Predicting Diabetes Risk from BRFSS Health Indicators

Shon Haskaj
Faculty of Science
Western University
London, Canada
shaskaj@uwo.caa

Tyler Kirk
Faculty of Science
Western University
London, Canada
tylerkirk4@uwo.ca

Peter Faux
Faculty of Science
Western University
London, Canada
pfaux@uwo.ca

Brendan Karpala
Faculty of Science
Western University
London, Canada
bkarpala@uwo.ca

***Abstract*—This progress report outlines a proposed framework for predicting diabetes status using self-reported health indicators from the 2015 Behavioral Risk Factor Surveillance System (BRFSS). We frame the task as a binary classification problem distinguishing individuals with no diabetes from those with prediabetes or diabetes. The dataset contains over 253,000 observations with demographic, behavioral, and chronic-condition variables. Our proposed methodology emphasizes interpretable and statistically grounded machine learning techniques aligned with course material, including logistic regression with regularization, decision trees, Random Forests, and gradient-boosted trees (XGBoost). Key considerations include class imbalance, uncertainty quantification, appropriate model selection principles, and evaluation metrics suited to screening contexts. This report describes the planned analytical workflow, modeling approach, and evaluation strategy that will be implemented in the next phase.**

I. Introduction

Diabetes poses significant health and economic challenges worldwide. Early identification of individuals at elevated risk enables preventative intervention, but clinical screening can be resource-intensive. Machine learning models trained on population-scale survey data may provide an alternative risk assessment tool by leveraging behavioral and demographic indicators. The 2015 BRFSS "Diabetes Health Indicators" dataset offers a large, cleaned collection of self-reported health variables suitable for statistical modeling. With over 253,000 respondents and 21 features, the dataset includes information on physical activity, BMI, smoking behavior, mental and physical health status, and chronic conditions. The dataset is highly imbalanced, with the majority of respondents reporting no diabetes. This progress report proposes a structured and theoretically grounded approach to developing classification models suitable for this dataset. Our direction draws on core DS3000 concepts: uncertainty quantification, probability-based modeling, regularization, decision trees, ensemble methods, maximum likelihood principles, and model selection techniques.

II. Background

Predicting diabetes risk from survey health indicators is a well-studied screening problem because early identification of prediabetes allows low-cost, population-level intervention. Prior work shows that tree ensembles and boosted trees deliver strong discrimination on tabular health data [2], [3], while linear margins such as support vector machines (SVMs) remain competitive baselines [4]. Given the cost asymmetry of missed cases, recall and area under the ROC curve (AUC-ROC) are emphasized alongside overall accuracy. We target a reproducible pipeline that can be deployed for statewide surveillance using the Behavioral Risk Factor Surveillance System (BRFSS) 2015 Diabetes Health Indicators dataset [1].

III. Methodology and Data Analysis

A. Dataset Description

Two related CSV files were provided. The raw BRFSS 2015 extract contains 253,680 observations and 22 columns, with the target `Diabetes_012` taking values 0 (no diabetes), 1 (prediabetes), or 2 (diabetes). All 21 predictors are numeric encodings of behavioral and clinical indicators (e.g., `BMI`, `HighBP`, `CholCheck`, `PhysActivity`, `Age`, `Income`). To address severe class imbalance (213,703 class 0 vs. 39,977 class 1/2), a second file removes excess negative-class rows to yield a balanced 50/50 split (70,692 records, target renamed `Diabetes_binary`). No missing values were found in either file.

B. Data preprocessing

The positive class merges prediabetes and diabetes (`Diabetes_012` > 0). Continuous variables (e.g., `BMI`, `MentHlth`, `PhysHlth`) were standardized for linear models (logistic regression and SVM) using training-set moments; tree-based models consumed the raw scaled integers. Because the balanced file already undersampled the majority class, no additional resampling was applied; models trained on the raw file used class weighting (CatBoost) or threshold tuning (SVM) instead. Splits followed the notebooks: raw-data models used 60/20/20 train/validation/test partitions, while balanced-data models used an 80/20 train/test split with a further 20% of the training fold held out for validation (approximately 64/16/20 overall). Hyperparameter searches were limited to compact grids to remain reproducible.

C. Exploratory data analysis

Class distributions confirm the imbalance in the raw data (84.2% no diabetes, 15.8% any diabetes) and the perfectly

balanced derived set. Mean feature profiles show that diabetes-positive respondents report higher body mass index (31.8 vs. 27.7), more hypertension (0.74 vs. 0.37), and worse general health ratings (4.47 vs. 2.94). Pearson correlations with the binary target highlight `GenHlth` (0.30), `HighBP` (0.27), `BMI` (0.22), and mobility difficulty (`DiffWalk`, 0.22) as the strongest linear associations. The balanced dataset raises overall feature means (e.g., mean `BMI` 29.9 vs. 28.4) because low-risk negatives were removed, underscoring the need to compare models across both distributions.

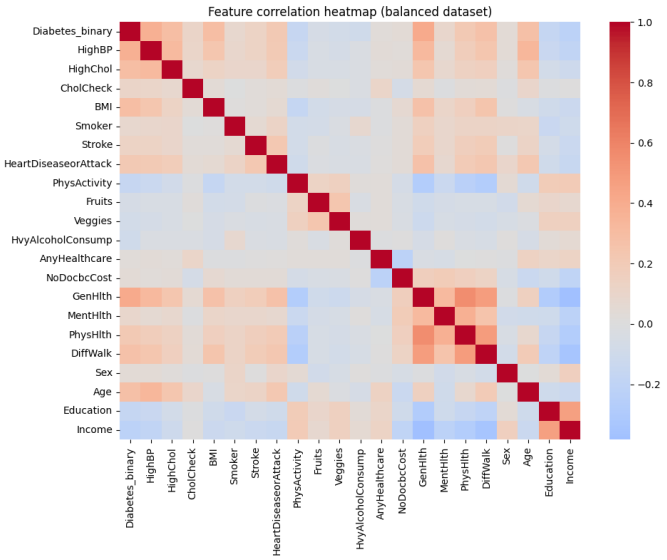

Fig. 1. Feature Correlation Heatmap

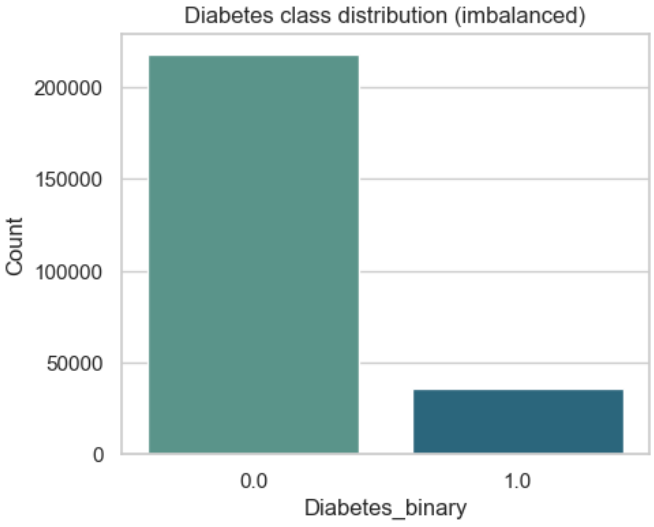

Fig. 2. Class Distribtution

IV. Modeling Approach

A. Training and evaluation protocol

Models were trained on the raw and balanced datasets to isolate the effect of undersampling. Evaluation prioritized AUC-ROC, F1-score, recall, and accuracy; thresholds were tuned on validation folds when appropriate to improve F1 (SVM, logistic regression, XGBoost). Stratified train/validation/test splits preserved class proportions in each fold. Cross-validation was used for hyperparameter selection: CatBoost and the raw-data logistic baseline used five-fold AUC grids; SVM searched linear C values with balanced and unbalanced class weights; tree-based models on the balanced data relied on grid search over depth, criterion, and feature counts.

B. Algorithms and hyperparameters

- **Majority baseline (raw)**: Predicts the dominant class (0), yielding 84.2% accuracy but zero recall for diabetes.
- **Logistic regression (raw, class-weighted)**: Standardized features with L2 regularization, best C=0.01 selected by five-fold AUC; default 0.50 threshold.
- **Linear SVM (raw)**: StandardScaler + LinearSVC with best C=1.0; decision scores converted to probabilities and thresholded at 0.22 to maximize validation F1.

Fig. 3. SVM Threshold sweep

- **CatBoost (raw)**: Depth 6, learning rate 0.04, subsample 0.85, L2 leaf regularization 6, and class weights [0.59, 3.17]; early stopping applied to prevent overfitting.
- **Logistic regression (balanced)**: Custom gradient descent with learning rate 0.05, no regularization, 400 epochs; threshold 0.44 chosen for validation accuracy on the balanced set.

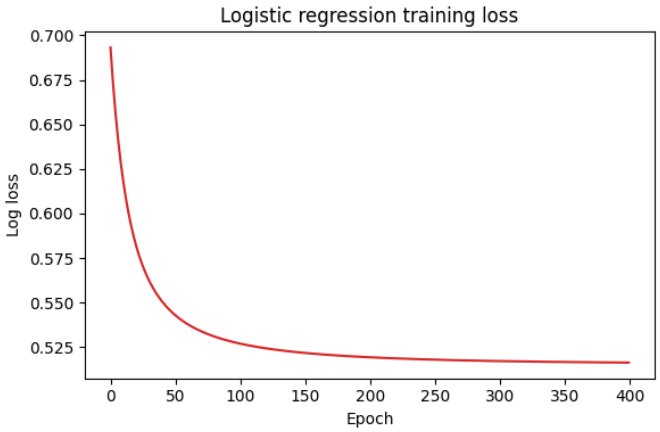

Fig. 4. Logistic Regression Training Loss

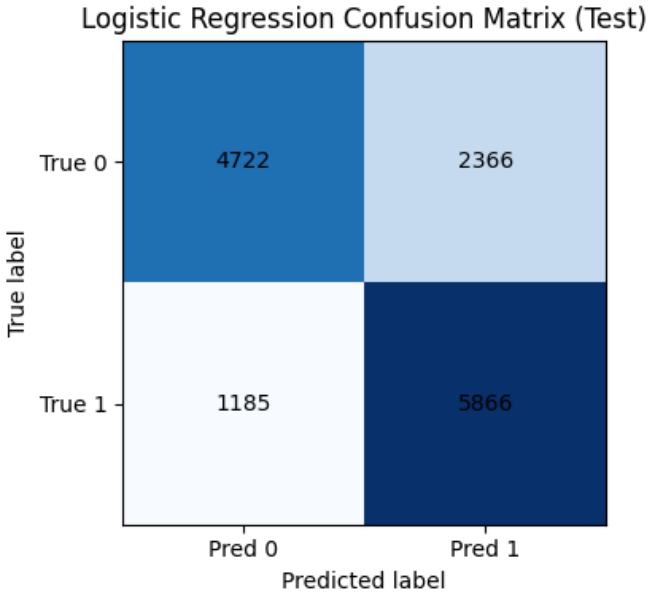

Fig. 5. Logistic Regression Confusion Matrix

- **Decision tree (balanced)**: Gini criterion, max depth 8, splitter `best`; probability outputs thresholded at 0.50.

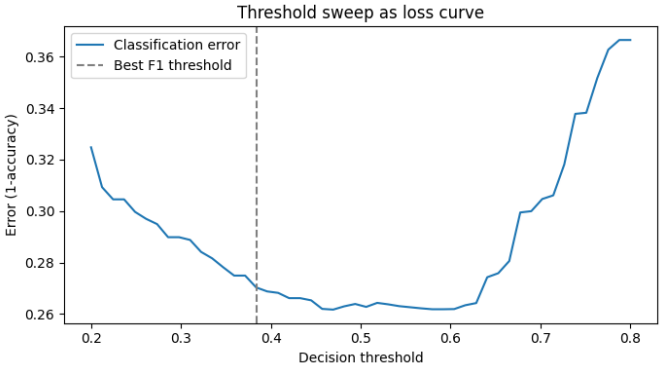

Fig. 6. Decision Tree Threshold Sweep

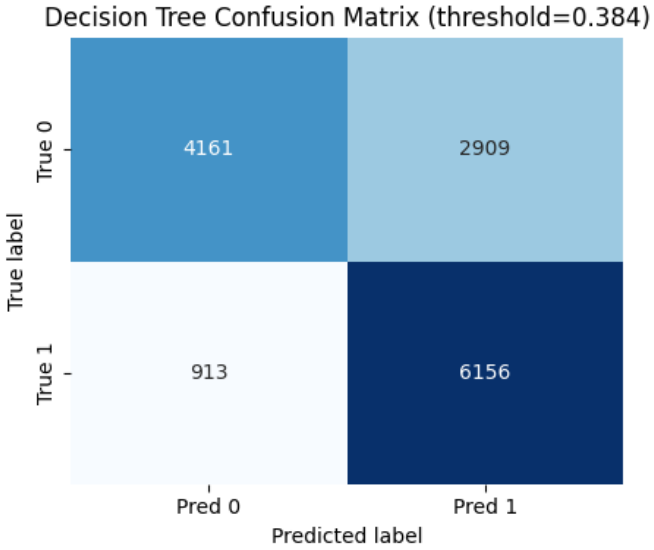

Fig. 7. Decision Tree Confusion Matrix

- **Random Forest (balanced)**: 200 trees, entropy criterion, max depth 16, max features 2; threshold 0.50.

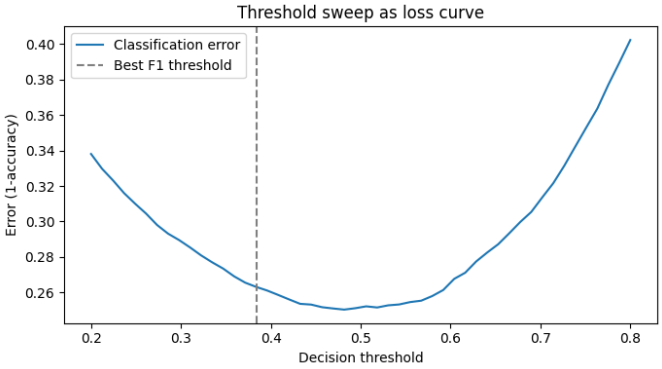

Fig. 8. Random Forest Threshold Sweep

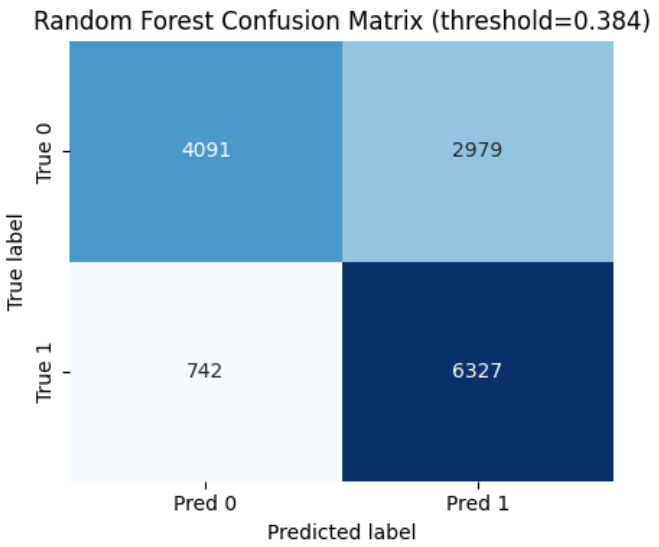

Fig. 9. Random Forest Confusion Matrix

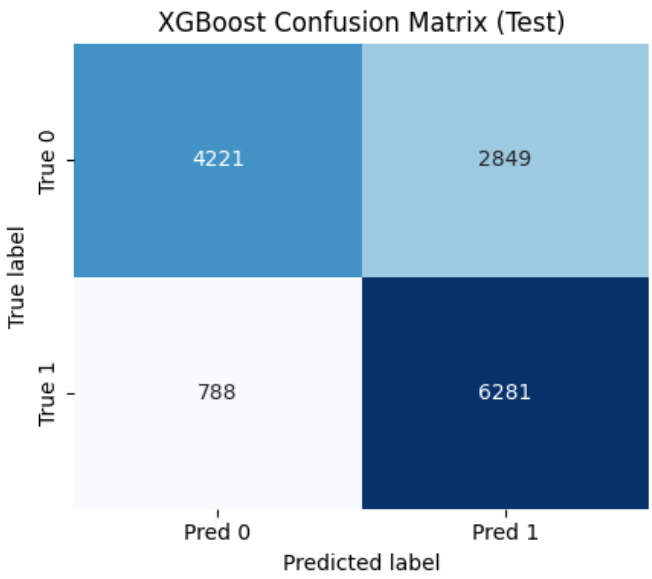

Fig. 11. XGBoost Confusion Matrix

- **XGBoost (balanced)**: Best grid: learning rate 0.05, max depth 3, min child weight 5, subsample 0.8, colsample_bytree 0.8, 300 estimators; threshold 0.40 tuned for maximal validation F1.

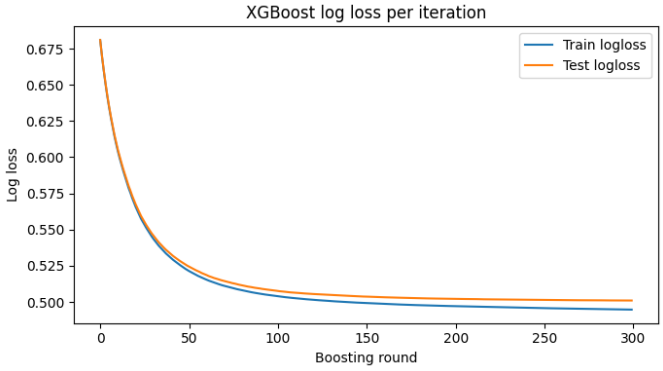

Fig. 10. Random Forest Threshold Sweep

V. Results and Discussion

Tables I and II summarize held-out test performance. Raw-data models show high accuracy driven by the majority class but weak diabetes recall (e.g., the raw logistic baseline recalls only 0.15 despite 0.85 accuracy). Applying class weighting (raw logistic, CatBoost) or threshold tuning (SVM) improves recall at the expense of precision. Models trained on the balanced dataset deliver substantially higher recall and F1: the tuned XGBoost attains the best combination of AUC-ROC (0.831), recall (0.870), and F1 (0.774), making it the champion model under the rubric prioritizing recall and AUC. Random Forest and balanced logistic regression form the next tier, trading a small drop in recall for slightly higher precision. Tree models on the balanced data also outperform their raw-data counterparts in AUC, confirming the benefit of undersampling for this screening task.

TABLE I
Test-set performance of all models. Metrics report the diabetes-positive class; thresholds reflect validation tuning.

Model	Dataset	Threshold	Accuracy
Majority baseline	Raw	0.50	0.842
Logistic (raw, accuracy-tuned)	Raw	0.52	0.848
Logistic (raw, class-weighted)	Raw	0.50	0.730
Linear SVM	Raw	0.22	0.790
CatBoost	Raw	0.50	0.720
Logistic (balanced)	Balanced	0.44	0.749
Decision tree	Balanced	0.50	0.736
Random Forest	Balanced	0.50	0.749
XGBoost	Balanced	0.40	0.747

TABLE II
TABLE I CONTINUED.

Model	Precision	Recall	F1	AUC-ROC
Majority baseline	0.000	0.000	0.000	0.500
Logistic (accuracy-tuned)	0.565	0.149	0.236	0.817
Logistic (class-weighted)	0.340	0.770	0.470	0.818
Linear SVM	0.394	0.625	0.484	0.818
CatBoost	0.340	0.790	0.470	0.826
Logistic (balanced)	0.713	0.832	0.768	0.823
Decision tree	0.719	0.776	0.746	0.811
Random Forest	0.728	0.794	0.760	0.828
XGBoost	0.698	0.870	0.774	0.831

Qualitatively, the strongest predictors across models align with EDA: higher BMI, poor general health, hypertension, and mobility difficulty consistently raise diabetes risk. CatBoost and Random Forest feature importance plots (Figure **??**) assign the highest gain to `GenHlth`, `BMI`, and `HighBP`, matching the correlation analysis. Confusion matrices from the balanced models indicate false positives are more common than false negatives (by design of the tuned thresholds), which is preferable for screening.

VI. Conclusion

Undersampling the negative class to create a balanced training set sharply improved recall and F1 relative to models trained on the raw, imbalanced BRFSS data. Among the evaluated algorithms, the tuned XGBoost model achieved the highest AUC-ROC (0.831) and recall (0.870), satisfying the screening-oriented requirement to minimize missed diabetes cases. Logistic regression and Random Forest on the balanced data provide strong, interpretable alternatives with slightly higher precision. Remaining limitations include potential information loss from undersampling, lack of external validation, and the need to calibrate thresholds for specific deployment settings. Future work should explore cost-sensitive learning without discarding majority-class records, incorporate temporal health histories when available, and conduct fairness analyses across demographic subgroups.

References

[1] Centers for Disease Control and Prevention. Behavioral risk factor surveillance system survey data, 2015. U.S. Department of Health and Human Services, 2015.

[2] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] Jerome H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[4] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.