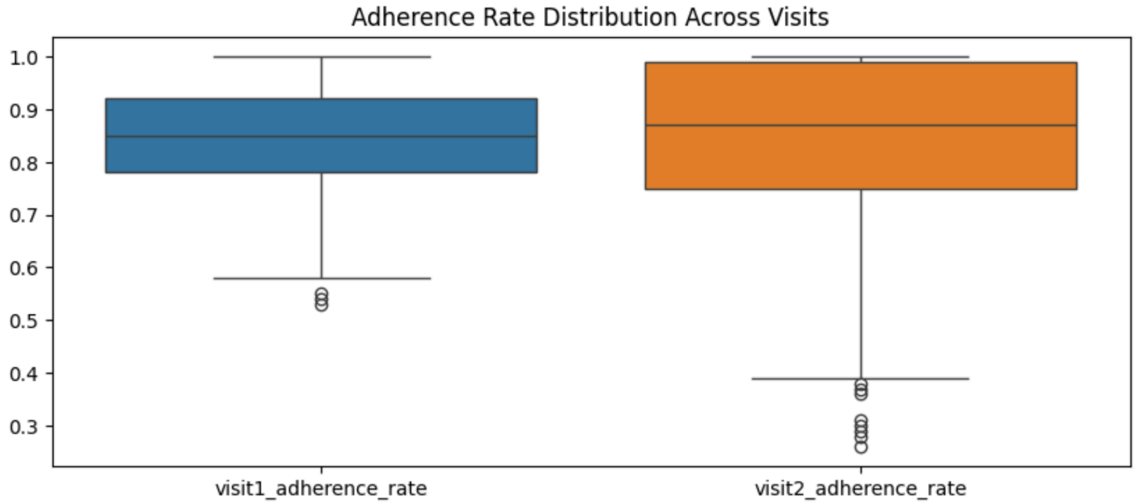
BMI, baseline lab score, disease severity, and prior treatments fall into plausible ranges for a chronic-disease style trial. These baseline variables capture overall health burden that can influence adherence, adverse events, and retention.

Figure 5. Symptom scores from Visit 1 to Visit 2



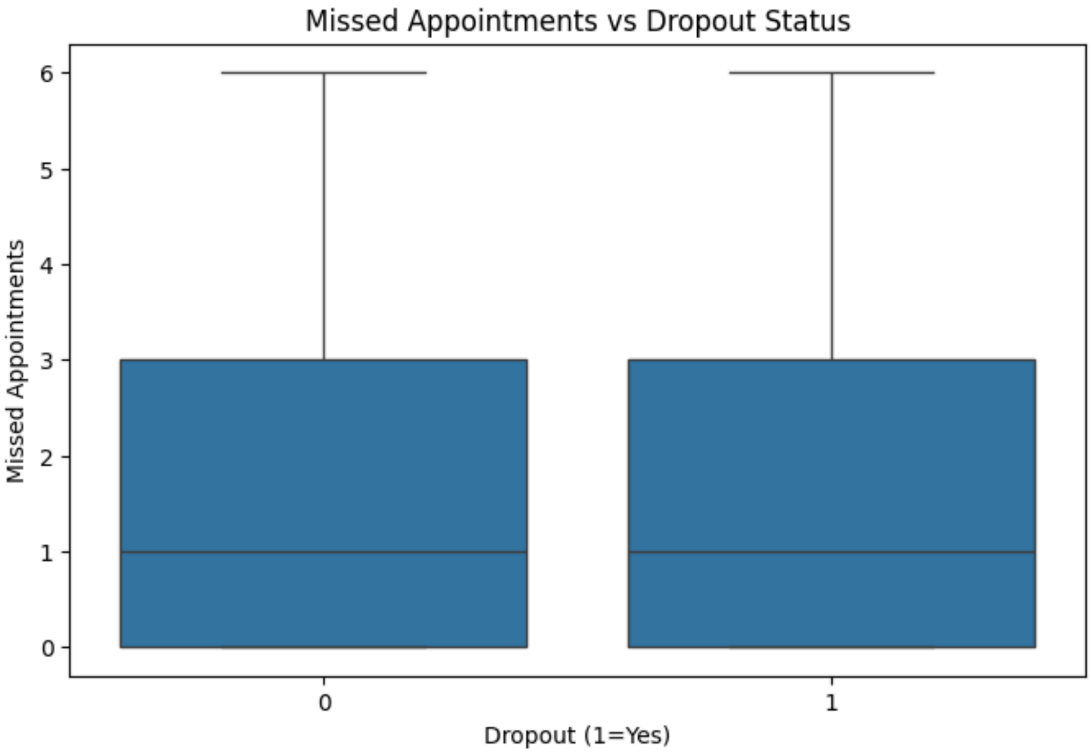
- There is a **moderate positive relationship** between Visit 1 and Visit 2 symptom scores.
- Participants with higher symptom severity at Visit 1 generally tend to have higher symptom severity at Visit 2.
- This indicates that **early symptom burden tends to persist** over the short term rather than change dramatically between visits.
- The points are **widely dispersed around the upward trend**, meaning there is substantial individual variability.
- This suggests that while symptoms are related across visits, **Visit 1 symptoms alone do not fully determine Visit 2 outcomes**.

Figure 6. Adherence rate distributions across visits



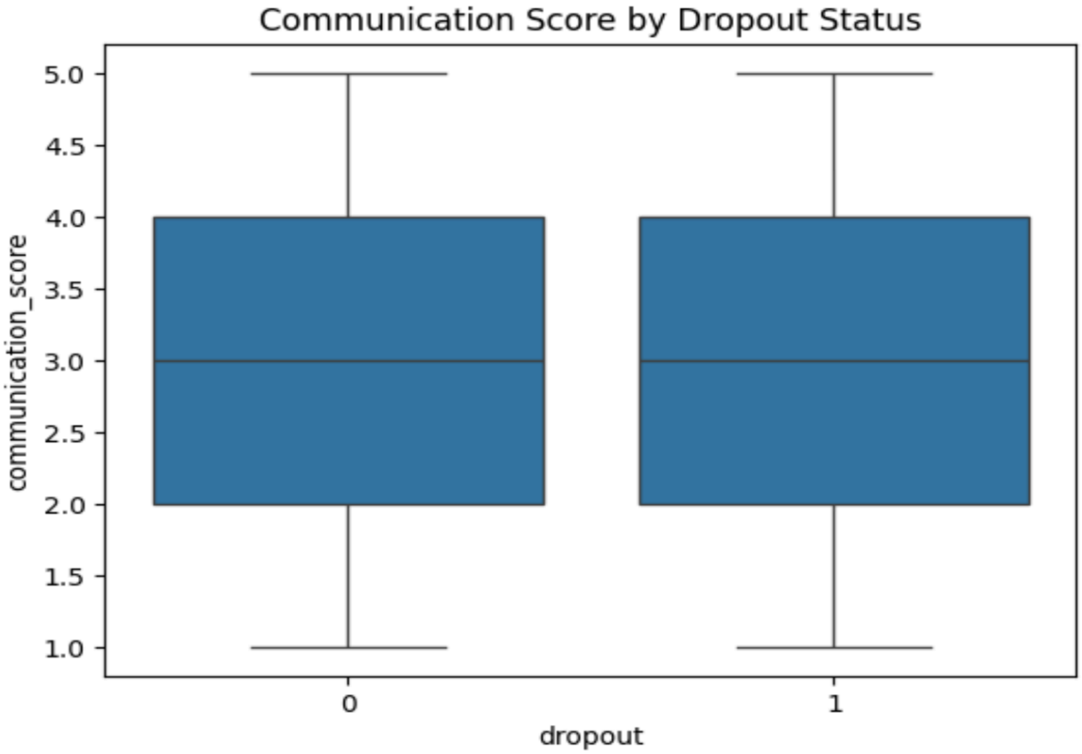
Adherence is generally high, but the spread at Visit 2 suggests some participants decline over time. In trials, declining adherence is often an early warning sign of disengagement, making adherence a strong candidate predictor.

Figure 7. Missed appointments by dropout status



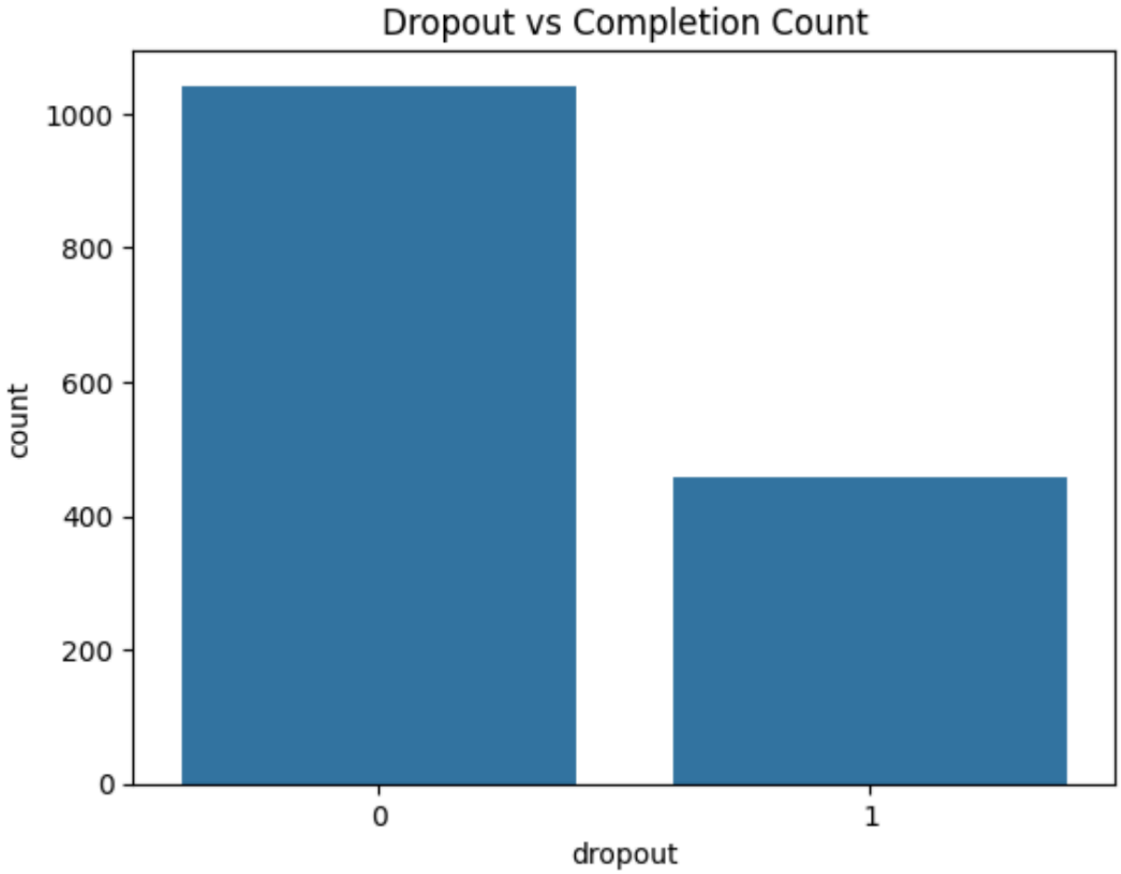
Based on this plot, missed appointments by itself do not show a strong visual separation between completers (0) and dropouts (1). Any difference, if it exists, is subtle and would need statistical testing or additional features.

Figure 8. Communication score by dropout status



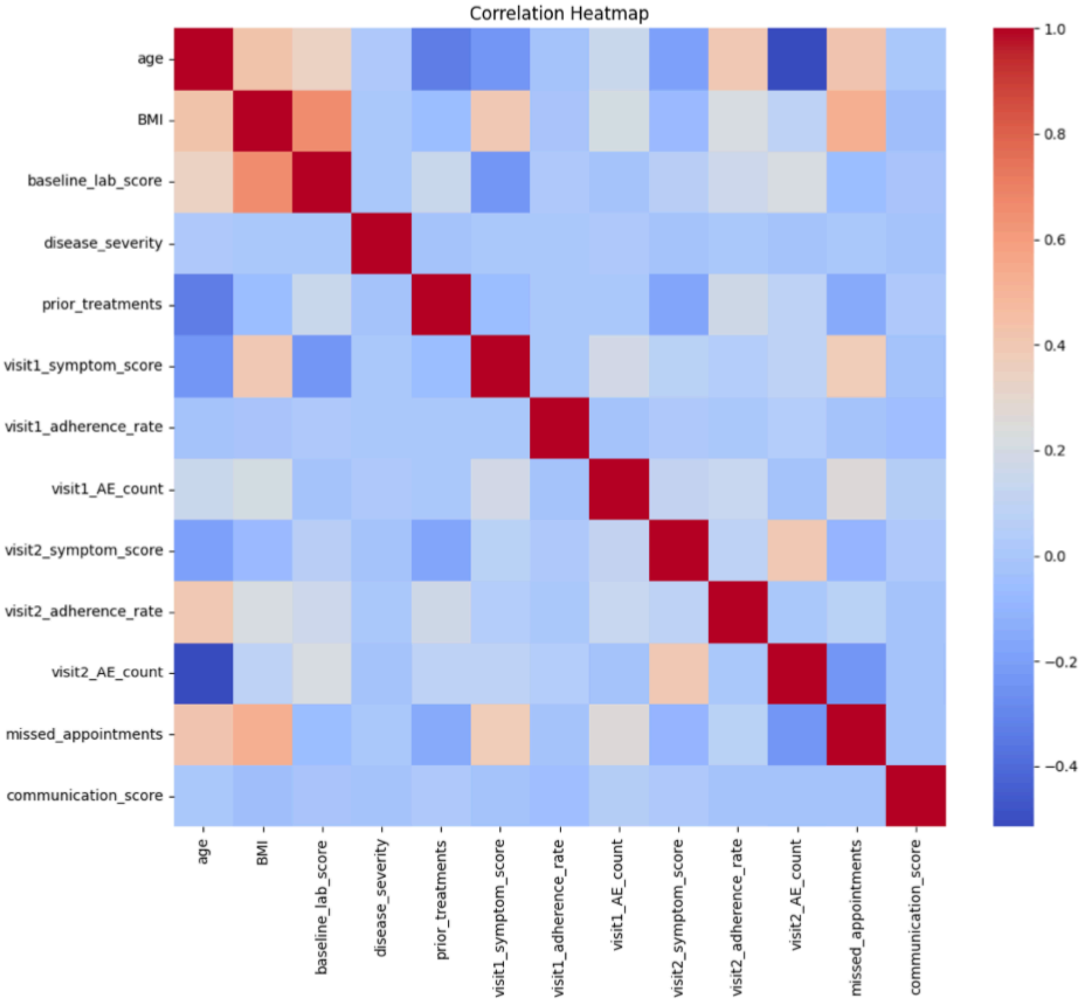
There is no visible difference in communication scores between participants who dropped out and those who completed the trial.
The two distributions completely overlap, indicating that communication score alone does not distinguish dropout behavior in this dataset.

Figure 9. Dropout vs completion class counts



The target variable shows that dropout is the minority class. Class imbalance motivates careful metric selection (ROC-AUC/PR-AUC, precision, recall) and the use of imbalance-handling techniques during training.

Figure 10. Correlation heatmap

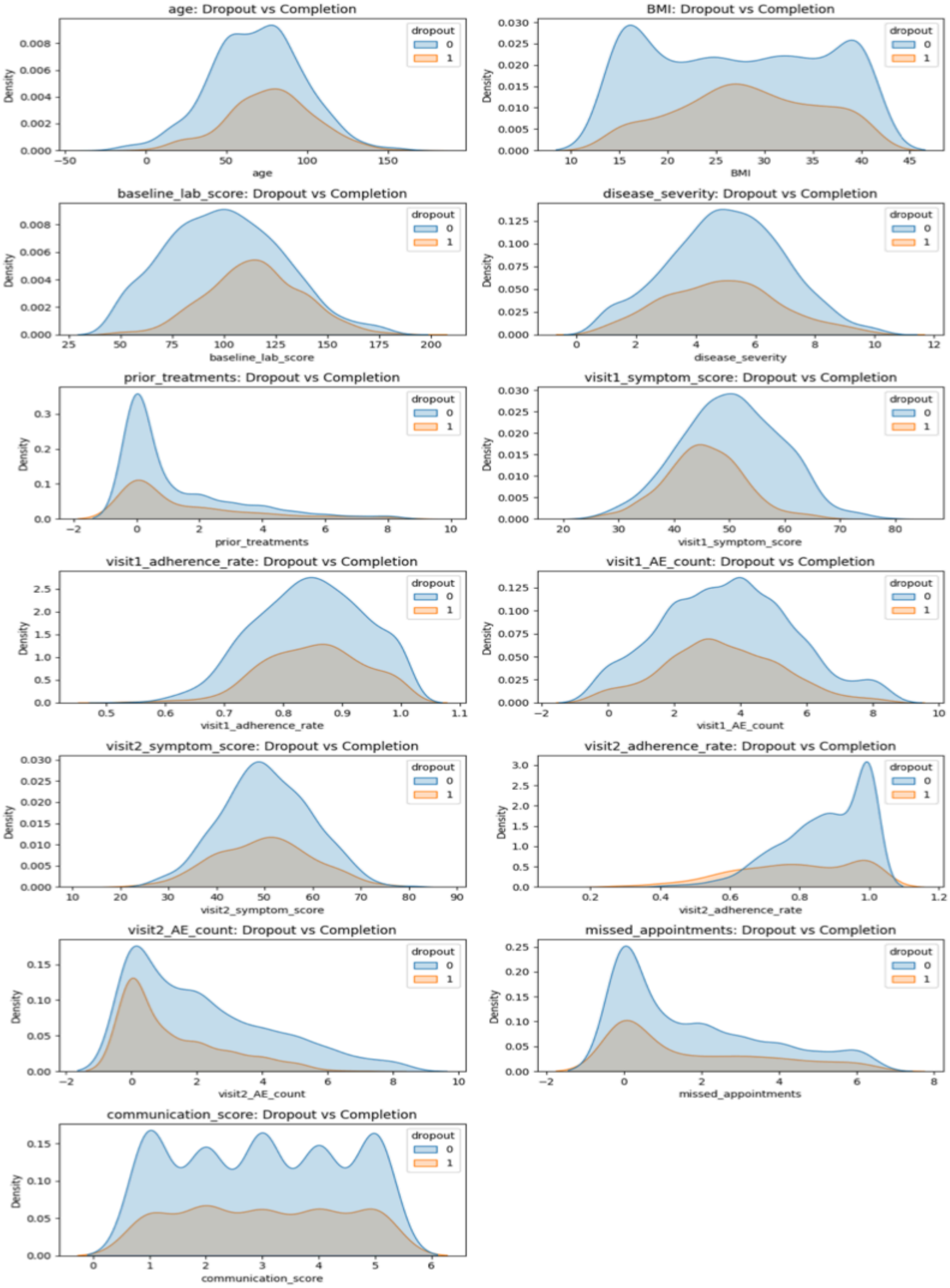


Overall: Most correlations are small, meaning many variables provide **distinct information** (low multicollinearity overall).

Strongest positive relationship: BMI and baseline_lab_score (r = 0.654) — higher BMI tends to align with higher baseline lab scores.

Strongest negative relationship: age vs visit2_AE_count (−0.515) — older participants tend to have fewer Visit 2 adverse events.

Figure 11. Feature distributions by outcome (dropout vs completion)



These density plots compare dropouts (orange) and completers (blue) across key variables. Visible shifts - especially for adherence, missed appointments, and some clinical measures - suggest those features carry signals. The plots also reveal overlap between groups, which is expected in real trial settings and reinforces why probability-based risk scoring is more appropriate than hard rules.

No single variable cleanly distinguishes participants who drop out from those who complete the trial. Instead, dropout appears to be associated with **patterns of engagement and behavior**—particularly adherence rather than baseline clinical severity. This supports the use of multivariate and nonlinear machine learning models to capture combined effects rather than relying on individual predictors.

4. Data preparation and cleaning

Preprocessing was implemented as a pipeline so that the exact same cleaning steps are applied during both training and deployment. Key steps included:

- **Clinical consistency rules:** adult-only age constraints (cap below 18 to 18; cap above 90 to 90) and a visit-timeline rule (if all Visit 1 fields are missing, Visit 2 fields are cleared).
- **Missingness as signal:** missing-indicator flags are added for selected variables, because in clinical operations missing data can reflect disengagement, missed visits, or data capture failures.

- **Imputation and encoding:** numeric variables are imputed and scaled; categorical variables (sex, race) are imputed and one-hot encoded.

5. Model training approach

This is a supervised binary classification task. Features include demographics, baseline clinical measures, visit-level outcomes, and engagement signals; the target is dropout (1) vs completion (0). Three models were evaluated using **cross-validated ROC-AUC** to compare performance and robustness across all three models:

LogReg CV ROC-AUC: **0.8056** ± 0.0253
RandomForest CV ROC-AUC: **0.9353** ± 0.0123
XGBoost CV ROC-AUC: 0.9412 ± 0.0125

What does this mean?

The higher the number the better.

It tells you how well each model separates dropouts from completers across all possible probability thresholds.

A value of: 0.50 = no better than random guessing

0.70–0.80 = acceptable

0.80–0.90 = strong

Greater than 0.90 = excellent (rare in messy clinical data)

The ± value is the standard deviation across CV folds, which reflects stability and generalizability.

As you can see XGBoost achieved the strongest cross-validated ROC-AUC, so it was selected as the final model.

Hyperparameters were then tuned using randomized search with 5-fold cross-validation for the XGBoost model keeping the one that performs the best under cross-validation.

```
Best XGB params: {'model__subsample': 1.0, 'model__n_estimators': 200,
'model__min_child_weight': 1, 'model__max_depth': 4,
'model__learning_rate': 0.1, 'model__gamma': 0.1,
'model__colsample_bytree': 1.0}
```

```
Best CV ROC-AUC: 0.940967557487052
```

6. Model evaluation and results on test set

On the held-out test set (300 participants), the final XGBoost model achieved:

- **Accuracy:** 0.92
- **Precision:** 0.8695652173913043 - **Recall:** 0.8695652173913043
- **F1:** 0.8695652173913043
- **ROC-AUC:** 0.9611204013377926

What does this mean?

The final XGBoost model was evaluated on a held-out **test set** of **300 participants**, consisting of **208 non-dropouts (69.3%)** and **92 dropouts (30.7%)**.

The model correctly classified **276 out of 300 participants**, achieving an overall **accuracy of 92%**.

From the confusion matrix, the model:

- Correctly identified **80 true dropouts**
- Correctly identified **196 non-dropouts**
- Produced **12 false dropouts** (unnecessary interventions)
- Missed **12 true dropouts**, representing the most clinically important errors

Performance on the dropout class was strong and well-balanced:

- **Precision = 0.87** →When the model says **“this person will drop out”**, it’s right about **87%** of the time.
- **Recall = 0.87** → 87% of all true dropouts were correctly identified

- **F1-score = 0.87** → Indicates a good balance between catching dropouts and limiting false alarms

The model achieved an excellent **ROC-AUC of 0.96**, demonstrating strong ability to distinguish between dropouts and non-dropouts across different decision thresholds.

Class-wise results show:

- **Non-dropouts (Class 0):** Precision, recall, and F1 ≈ 0.94
- **Dropouts (Class 1):** Precision, recall, and F1 ≈ 0.87

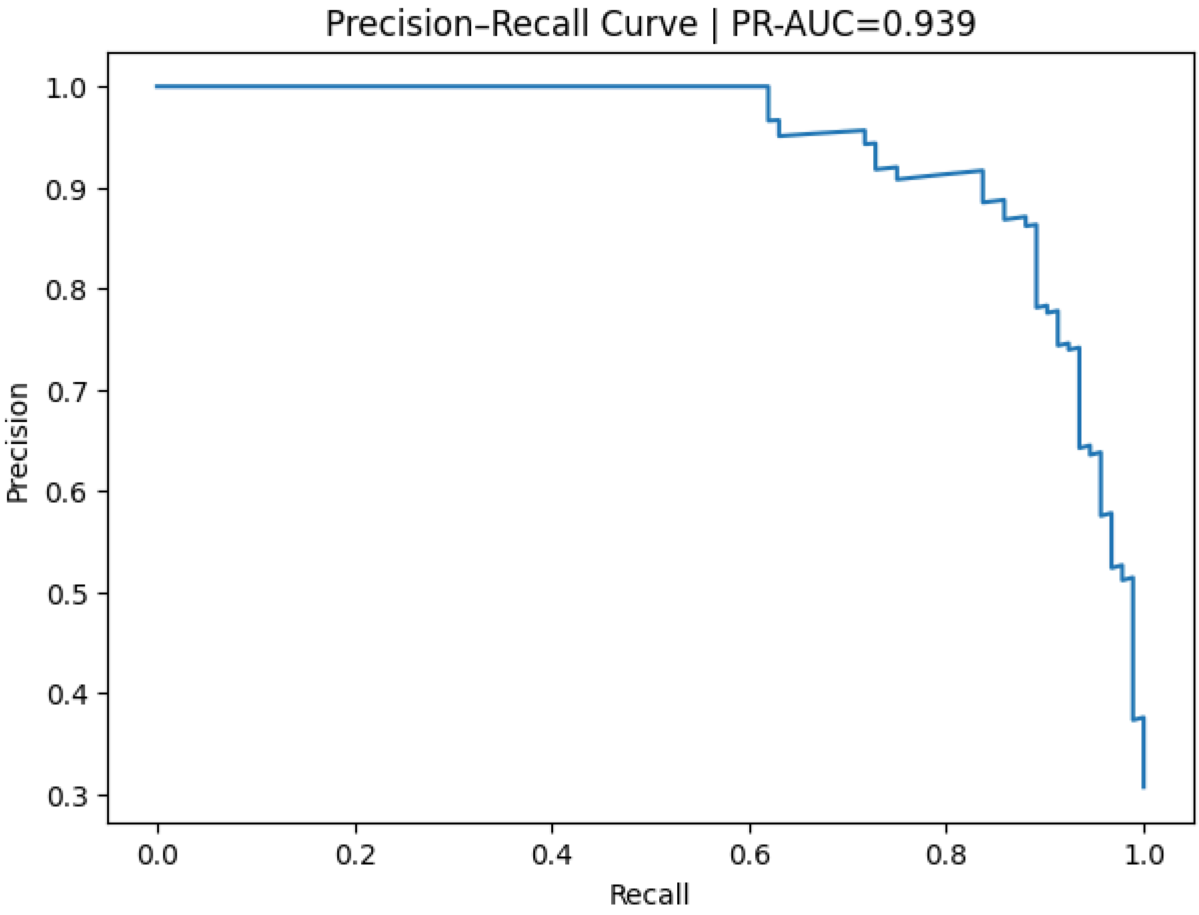
Overall, these results indicate that the model performs reliably on unseen data and is effective at identifying participants at high risk of dropping out, making it suitable as a clinical decision-support tool for early intervention.

PR-AUC (Average Precision)

This measures how well the model ranks dropouts above non-dropouts, using the Precision–Recall framework.
PR-AUC = 0.9390

What does this mean?
A PR-AUC of **0.939** shows that the model is **highly effective at identifying and prioritizing participants at risk of dropout,** even in an imbalanced clinical dataset.

Figure 12. Precision-Recall curve (XGBoost)



PR-AUC = 0.939 indicates strong performance on the positive class (dropout), which is especially useful under class imbalance. The visualizes that across *all* possible thresholds, this model is very good at **ranking true dropouts higher than non-dropouts**, and it keeps **precision reasonably high even as you push recall up** (which matters a lot when dropouts are the minority class).

Threshold tunning

I tuned the probability cutoff (threshold) for the XGBoost dropout model, then locked in the best cutoff for the Streamlit app, and finally evaluated and visualized the results:

I tested thresholds from **0.05 to 0.95** (step = 0.05).
At each threshold, it turned probabilities into predictions.
For each threshold, I calculated **precision, recall, and F1**, and stored them in a table.

Then i applied a rule:
Goal: choose a threshold with **recall ≥ 0.90**
If multiple meet that goal, pick the one with the **best precision** (and then best F1)
If none meet it, choose the threshold with the **best F1 overall**

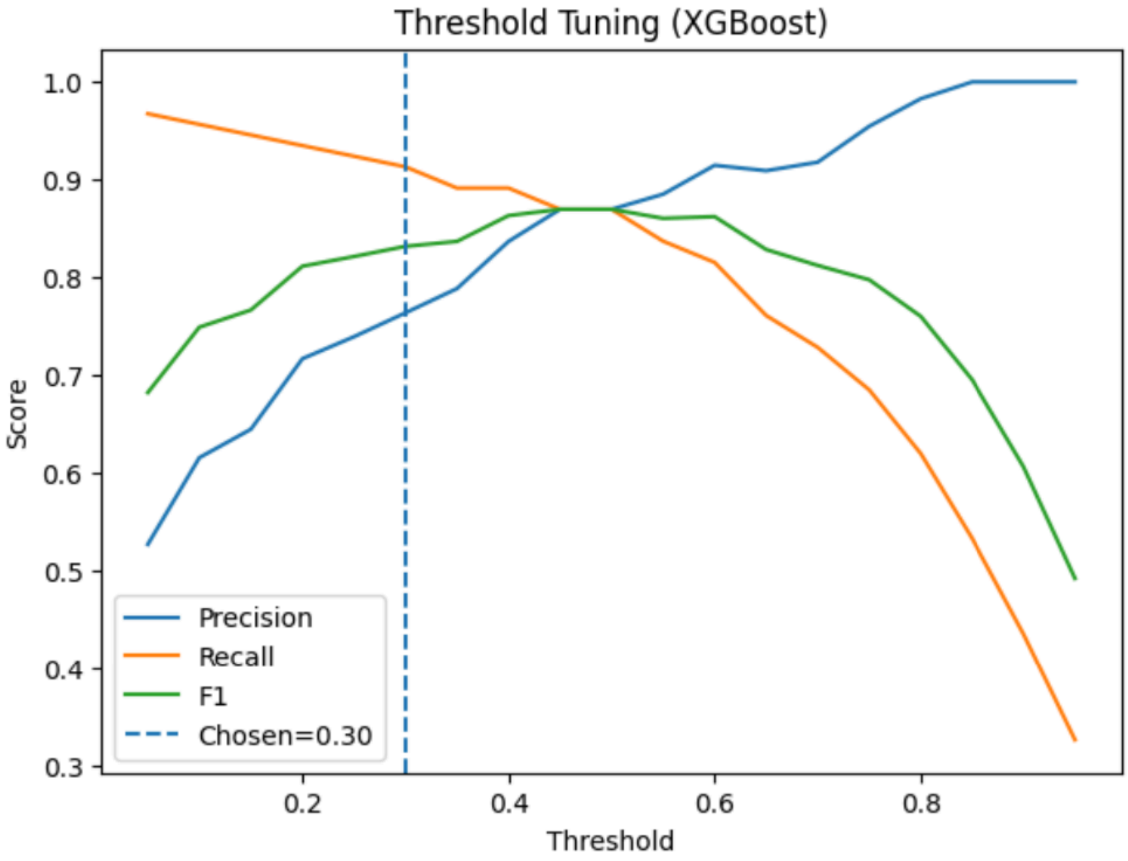
	threshold	precision	recall	f1
9	0.50	0.869565	0.869565	0.869565
8	0.45	0.869565	0.869565	0.869565
7	0.40	0.836735	0.891304	0.863158
11	0.60	0.914634	0.815217	0.862069
10	0.55	0.885057	0.836957	0.860335
6	0.35	0.788462	0.891304	0.836735
5	0.30	0.763636	0.913043	0.831683
12	0.65	0.909091	0.760870	0.828402
4	0.25	0.739130	0.923913	0.821256
13	0.70	0.917808	0.728261	0.812121

```
Chosen threshold when recall ≥ 0.9:
threshold    0.300000
precision    0.763636
recall       0.913043
f1           0.831683
```

What does it mean?
I selected a recall-first threshold of 0.30 to catch ≥90% of dropouts (recall=0.913). At this threshold, the model flags 84/92 dropouts while missing only 8, with 26 false alarms. This is

very much appropriate for a clinical setting where missed dropouts are more costly than extra outreach.

Figure 13. Threshold tuning (XGBoost)

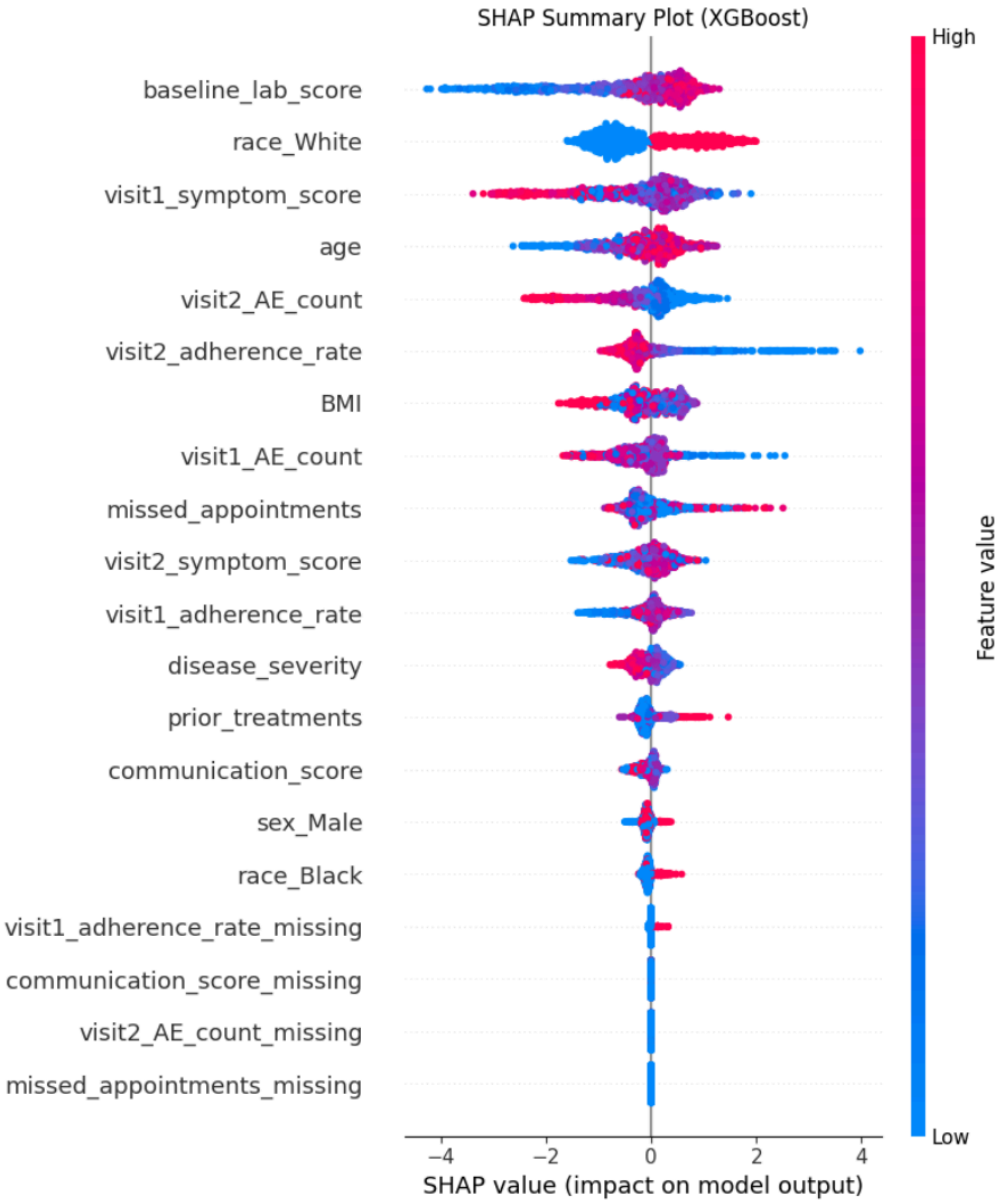


The threshold sweep shows how precision, recall, and F1 change as the cutoff moves. Lower thresholds increase recall (catch more potential dropouts) but reduce precision; higher thresholds do the opposite. The notebook highlights a recall-focused threshold around 0.30.

7. Model interpretability (SHAP)

To make predictions transparent, SHAP values were used to explain feature contributions. SHAP does not just rank features; it also shows directionality (whether high values increase or decrease predicted dropout risk).

Figure 14. SHAP summary plot (XGBoost)



How to read this plot

- Rows (top → bottom) = features are ranked based on overall importance (top features will influence a predictions the most)
- Dots = individual participants
- X-axis (SHAP value):

- Right (+) → increases predicted dropout risk
- Left (–) → decreases predicted dropout risk

- Color

- Red = high value of the feature
- Blue = low value of the feature

- Key drivers of dropout risk

Baseline lab score (most important)

- High baseline lab score (red → right) → higher dropout risk
- Low baseline lab score (blue → left) → lower dropout risk

Interpretation: Participants entering the trial with more abnormal or severe baseline labs are more likely to drop out.

8. User interface integration

The final trained pipeline was saved and deployed in a Streamlit web application embedded within a personal website. The website includes: (1) a biographical homepage, (2) a resume page, (3) a general projects page, and (4) a dedicated Dropout Project page where users can enter participant values and receive a predicted dropout probability. The app also provides risk buckets (low/moderate/high) and SHAP-based explanations to support non-technical decision making.

9. Challenges and lessons learned

Key challenges included managing synthetic data artifacts (for example, unrealistic ages), handling missing visit information without discarding data, and keeping preprocessing consistent between training and deployment. Building the Streamlit interface reinforced that a model is only useful if it can be used safely and clearly by real users.

10. Conclusion

This project demonstrates a complete applied data science workflow: problem framing, EDA, preprocessing, model training, evaluation, interpretability, and deployment. The resulting system provides an interpretable risk score that can help clinical operations teams prioritize retention interventions after Visit 2.

11. References

Machine Learning Algorithm: XGBoost

Software / Documentation
XGBoost Developers. (2024). ***XGBoost Documentation*****.**
https://xgboost.readthedocs.io/

Peer-Reviewed Foundation
Chen, T., & Guestrin, C. (2016). ***XGBoost: A scalable tree boosting system*****. Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 785–794.**
https://doi.org/10.1145/2939672.2939785

Model Interpretability (SHAP)

Peer-Reviewed Foundation
Lundberg, S. M., & Lee, S. I. (2017). ***A unified approach to interpreting model predictions*****. Advances in Neural Information Processing Systems (NeurIPS).**

Data Manipulation & Analysis

Software / Documentation

pandas development team. (2024). *pandas Documentation*.
https://pandas.pydata.org/docs/

NumPy Developers. (2024). *NumPy Documentation*.
https://numpy.org/doc/

Peer-Reviewed Support

Rajkomar, A., Dean, J., & Kohane, I. (2019). *Machine learning in medicine*. New England Journal of Medicine, 380(14), 1347–1358.
https://doi.org/10.1056/NEJMra1814259

Data Visualization

Software / Documentation

Matplotlib Developers. (2024). *Matplotlib Documentation*.
https://matplotlib.org/stable/

Waskom, M. (2024). *Seaborn Documentation*.
https://seaborn.pydata.org/

Model Evaluation & Class Imbalance

Peer-Reviewed Reference

Saito, T., & Rehmsmeier, M. (2015). *The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets*. PLOS ONE, 10(3), e0118432.
https://doi.org/10.1371/journal.pone.0118432

Clinical Trials & Healthcare Context

Peer-Reviewed References

Obermeyer, Z., & Emanuel, E. J. (2016). *Predicting the future—big data, machine learning, and clinical medicine*. New England Journal of Medicine, 375(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

Fogel, D. B. (2018). *Factors associated with clinical trials that fail and opportunities for improving the likelihood of success*. Contemporary Clinical Trials Communications, 11, 156–164. https://doi.org/10.1016/j.conctc.2018.08.001

Appendix: Learning Platforms (Not Scientific Evidence)

The following resources were used strictly for coding support and conceptual clarification.

Stack Overflow. https://stackoverflow.com/
freeCodeCamp. https://www.freecodecamp.org/
YouTube tutorials (various creators)
Reddit (r/datascience, r/MachineLearning)