

colorectal cancer prediction

Machine Learning

Focusing on sleep disorders



Compliance – Machine Learning

Presented by: Avraham Levi and Noa Tsivoni

Table of *Contents*

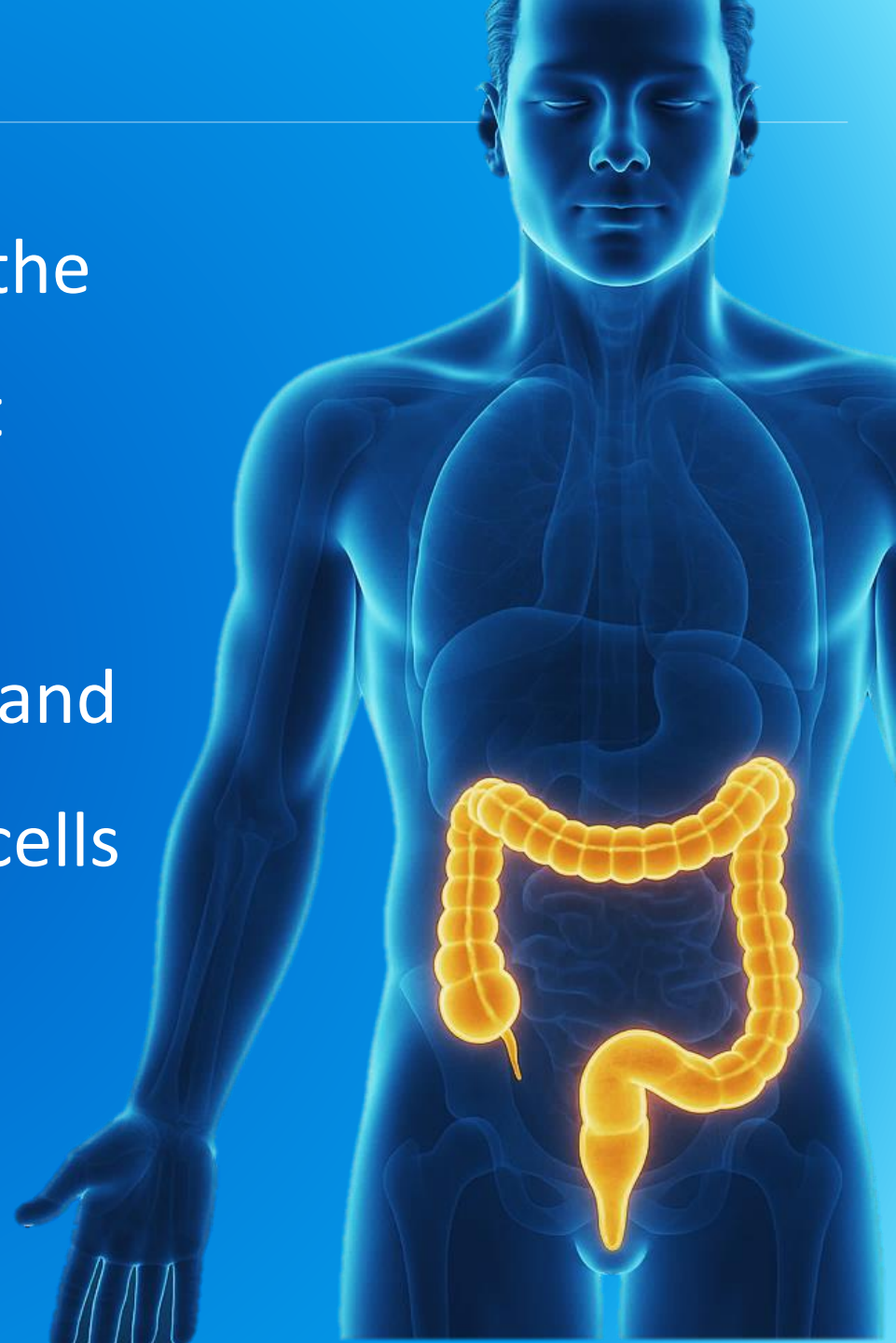


Introduction



◆ Cancer is a disease where certain cells in the body **start growing out of control** and don't stop when they should.

Normally, the body controls how cells grow and die. But in cancer, this control is lost — the cells keep growing, even when the body doesn't need them.





- ◆ **Colon cancer is the 3rd most common cancer in the world**
- ◆ **1.93 million new cases in 2022 (WCRF)**
- ◆ **Highest rates:** Europe, North America, Australia
- ◆ **Mostly affects people aged 50+ (WHO)**
- ◆ **Second most common cancer in Israel for both men and women**

Early detection of colorectal cancer can **significantly increase survival rates**, often allowing for less aggressive treatment and a better quality of life.

Unfortunately, symptoms in the early stages are usually **mild or go unnoticed**, such as changes in bowel habits, fatigue, or slight bleeding, which many people ignore or mistake for other issues.



Our Mission

- ◆ **Primary Goal:** To develop predictive models for colorectal cancer (CRC) risk using the UK Biobank dataset.
 - ◆ **Specific Focus:** To try specifically investigate the predictive value of **formal clinical sleep disorder diagnoses** within this cohort.
-

Literature Review

Xie, H., Xi, Z., Wen, S., et al. (2025). Associations between chronotype, genetic susceptibility and risk of colorectal cancer in UK Biobank. *Journal of Epidemiology and Global Health*, 15(1), 84.

<https://pubmed.ncbi.nlm.nih.gov/40208451/>

Lee, D. B., An, S. Y., Pyo, S. S., et al. (2023). Sleep fragmentation accelerates carcinogenesis in a chemical-induced colon cancer model. *International Journal of Molecular Sciences*, 24(5), 4547.

<https://www.mdpi.com/1422-0067/24/5/4547>

Lin, C. L., Liu, T. C., Wang, Y. N., Chung, C. H., & Chien, W. C. (2019). The association between sleep disorders and the risk of colorectal cancer in patients: A population-based nested case-control study. *In Vivo*, 33(2), 549–555.

<https://iv.iarjournals.org/content/33/2/573>

Chen, J., Chen, N., Huang, T., Huang, N., Zhuang, Z., & Liang, H. (2022). Sleep pattern, healthy lifestyle and colorectal cancer incidence. *Scientific Reports*, 12(1), 18317.

<https://www.nature.com/articles/s41598-022-21879-w>

UK Biobank



UK Biobank

Over
500,000

volunteers recruited

Lifestyle & Health
Questionnaires

Covers sleep habits, diet, physical activity, mental health, and more

Medical
Diagnostics

Linked hospital, cancer, and death records for long-term health tracking

Biological
Samples

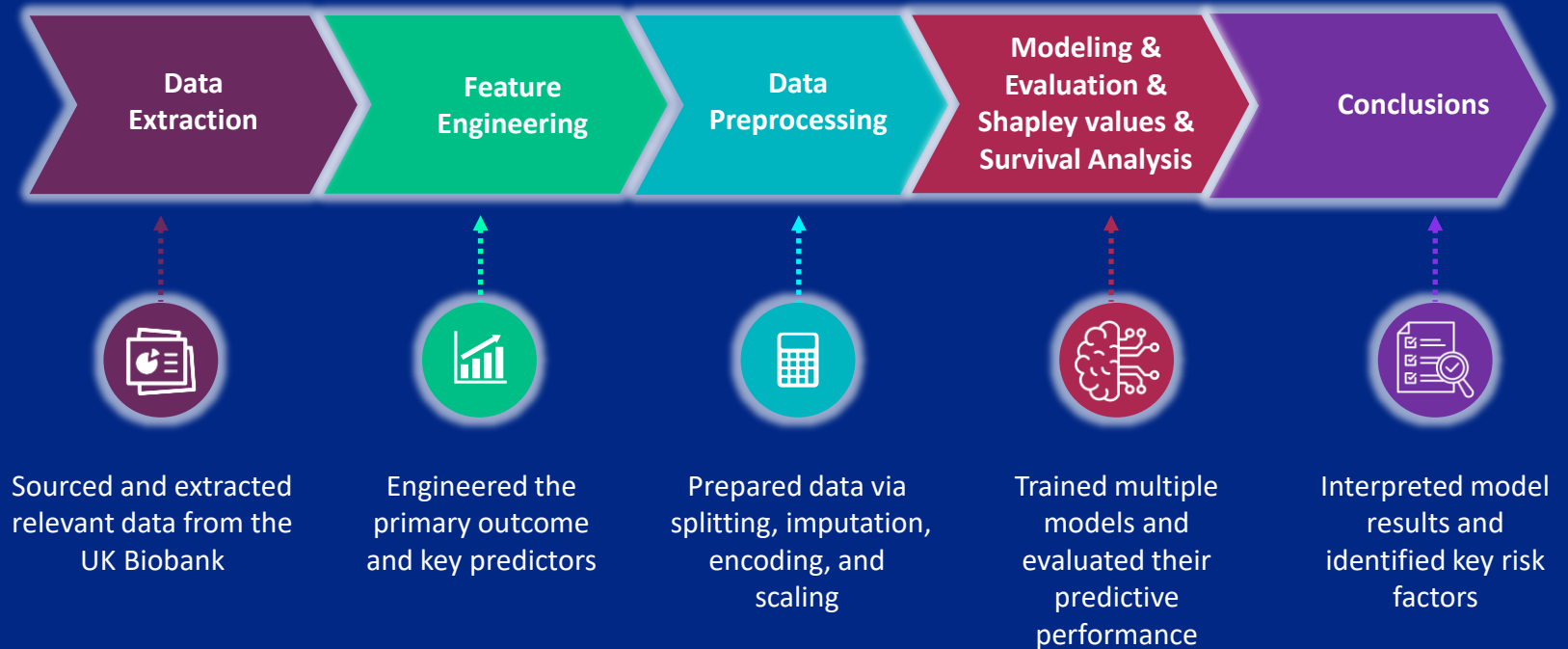
Includes blood, urine, and saliva samples and more

Process



Process

We began with the raw UK Biobank data, from which we engineered our outcome variable and key predictors. The data was then carefully prepared for modeling by splitting it into sets, handling missing values, and scaling features. Finally, we trained several machine learning models to predict CRC risk and understand the most significant factors.



Our Data

Our Data



Clinical Picture

Blood Markers
Waist circumference
Blood pressure
And more...



Lifestyle & Medical History

Age
Sex
smoking years
Alcohol consumption
Family history of colon cancer
diabetes
And more...



Sleep Profile

Clinical Diagnosis
(sleep apnea, insomnia, e.g.)

Self-Reported
(sleep duration, snoring habits,
daytime sleepiness, e.g.)

Our Data

Ensuring a Correct Timeline:

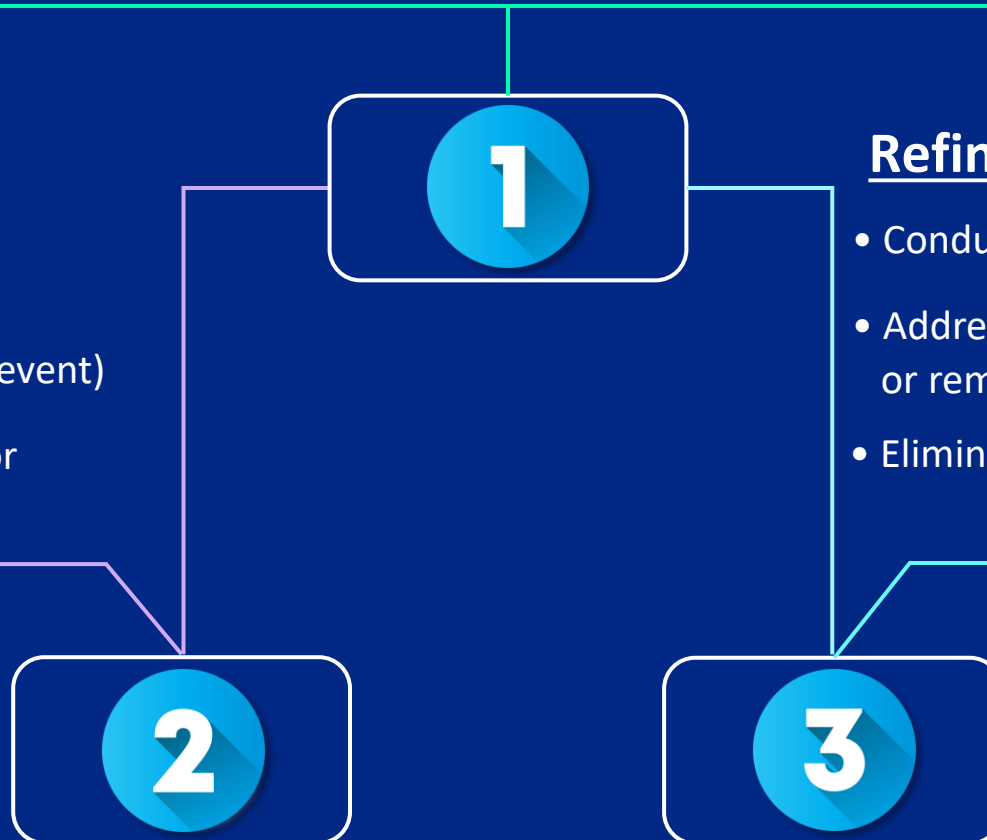
- Guaranteeing all predictors were measured before the outcome developed
 - Eliminating Reverse Causation by excluding 2,317 prevalent CRC cases
 - Defining a personal "Time Zero" (Baseline) for each participant

Feature Engineering

- Distilled arrays into meaningful binary flags (1/0)
- Engineering the Outcome Variable (CRC_event)
- Unifying Sleep Data into a Single Predictor

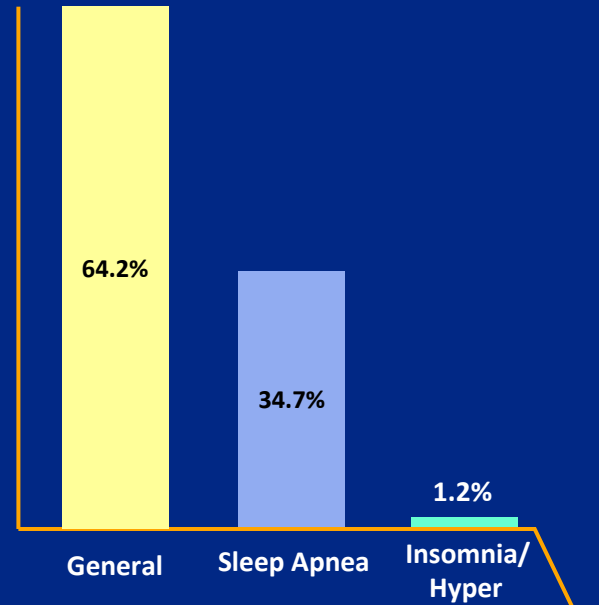
Refining the Data Based on EDA

- Conducted EDA to identify weaknesses
- Addressed high missingness by either transforming or removing features
- Eliminated redundant source data

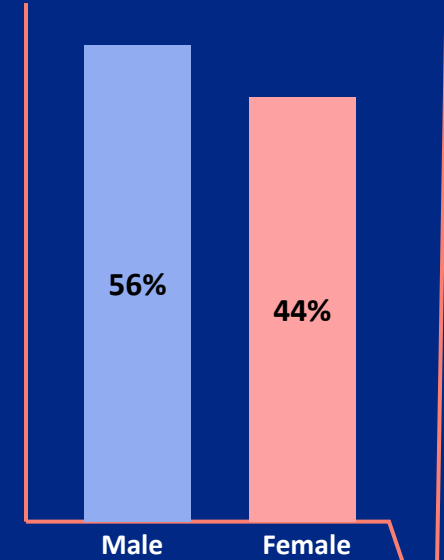


Our Data

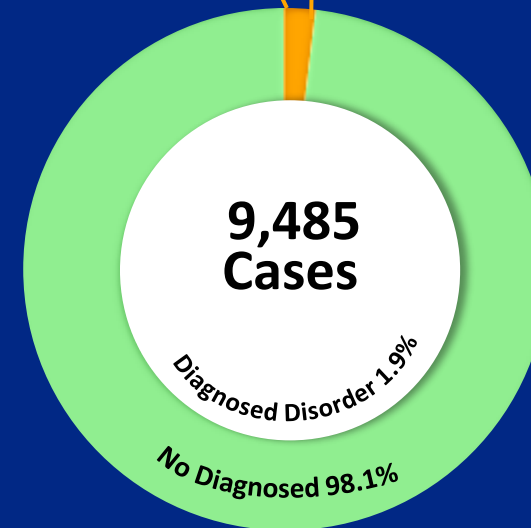
- ◆ We analyzed data from **500,053** participants
- ◆ The average participant was **57 years old**
- ◆ 54.5% women and 45.5% men



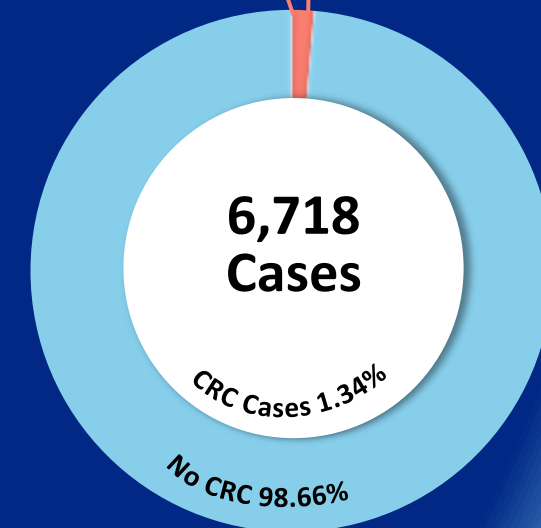
CRC Cases by Sex



Both CRC and diagnosed sleep disorders are rare events (<2%), highlighting the prediction challenge



Sleep Disorders



Colorectal Cancer

Our Data

Our Preprocessing Strategy: The Leak-Proof Framework

Preventing Data Leakage

Our guiding principle: the test set was kept separate and untouched, ensuring our final score is honest and not inflated

Splitting the Data First

We created our Train (70%), Validation (15%), and Test (15%) sets before any other processing step.

Learning All Rules from the Training Set

All preprocessing rules—from **imputation values to encoding categories**—were learned only from the training data

Missing Values (Imputation)

Filled numerical features using the median.
Filled categorical features using the mode.

Categorical Features

Used One-Hot Encoding

Outliers

Performed Outlier Capping on numerical features

Modeling Performance

Modeling

Our primary challenge was the severe class imbalance in the data. To thoroughly test our model's stability and capabilities, we designed **three distinct experimental scenarios**:

Scenario 1:

Perfect Balance

Small, perfectly balanced dataset by taking all 6,718 cancer patients and sampling an equal number of 6,718 healthy controls.

Scenario 2:

Balanced Training, Realistic Test

Fully balanced training set . The validation and test sets, however, remained imbalanced to reflect real-data conditions.

Scenario 3:

Full Imbalance

Data in its natural state, preserving the imbalanced state

Modeling

```
graph TD; Modeling[Modeling] --> LR[Logistic Regression]; Modeling --> RF[Random Forest]; Modeling --> XGB[XGBoost (tuned)];
```

Logistic Regression

classic, highly interpretable model. provides an excellent **baseline** for comparison and helps identify key predictors

It uses a logistic function to estimate the probability of a binary outcome

Random Forest

A powerful and flexible ensemble model capable of capturing complex interactions. By averaging many decision trees

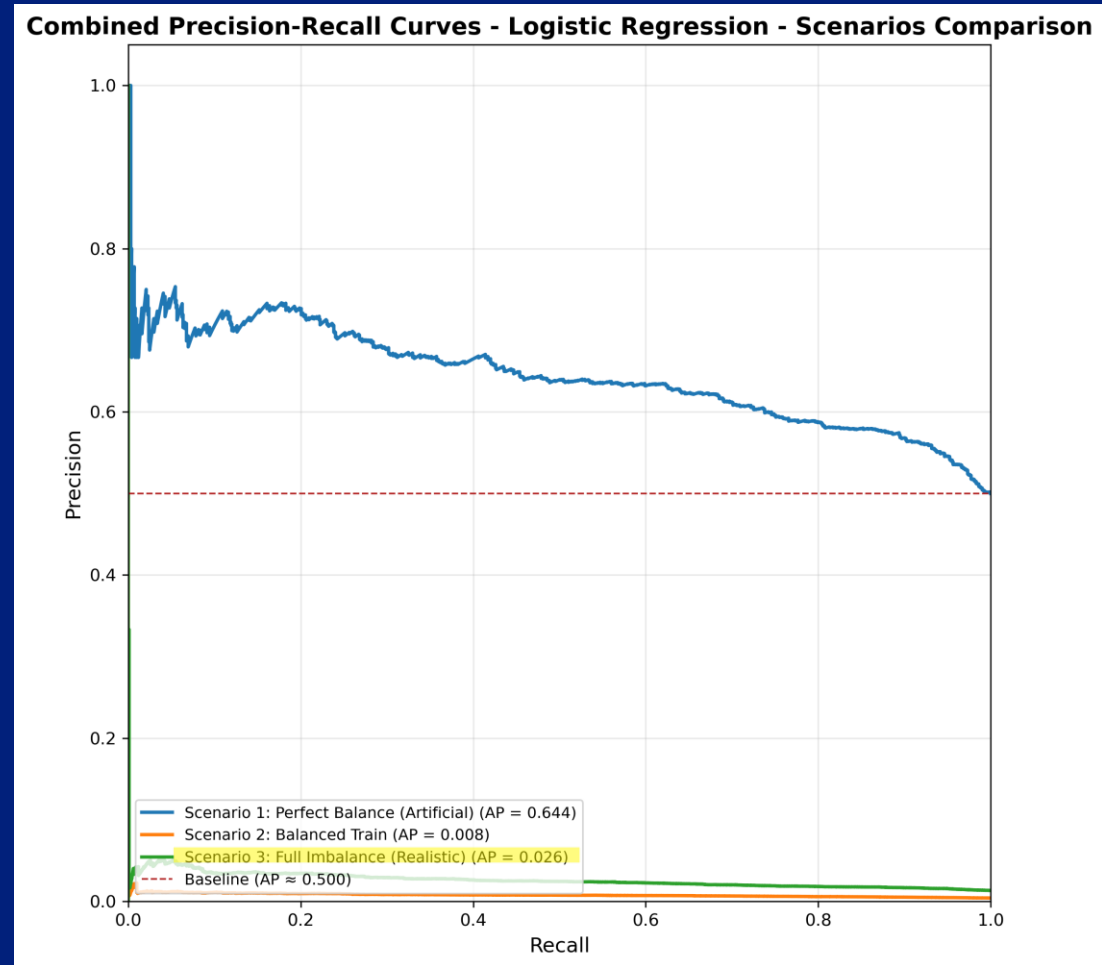
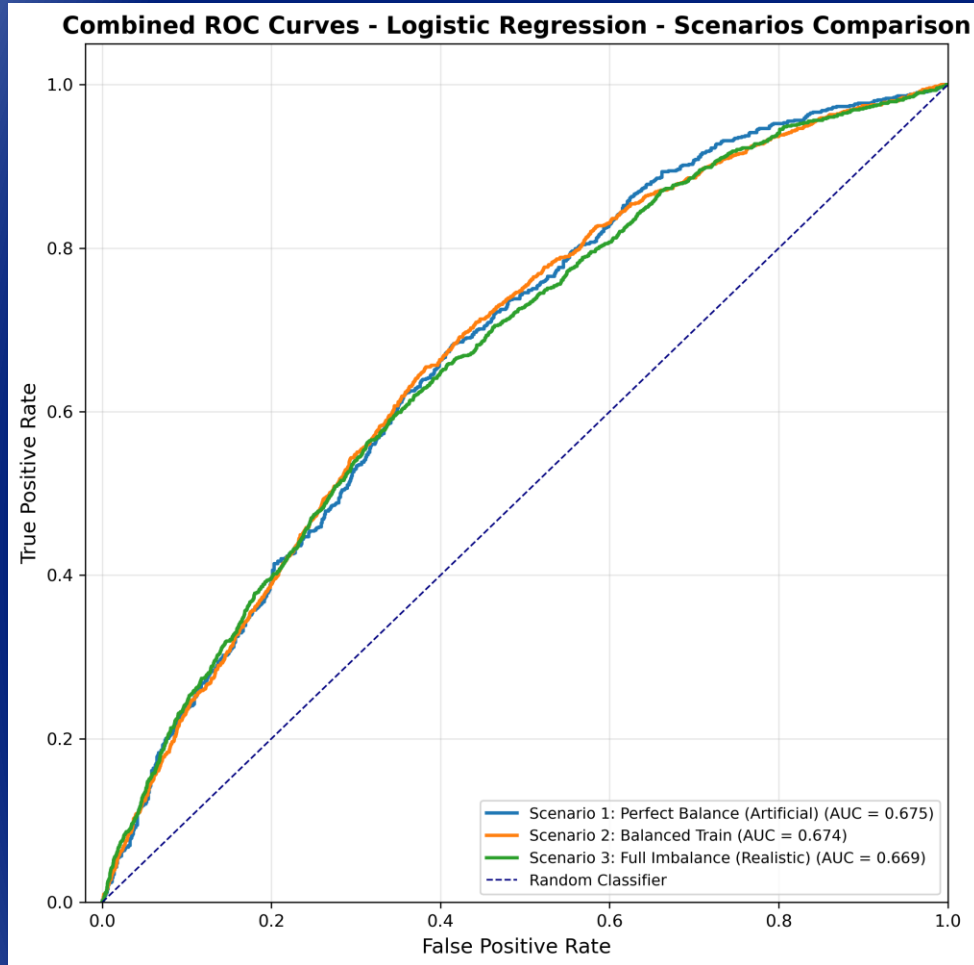
It builds a "forest" of many individual decision trees. Each tree is trained on a random subset of the data and features. To make a prediction, it collects the predictions from all its trees and takes the majority vote

XGBoost (tuned)

Considered as a superior performance and high accuracy model

It's a "boosting" algorithm, meaning it builds a sequence of decision trees. Each new tree in the sequence tries to correct the errors made by the previous trees.

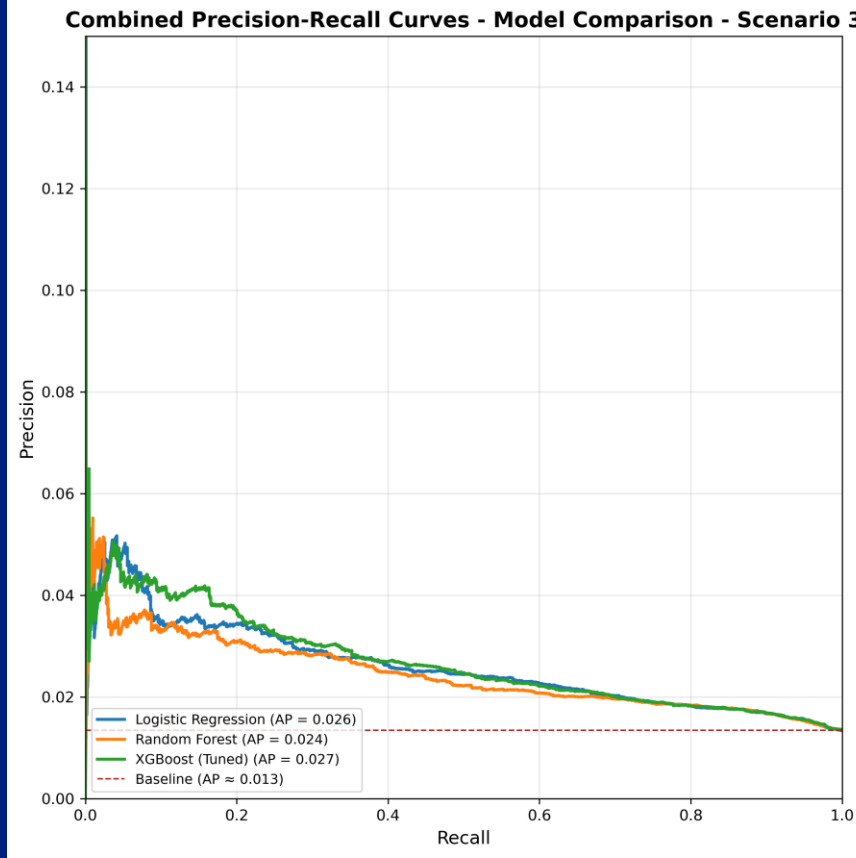
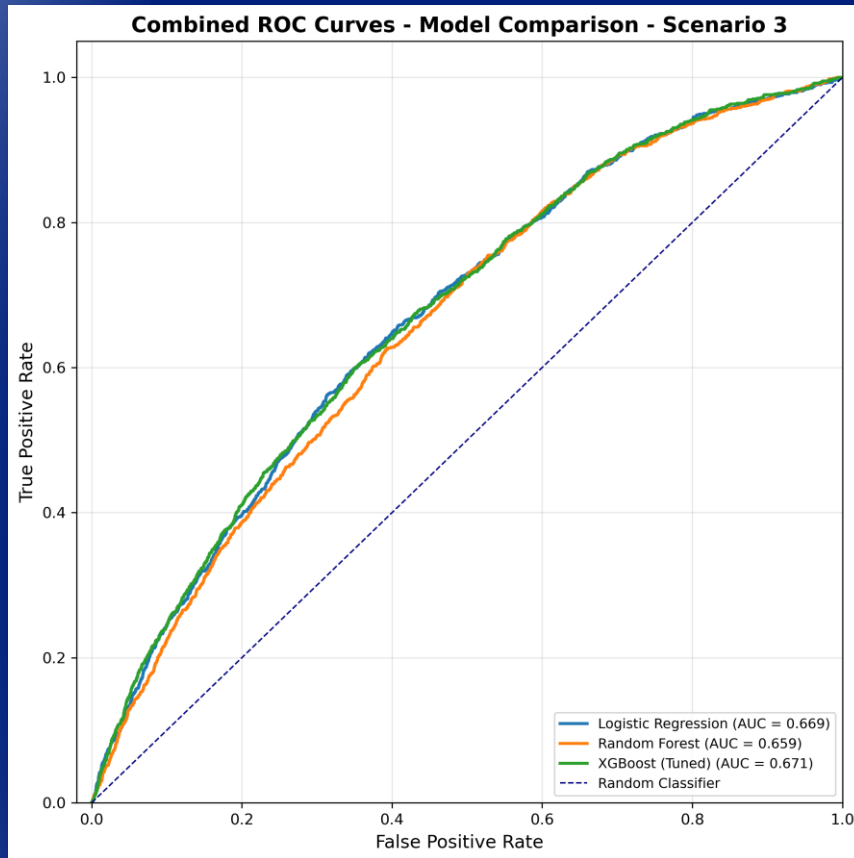
Logistic Regression



- ◆ **The ROC Deception:** ROC curves show **virtually identical performance** across all scenarios (AUC $\approx$ 0.67)
- ◆ **Scenario 3:** the most **realistic** approach, which incurred **no performance penalty** and demonstrated a very slight **advantage** on the PR curve

**View Full
Validation
Results
(Appendix)
—
Click Me**

Model Performance Comparison



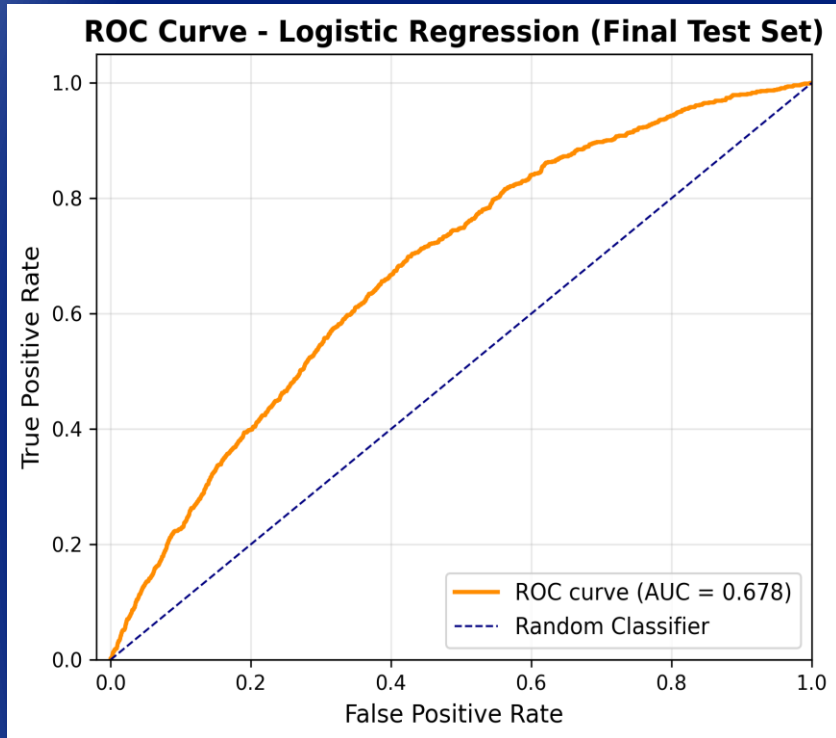
◆ **Consistent Performance**

Across Models: All models, from simple logistic regression to complex XGBoost, show remarkably similar performance on both the ROC and Precision-Recall curves.

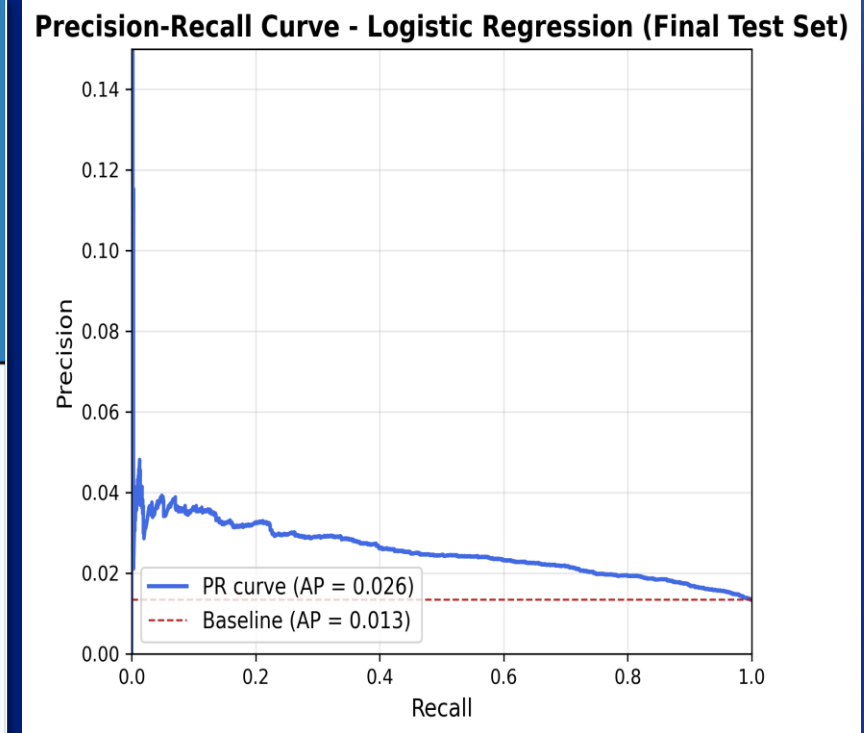
- ◆ The similar results indicate that a more complex model does not necessarily lead to better performance with this data.
- ◆ **The Takeaway:** The simpler, more interpretable **Logistic Regression** model performs just as well the advanced models, making it a highly valuable and efficient tool for this problem.

Final Verdict: Logistic Regression on Unseen Test Data

Evaluating the Model's Real-World Performance



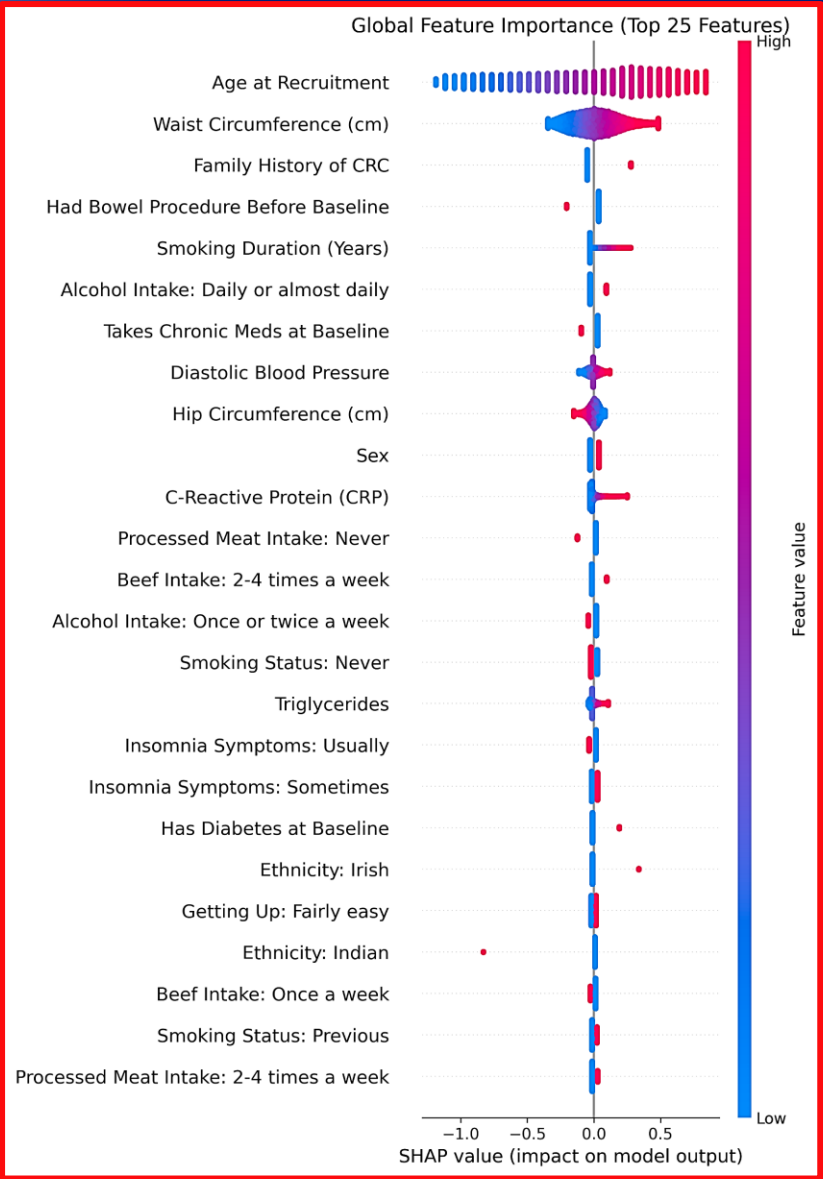
Confusion Matrix			
Actual: No CRC	True Negative 43,618 (58.2%)	False Positive 30,383 (40.5%)	
	Actual: CRC	False Negative 320 (0.4%)	True Positive 687 (0.9%)
Predicted: No CRC		Predicted: CRC	



- ◆ **The model successfully identifies over two-thirds of all actual cancer cases in the unseen data (68%)**
- ◆ **High Recall with a Big Trade-Off:** The model successfully identifies 68% of cancer cases (High Recall), but at the cost of a very low Precision (2.2%), generating a large number of false alarms.
- ◆ **A Low-Cost, First-Stage Assessment Tool.** The model's value is not in its precision, but in its use of **non-invasive, readily available data**. It shows potential as a preliminary tool for risk stratification, not a direct screening recommendation.

SHAP

Understanding the Key Drivers in our Logistic Regression Model

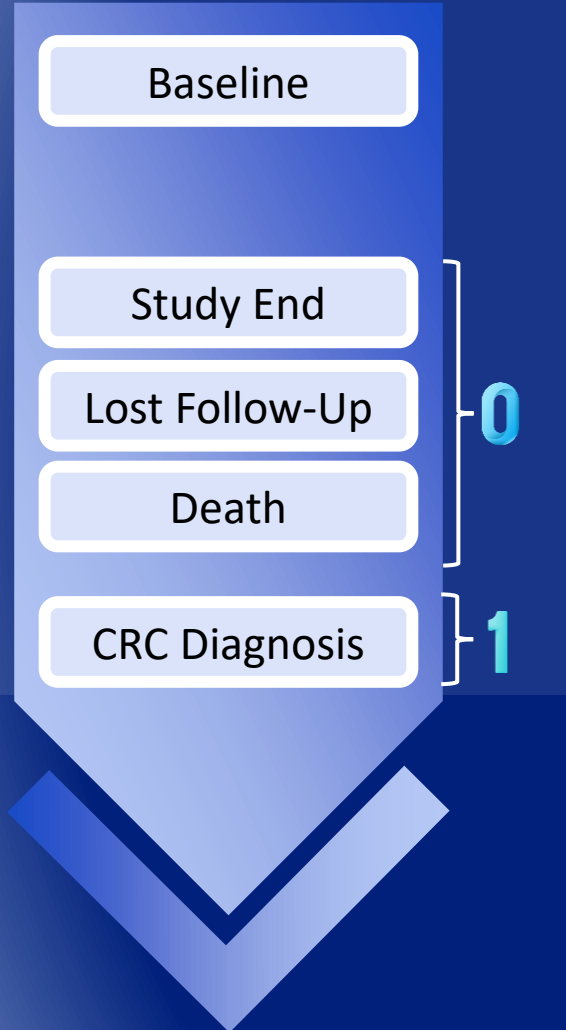


	coef	P > z	[0.025	0.975]
Age at Recruitment	0.557	<0.01	0.515	0.589
Waist Circumferences	0.1797	<0.01	0.115	0.244
Family History	0.0921	<0.01	0.067	0.117
Had Bowel Procedure	-0.073	<0.01	-0.103	-0.043
...	...	...	...	...
Sleep Disorder: Apnea	0.0189	0.108	-0.004	0.042
Sleep Disorder: Insomnia/Hypersomnia	-0.001	0.946	-0.031	0.029
Sleep Disorder: General/Unspecified	-0.003	0.838	-0.032	0.026

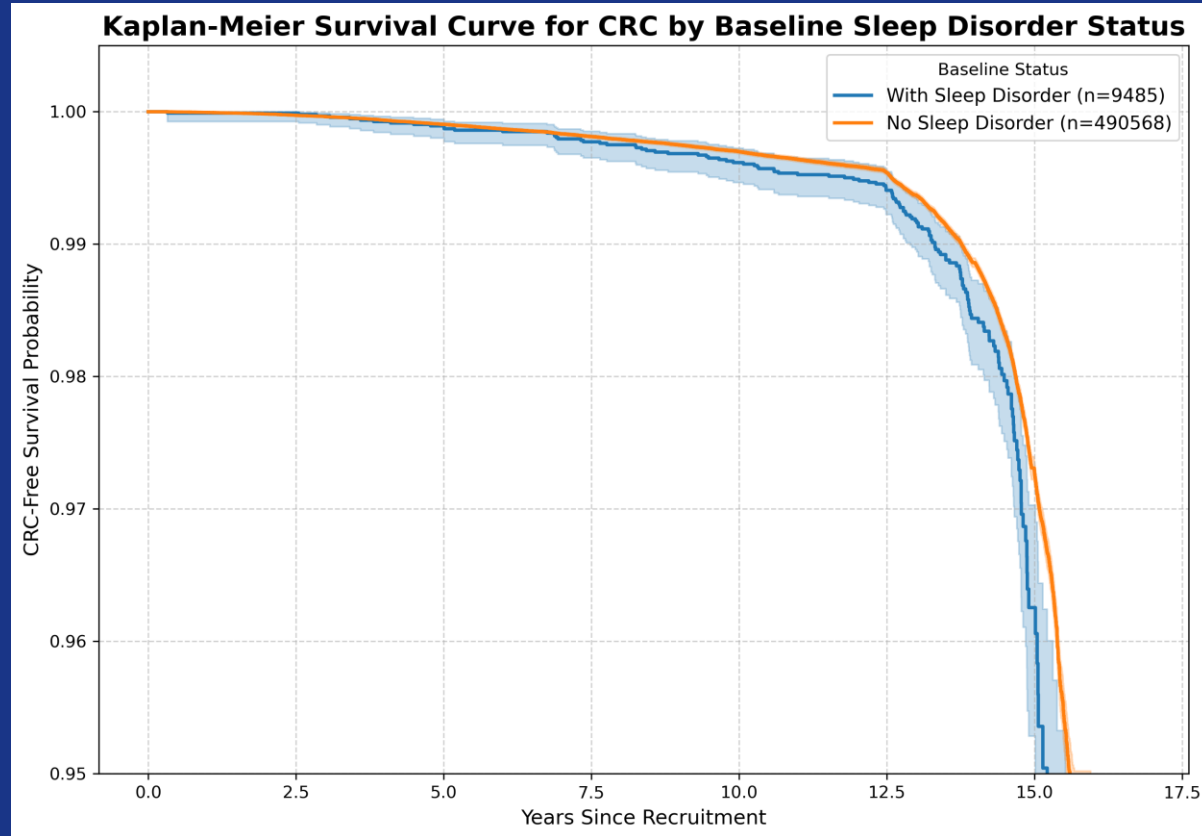
- ◆ Established risk factors dominate: Age, Family History, and Waist Circumference are confirmed as strong predictors
- ◆ Self-reported sleep variables appear in the top 25 features but have a demonstrably smaller impact than the primary risk factors.
- ◆ Key Insight: Our core variable, clinical Sleep Disorder Status, was not statistically significant

Survival Analysis

Do Sleep Disorders Accelerate CRC Risk Over Time ?



Kaplan - Meier



Cox Model

	HR	P > z	HR Lower 95% CI	HR Upper 95% CI
Has_Sleep_Disorder	1.099	0.26	0.932	1.297

◆ Kaplan-Meier curves show nearly identical survival probability, suggesting no strong, unadjusted effect between the groups.

◆ After adjusting for confounders, the Cox model confirms a **non-statistically significant** impact of sleep disorders on CRC risk (p=0.26).

Overall Conclusion: Consistent with our classification models, a formal clinical diagnosis was **not a primary driver of CRC risk over time** in this cohort.

Conclusions

Conclusions

Conclusions & Key Takeaways

A Reasonable Screening Tool for This Cohort

The model's value is not in its precision, but in its ability to identify a high-risk group using **non-invasive, readily available data**. It serves as a **potential preliminary tool** for risk stratification, not a diagnostic test.



No Statistically Significant Link for Clinical Diagnoses

Across all our analyses of **this specific dataset**, a **formal clinical diagnosis** of a sleep disorder was **not a statistically significant predictor** of colorectal cancer risk.



Model Validity Confirmed on Our Data

Within our dataset, the model correctly identified **Age, Family History, Waist Circumference, and Smoking Duration** as the most dominant risk factors, validating its ability to learn medically-recognized patterns.

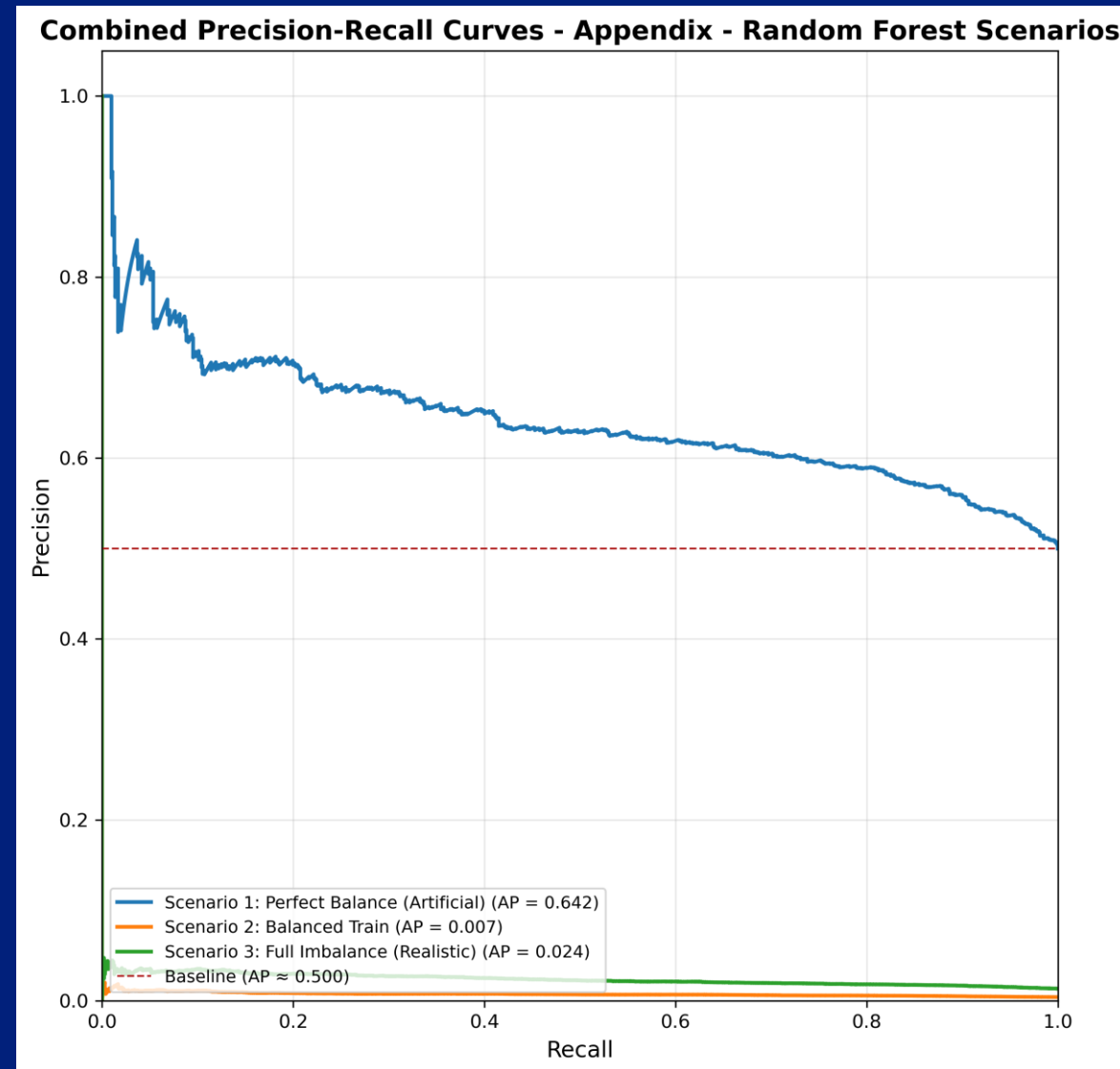
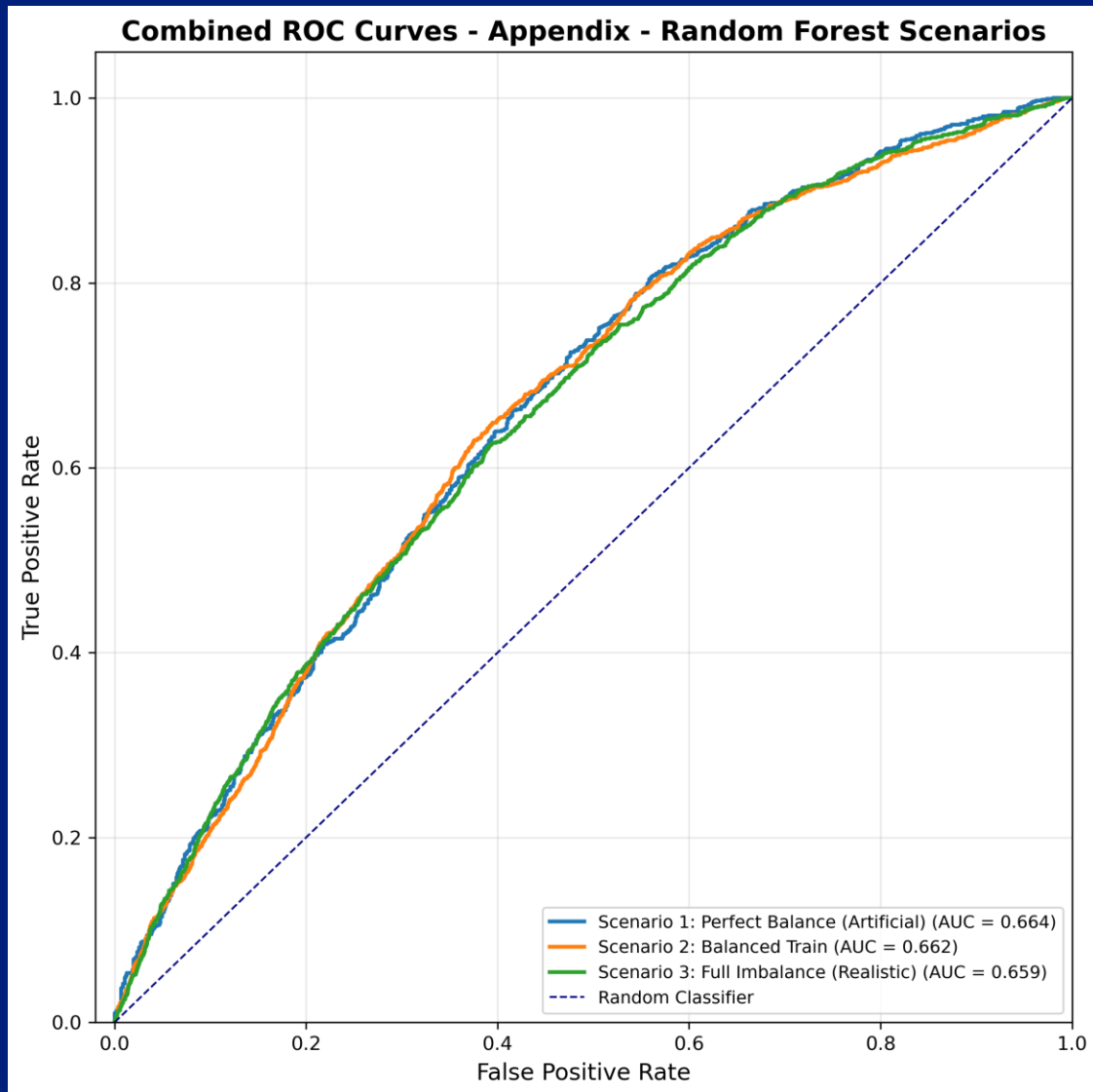


Questions



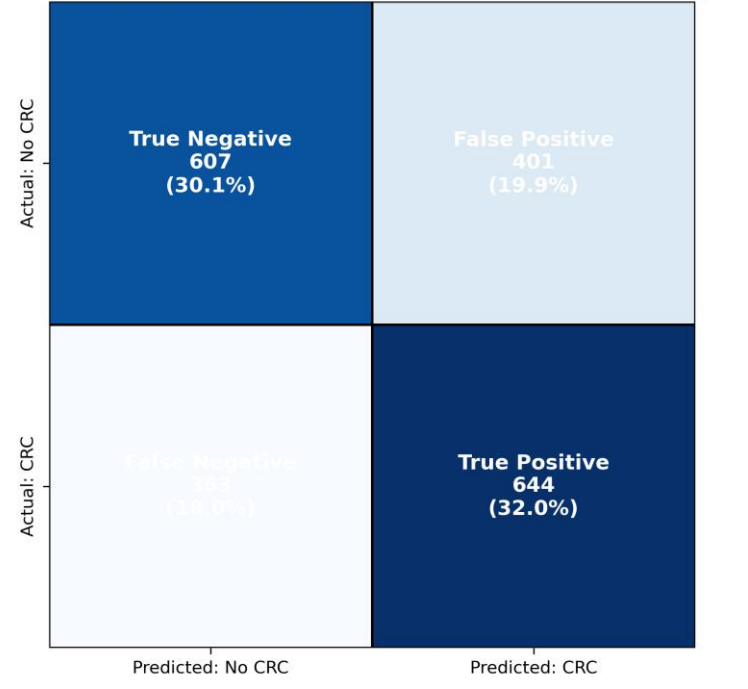
Appendix

Random Forest

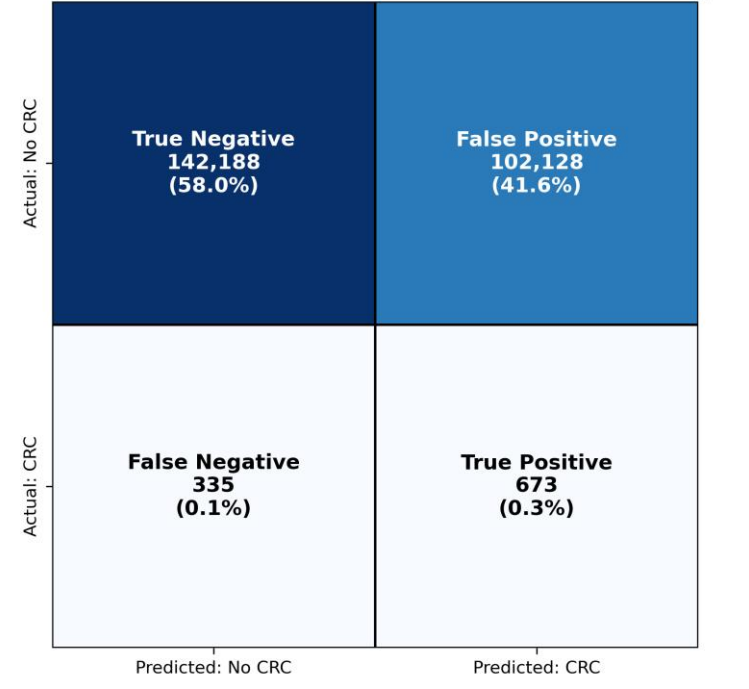


Random Forest Confusion Matrixes

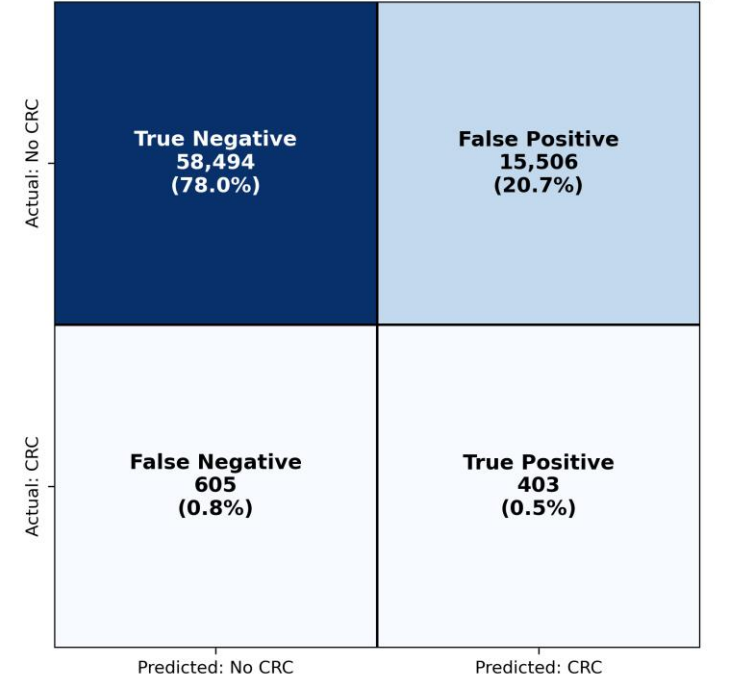
Confusion Matrix - Random Forest (scenario_1)



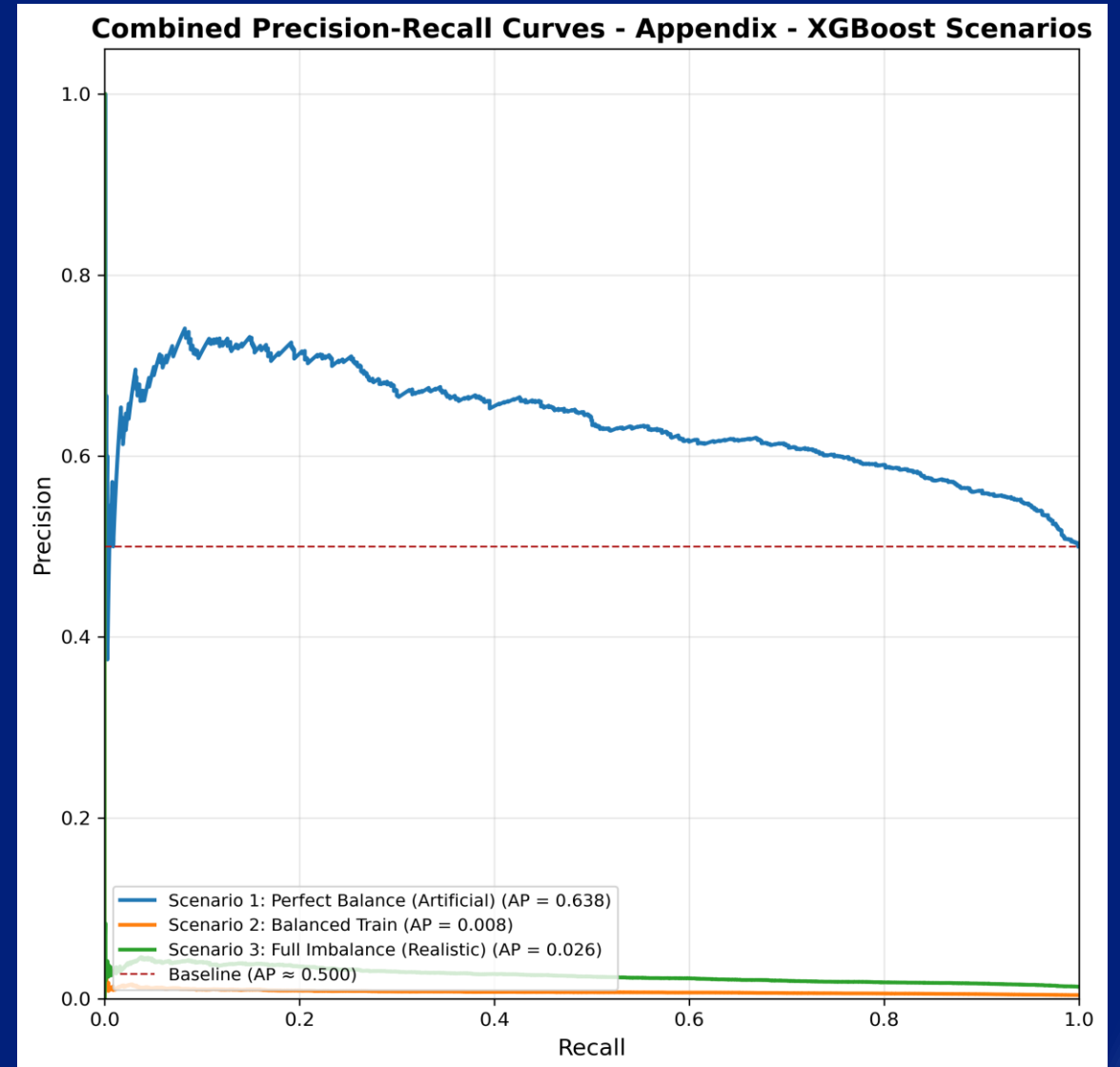
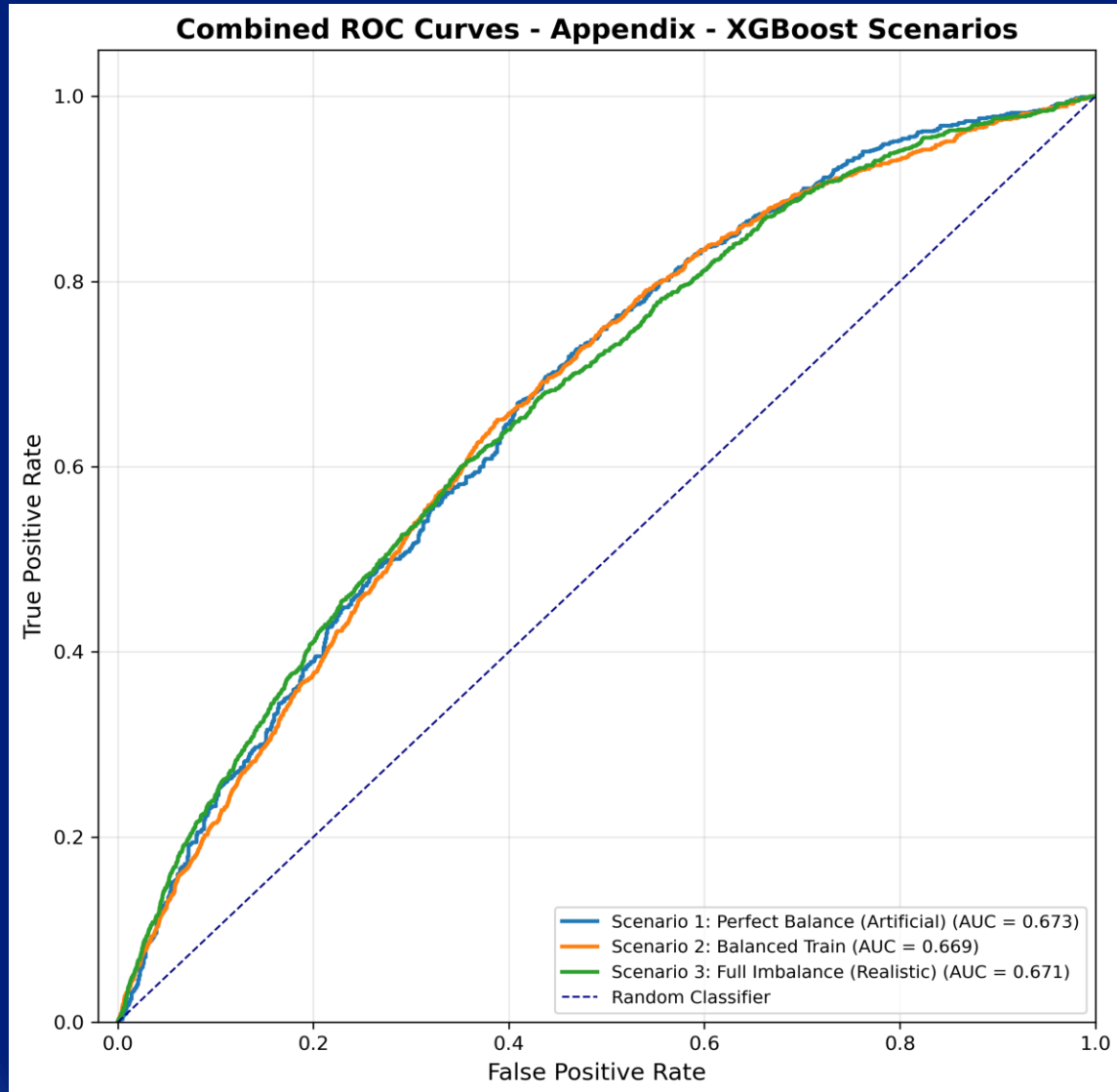
Confusion Matrix - Random Forest (scenario_2)



Confusion Matrix - Random Forest (scenario_3)



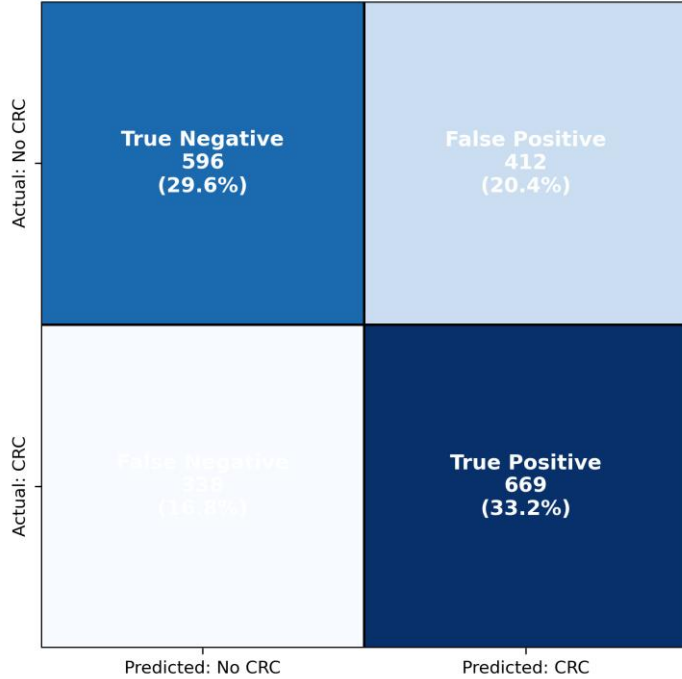
XG Boost



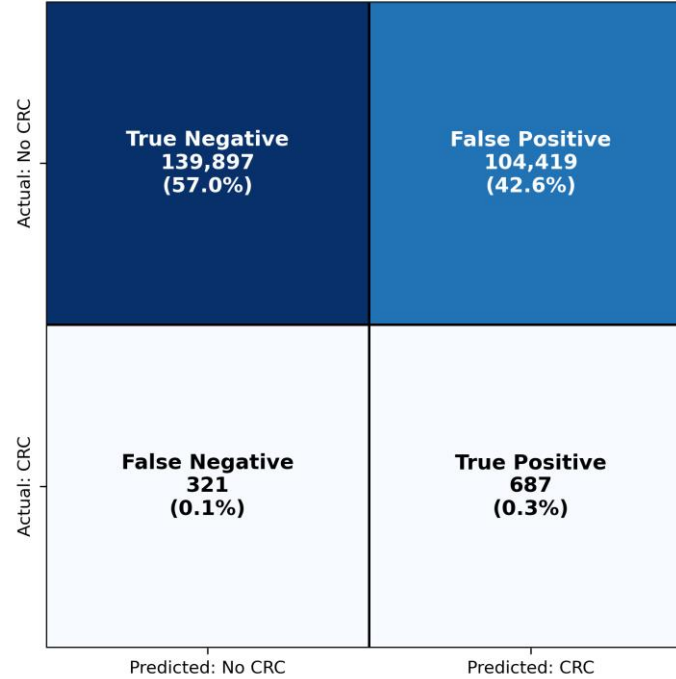
XG Boost

Confusion Matrixes

Confusion Matrix - XGBoost (Tuned) (scenario_1)



Confusion Matrix - XGBoost (Tuned) (scenario_2)



Confusion Matrix - XGBoost (Tuned) (scenario_3)

