GlucoPredict

Machine Learning for Noninvasive Glucose Tracking

Prepared By: Michael Walker

1. Introduction

Problem:

- Prediabetes affects 1 in 3 people, with a 10% annual risk of developing type 2 diabetes.
- No noninvasive, commercially available methods exist for self-management.

Project Highlights:

- Explored the feasibility of using smartwatches and food logs to predict glucose levels.
- Leveraged 25,000 simultaneous glucose, food log, and smartwatch measurements.

Outcome:

- Developed a machine learning model achieving a 13% Mean Absolute Percent Error in real-time glucose prediction.

Target Audience

- Healthcare Professionals
- Researchers: Data scientists and academics.
- Patients: Individuals managing prediabetes.
- Engineers: Those working on health devices.
- Investors: Those interested in health tech.

2. Data Overview + Project Outline

This dataset is downloaded from a study conducted at Duke University:

- **16 participants**
- **8-10 days** using **Dexcom G6** and **Empatica E4 devices**.
- **25,000+ interstitial glucose readings**, along with PPG, EDA, skin temperature, heart rate, interbeat interval, and triaxial accelerometry data, all stored in CSV files.
- **Food logs** were included
- **Demographic details** and HbA1c values recorded.

Files were available for each of the 16 patients and were merged for comprehensive analysis:

- ACC.csv (Accelerometer data)
- BVP.csv (Blood Volume Pulse data)
- Dexcom.csv (Glucose readings)
- EDA.csv (Electrodermal activity)
- Food_Log.csv (Food intake log)
- HR.csv (Heart rate data)
- IBI.csv (Interbeat interval data)
- TEMP.csv (Temperature data)
- DEMOGRAPHICS.csv (Demographics data)

Project Outline:

Data Wrangling + Preprocessing
  - Dataset: Over 25,000 glucose readings from Dexcom G6 and physiological data from Empatica E4, plus food logs.
  - Tasks: Clean, synchronize, and integrate.

Feature Engineering
  - Features: Derived from wearables (PPG, EDA, heart rate, accelerometry) and dietary logs.
  - Goal: Convert raw data into inputs for modeling.

Model Training
  - Models: Various machine learning algorithms were trained to predict glucose levels.
  - Metrics: Evaluated using Root Mean Square Error (RMSE) and Mean Absolute Percent Error (MAPE).

Real-Time Glucose Prediction
  - Application: Provides what would be real-time glucose predictions based on current wearable and food log data.

Evaluation
  - Validation: Assessed through Leave-One-Group-Out cross-validation (LOGO-CV) and error metrics.

3. Data Wrangling and Cleaning

Wearable Data

- **Issue**: Align patient data with 5-minute glucose readings.
- **Solution**: Early Feature Engineering with statistics, resampled to 5-minute intervals, and applied universal wrangling.

Food Log Data

- **Issue**: Inconsistent column names and formats.
- **Solution**: Standardized columns, merged logs, forward filled.

Wearable + Food Log + Demographic Data Integration

- **Issue**: Required alignment of multiple data types.
- **Solution**: Preprocessed, merged datasets, encoded categories.

Wrangling Overview

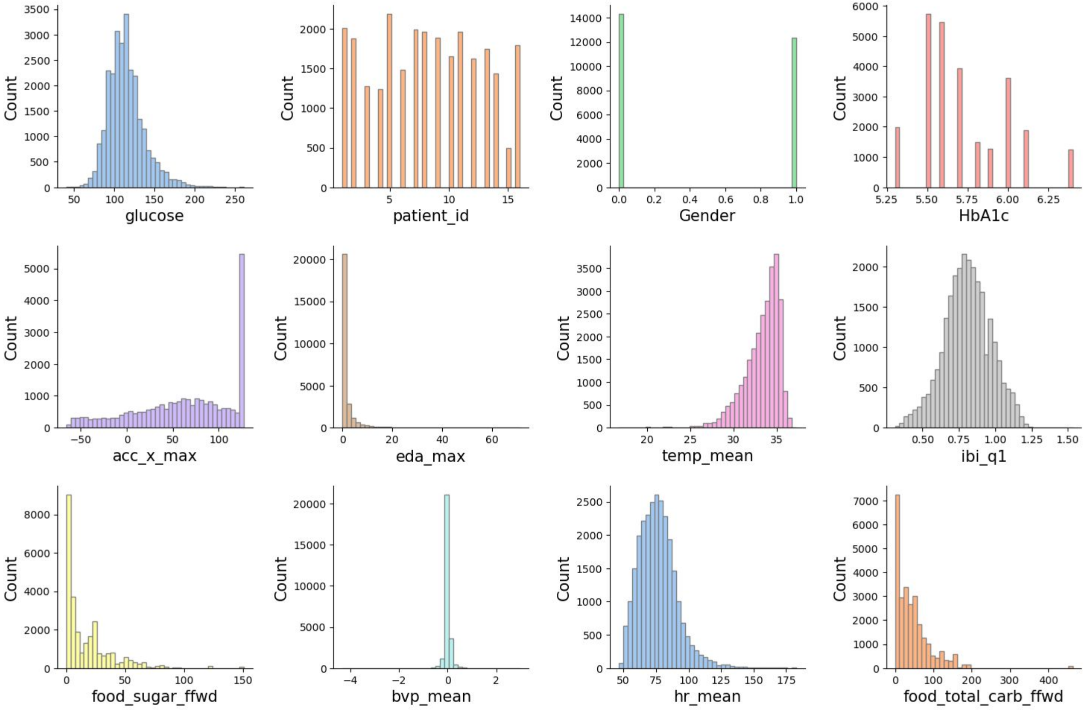
Wearables df	Food Log df	Demographics df	Combined df
Resampled aligned with Glucose sampling	Removed unnecessary columns	Sex, HbA1c, and Patient ID information	Integrated each df into a single, larger df

Required Resampling on Glucose

CSV	Description	Source	Sampling Period
ACC_001	Tri-axial accelerometry (X-Y-Z)	Empatica E4	0.03 s
BVP_001	Blood volume pulse	Empatica E4	0.02 s
Dexcom_001	Interstitial glucose concentration (mg/dL)	Dexcom G6	300.00 s
EDA_001	Electrodermal activity	Empatica E4	0.25 s
HR_001	Heart Rate	Empatica E4	1.24 s
IBI_001	Interbeat interval	Empatica E4	0.98 s
TEMP_001	Skin Temperature	Empatica E4	0.25 s
food_log	Food intake with time and nutritional information	User input	As needed
demographics_csv	Sex, HbA1c, Patient ID	User input	One time

4. Exploratory Data Analysis (EDA)

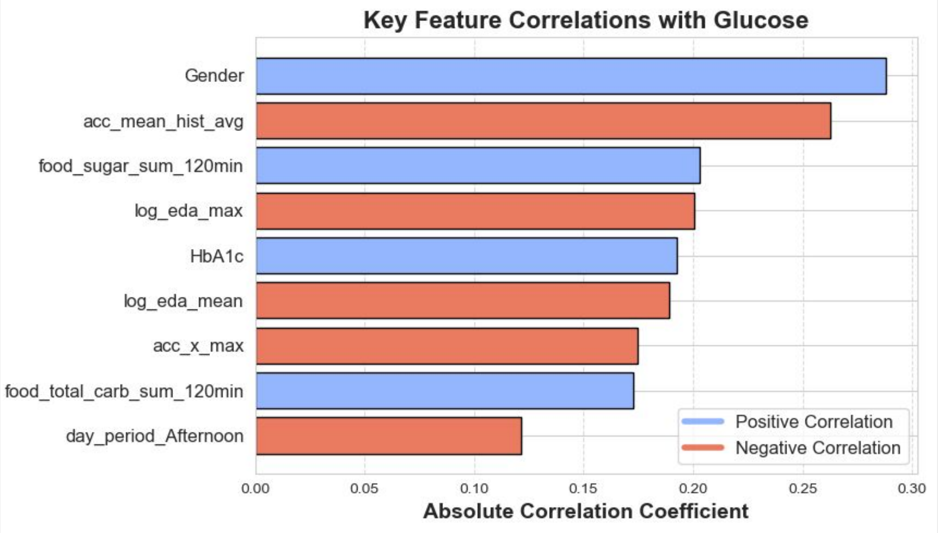
- **Glucose**: Roughly normal distribution, centered 100-120 mg/dL, mostly 70-150 mg/dL.
- **Patient ID**: Most patients have ~1,500 samples; patient 15 has 500 samples.
- **acc_x_max**: Right-skewed with many high values.
- **eda_max**: Highly skewed, mostly near zero.
- **temp_mean**: Left-skewed, mostly 32-34°C.
- **ibi_q1**: Slightly left-skewed, centered around 0.75s
- **food_sugar_ffwd**: Heavily right-skewed, mostly near zero with a long tail.
- **bvp_mean**: Tightly centered around zero.
- **hr_mean**: Normally distributed around 75-85 bpm.
- **food_total_carb_ffwd**: Heavily right-skewed.



5. Feature Engineering

Feature Enhancements:

- **Log Transformations**: Improved alignment with glucose levels.
- **Time-Based Features**:
  - Time since midnight
  - Day of the month
  - Weekend status
  - Total elapsed time
- **Categorical Features**: One-hot encoding for times of day (Night, Morning, Afternoon, Evening).
- **Rolling Statistics**:
  - Cumulative sums
  - Dietary intake metrics
  - Rolling sum windows for meal counts, wake time, and activity bouts

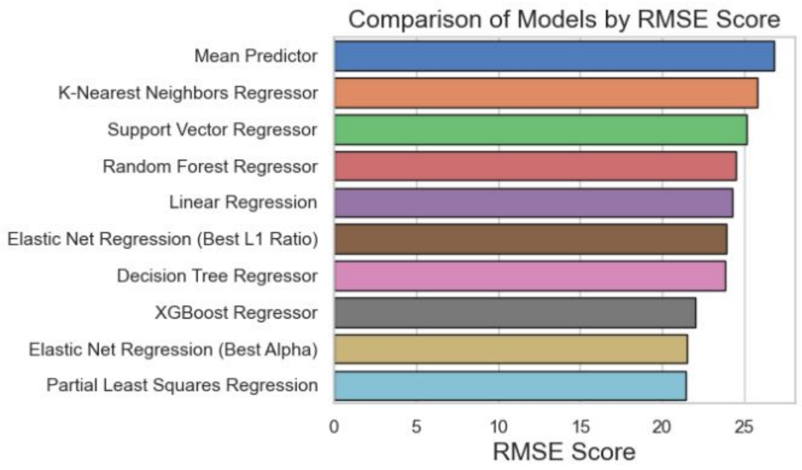


1. **Gender**: Highest correlation with glucose.
2. **acc_mean_hist_avg**: Negatively correlated; lower glucose with higher historical activity.
3. **food_sugar_sum_120min**: Positively correlated; reflects glucose spikes post-sugar intake.
4. **HbA1c**: Positively correlated; indicates higher long-term glucose levels.
5. **log_eda_mean**: Negatively correlated; lower glucose with higher mean electrodermal activity.
6. **log_eda_max**: Negatively correlated; lower glucose with higher electrodermal peaks.
7. **acc_x_max**: Negatively correlated; lower glucose with higher max acceleration.
8. **food_total_carb_sum_120min**: Positively correlated; glucose increases with carb intake.
9. **day_period_Afternoon**: Negatively correlated; lower glucose in the afternoon.

6. Pipeline & Model Survey

Model Pipeline Overview:

1. **Data Extraction**: First 25% of data
   a. **Training Data**: Patients 3-16
   b. **Testing Data**: Patient 2
2. **Preprocessing**:
   a. Impute missing values with mean
   b. Standardize features with StandardScaler
3. **Model Evaluation**:
   a. **Function**: evaluate_model()
   b. **Metrics**: MAE, RMSE, R^2, MAPE
4. **Visualization**:
   a. Compare true vs. predicted values
   b. Plot RMSE vs. parameter values
5. **Hyperparameters Tested**: Max depth, number of neighbors, C, etc.



Model Survey Discussion:

Partial Least Squares Regression had the best RMSE score.

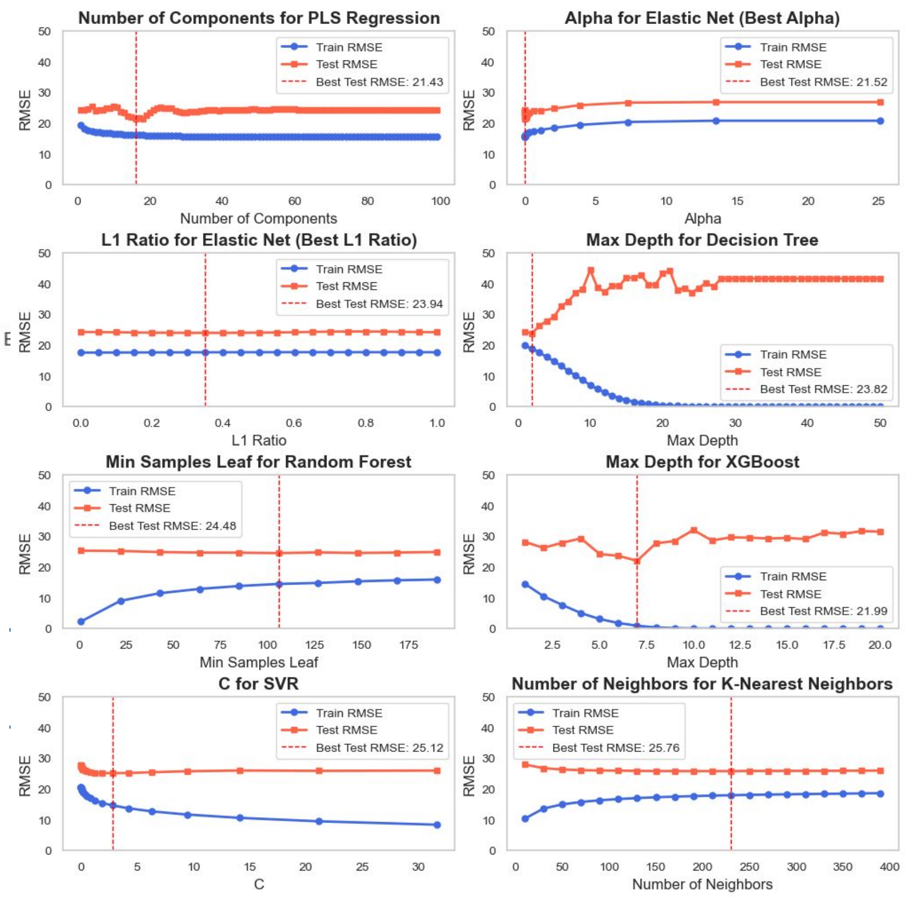
Next Steps:

1. Optimize PLS on full dataset
2. Choose a more complex model (XGBoost) and optimize hyperparameters on full dataset

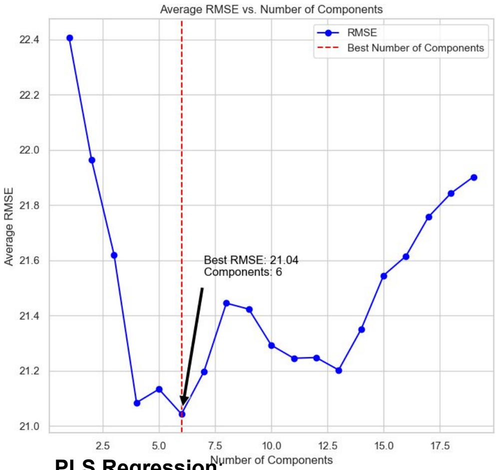
6b. Pipeline & Model Survey

Model Pipeline Overview:

1. **Data Extraction**: First 25% of data
   a. **Training Data**: Patients 3-16
   b. **Testing Data**: Patient 2
2. **Preprocessing**:
   a. Impute missing values with mean
   b. Standardize features with StandardScaler
3. **Model Evaluation**:
   a. **Function**: evaluate_model()
   b. **Metrics**: MAE, RMSE, R^2, MAPE
4. **Visualization**:
   a. Compare true vs. predicted values
   b. Plot RMSE vs. parameter values
5. **Hyperparameters Tested**: Max depth, number of neighbors, C, etc.



7. Final Model Optimization



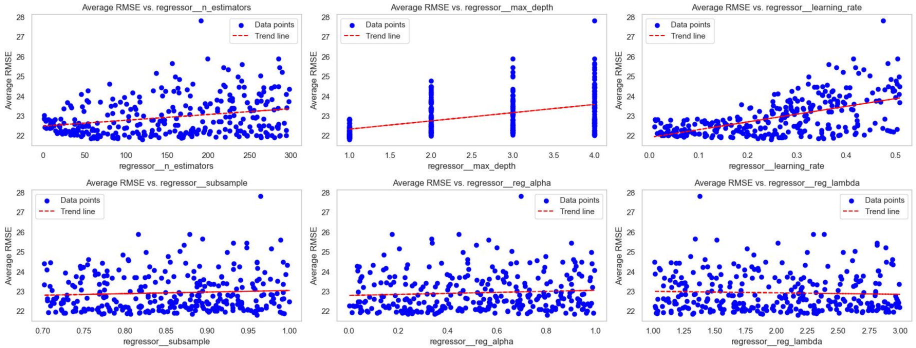
Optimization Process:

1. **Models**: PLS Regression & XGBoost
2. **Method**: RandomizedSearchCV with Leave-One-Group-Out Cross-Validation (LOGO CV)
3. **Data**: Patients 2-16 (excluding Patient 1)
4. **Pipeline**: Imputation and Standardize features
5. **Hyperparameters**: Tuned for optimal performance
   a. PLS: N_Components
   b. XGBoost: n_estimators, max_depth, learning_rate, reg_alpha, reg_lambda, subsample

PLS Regression:
Optimized N_Components: 6
RMSE: 21.04 (Best Performance)

XGBoost:
- **Optimized Parameters**:
- Learning rate: 0.247
- Max depth: 1
- 120 estimators
- reg_alpha: 0.806
- reg_lambda: 2.171
- Subsample: 0.960

RMSE: 21.80



8. Final Model Performance

Test Setup:

- Patient 1 excluded from training
- Models: XGBoost & PLS Regression
- Tuning: RandomizedSearchCV with LOGO CV (patients 2-16)

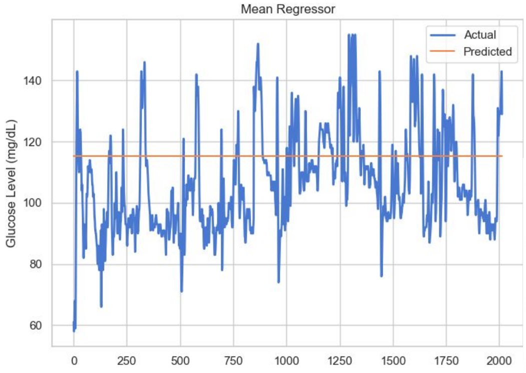
Results:

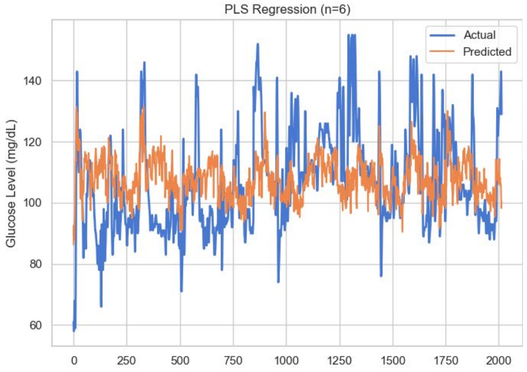
- **XGBoost**: RMSE 15.84 mg/dL, MAPE 11.56%
- **PLS**: Similar performance
- Both significantly outperformed Mean Regressor

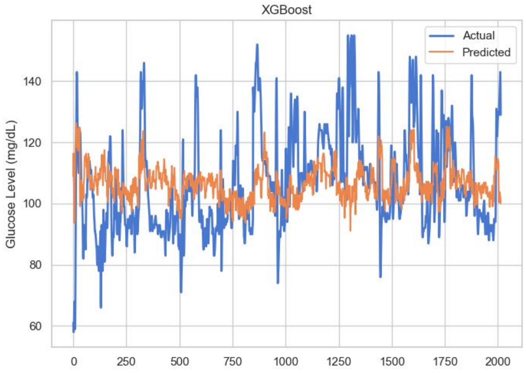
Takeaway:

- XGBoost Model performs best with lowest RMSE
- Further refinement is needed before clinical use

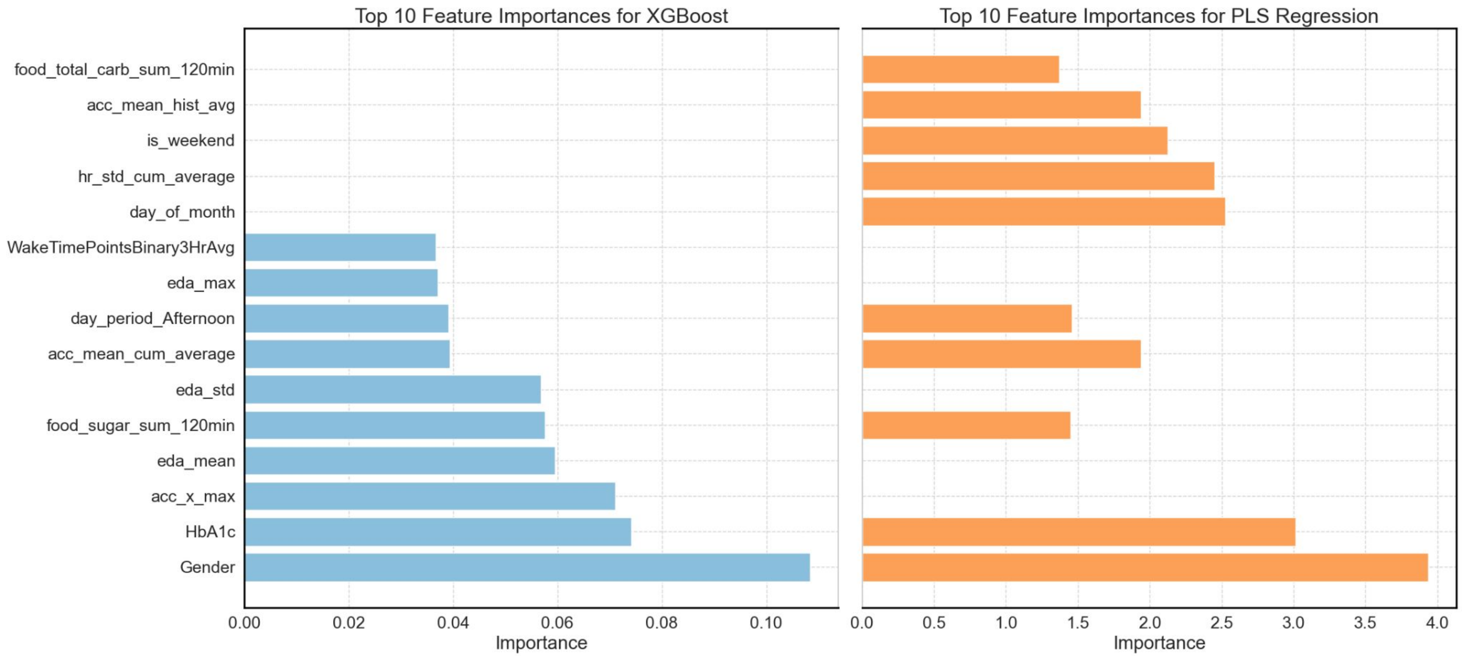
Model	RMSE (mg/dL)	MAPE (%)
Mean Regressor	18.42	15.82
PLS Regression	15.91	11.97
XGBoost	15.84	11.56





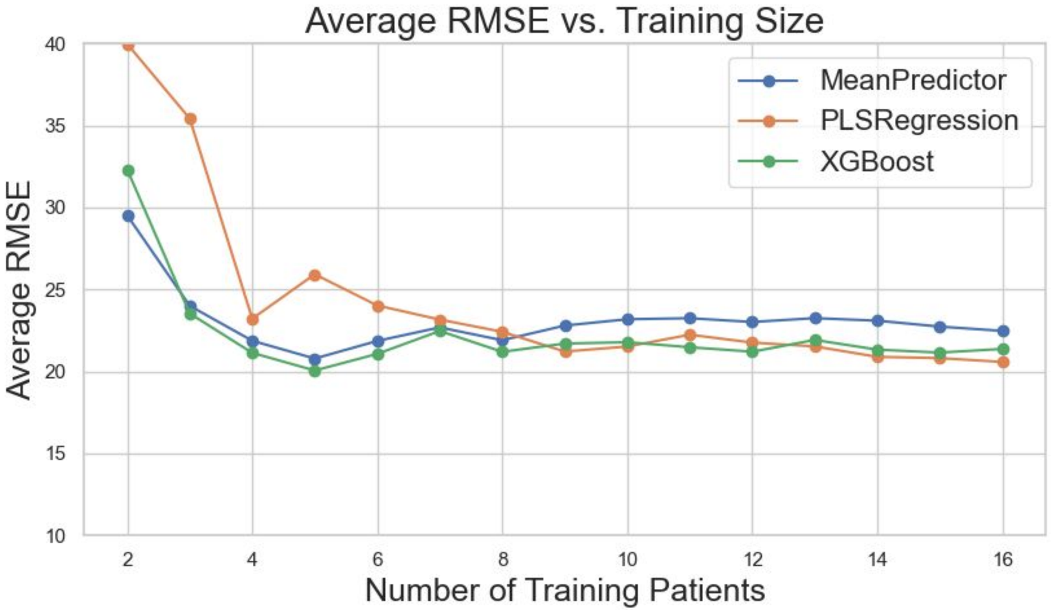


9. Feature Importance



- **Top Features:** Gender and HbA1c as crucial for predicting glucose levels.
- **XGBoost** emphasizes EDA and accelerometer features, while **PLS Regression** focuses on temporal factors and heart rate variability.
- **Food-related metrics are important**, confirming their role in glucose prediction.

10. Model Performance vs. Training Data Size



- Models struggle early on with just a few patients.
- Significant accuracy improvement with 3-5 patients.
- Diminishing returns beyond 8 patients.

11. Conclusions and Future Work

Project Overview and Key Takeaways

1. **Developed a model capable of capturing complex glucose patterns.**
2. **XGBoost** with optimized hyperparameters achieved the best performance: RMSE of 15.84 mg/dL and **MAPE of 11.56%** on the test patient.
3. Despite strong results, the model is **not yet suitable for clinical use.**

Future Work for Model Enhancements:

1. **Expanded Dataset**: Thousands of patients to boost generalization.
2. **Enhanced Food Logs**: Ensure consistent, accurate tracking
3. **Stress & Sleep Data & O2 Levels**: Add wearable/self-reported data on stress, sleep, and O2 levels.
4. **Environmental**: Include temperature and humidity for glucose impact.
5. **Hydration Tracking**: Integrate hydration data into prediction models.
6. **Advanced Features**: Explore new features to capture rapid glucose changes.