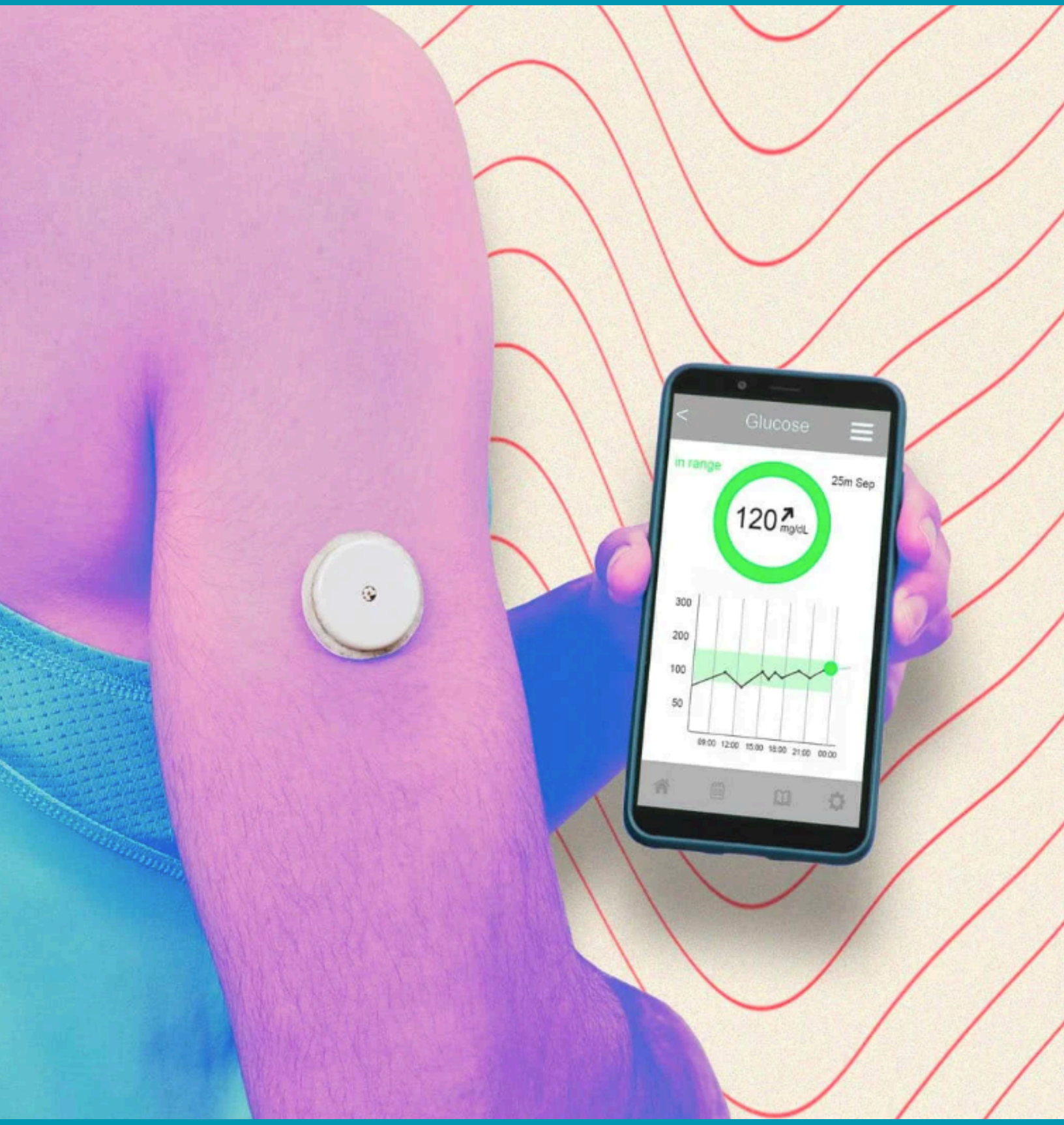
Machine Learning for Noninvasive Glucose Tracking



Prepared by Michael Walker

GlucoPredict - What it is

Prediabetes affects one in three people, with a 10% annual risk of progression to type 2 diabetes without intervention. Effective glycemic monitoring is crucial for prevention; however, no noninvasive, commercially available methods currently exist for self-management.

This study demonstrates the feasibility of using noninvasive methods, such as smartwatches and food logs, to monitor and predict glucose levels. Leveraging a dataset of 25,000 simultaneous glucose and smartwatch measurements, I developed a machine learning model that achieved a 13% Mean Absolute Percent Error in real-time glucose prediction.

Target Audience

1. Healthcare Professionals
2. Researchers: Data scientists and academics.
3. Patients: Individuals managing prediabetes.
4. Engineers: Those working on health devices.
5. Investors: Those interested in health tech.

Dataset

This dataset is downloaded from a study conducted at Duke University, and includes data from **16 participants** monitored over **8-10 days** using **Dexcom G6** and **Empatica E4 devices**. It features over **25,000 interstitial glucose readings**, along with PPG, EDA, skin temperature, heart rate, interbeat interval, and tri-axial accelerometry data, all stored in CSV files. **Food logs** and **demographic details** are also included. The data was collected under ethical approval from the Duke University Health System.

The following 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)

Methods Overview:

This project aims to predict glucose levels using noninvasive wearable data. The process includes:

1. **Data Preprocessing**
   - Dataset: Over 25,000 glucose readings from Dexcom G6 and physiological data from Empatica E4, plus food logs.
   - Tasks: Clean, synchronize, and integrate.
2. **Feature Engineering**
   - Features: Derived from wearables (PPG, EDA, heart rate, accelerometry) and dietary logs.
   - Goal: Convert raw data into inputs for modeling.
3. **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).
4. **Real-Time Glucose Prediction**
   - Application: Provides what would be real-time glucose predictions based on current wearable and food log data.
5. **Evaluation**
   - Validation: Assessed through Leave-One-Group-Out cross-validation (LOGO-CV) and error metrics.

Data Wrangling and Cleaning

In this project, data cleaning and preprocessing were critical for ensuring the accuracy and consistency of the glucose monitoring dataset. Here's a summary of the key steps and challenges addressed:

- **Wearable Data:**
  - Issue: Each patient's csv data required individual sampling period handling and alignment with the 5 minute glucose sampling.
  - Solution: Feature Engineered simple statistics prior to resampling each feature to 5-minute intervals aligned with glucose. Applied universal wrangling code to ensure consistency across all patients and saved the final dataset as patient_df.csv.
- **Food Log Data:**
  - Issue: Initial food log DataFrames required column removals and standardization.
  - Solution: Fixed column names for inconsistencies across different patients. Concatenated all food logs into a single DataFrame, converted columns to appropriate data types, and saved the cleaned data as food_df.csv.
- **Wearable + Food Log + Demographic Data Integration:**
  - Issue: Combining wearables and demographic data required careful alignment.
  - Solution: Imported and preprocessed patient data, set the index to datetime, and encoded categorical data. Merged food log data with patient information, forward-filled missing instances, and adjusted column order for clarity. Saved the final DataFrame as combined_df.csv.

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

Patient CSV Overview:

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

Data Wrangling Special Note

Early Feature Engineering: Maximizing Data Utilization

In this glucose monitoring project, data consists of glucose measurements from Dexcom sampled every 5 minutes, while other features are sampled at much higher frequencies. Simply resampling all features to match the Dexcom intervals would lead to a loss of valuable high-frequency information. To avoid this, early feature engineering was implemented on the original sampling rates before merging dataframes and resampling.

Domain-Specific Features

Based on domain knowledge and supported by relevant publications, the following features were developed to capture intricate details:

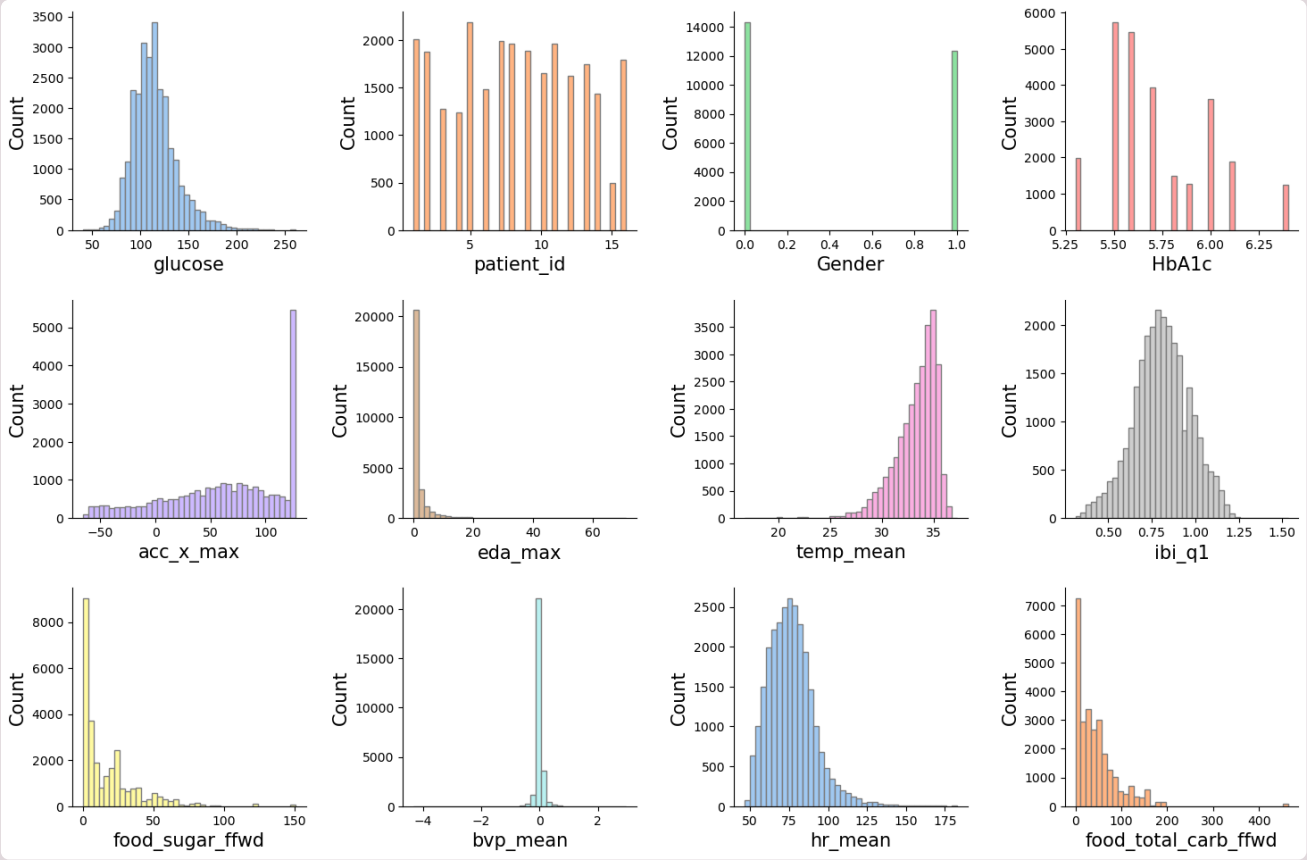
1. Electrodermal Peak Detection: Peaks in electrodermal activity were identified to capture physiological responses.
2. Statistical Summaries (Mean, Std, Min, Max, Q1, Q3, Skew): Key statistical metrics were calculated to summarize the distribution of each feature within the 5-minute intervals.
3. Rolling 2-Hour Mean and 2-Hour Max: Rolling statistics over a 2-hour window were computed to capture temporal trends and fluctuations.

These features were calculated over 5-minute chunks to preserve the granularity of the original data.

Special Note on Data Wrangling

Creating these features before resampling at 5-minute intervals was essential. This approach ensured that intricate details from the higher-frequency data were captured, preventing the loss of critical information during the resampling process.

Exploratory Data Analysis

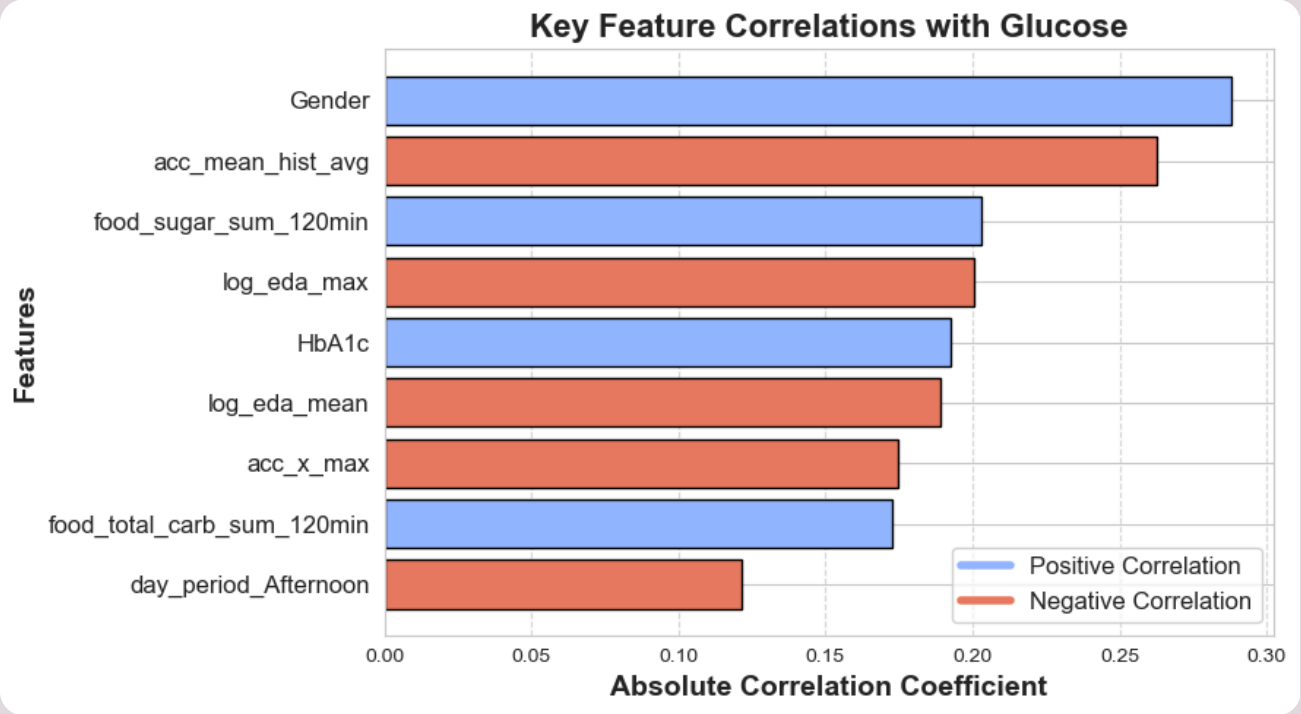


EDA Findings and Commentary:

- **Glucose:** Levels follow a roughly normal distribution, centered between 100-120 mg/dL, with most values ranging from 70-150 mg/dL.
- **Patient ID:** Most patients have around 1,500 samples, with patient 15 notably underrepresented at just 500 samples.
- **acc_x_max:** Maximum x-axis acceleration is right-skewed, with many high values.
- **eda_max:** Electrodermal activity is highly skewed, with most values near zero.
- **temp_mean:** Mean temperature is left skewed mostly around 32-34°C.
- **ibi_q1:** Inter-beat intervals are slightly left-skewed, centered around 0.75 seconds.
- **food_sugar_ffwd:** Food sugar intake is heavily right-skewed, with most values near zero and a long tail.
- **bvp_mean:** Blood volume pulse is tightly centered around zero.
- **hr_mean:** Heart rate is normally distributed around 75-85 bpm.
- **food_total_carb_ffwd:** Total carb intake is also heavily right-skewed, similar to sugar intake.

Feature Engineering

To enhance machine learning models, several advanced features were created to improve glucose analysis. Log transformations were used to better align features with glucose levels. Time-based features were added to capture time since midnight, day of the month, weekend status, and total elapsed time. Categorical features were one-hot encoded to represent different times of day (Night, Morning, Afternoon, Evening). Rolling statistics, including accumulative sums and metrics for dietary intake were calculated. Additionally, rolling sum windows for meal counts, wake time calculations, and activity bout statistics were included to provide a detailed view of the data.



Key Feature Correlations with Glucose

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

Model Survey, Optimization, and Performance Overview

Step 1: Model Survey

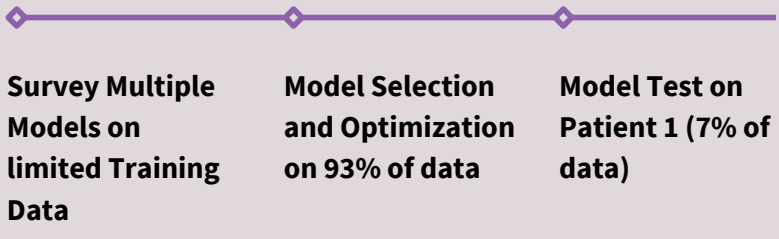
The first phase focused on surveying various models by training them on the initial 25% of data from patients 3-16 and testing on the first 25% of data from patient 2. Models such as Linear Regression, Partial Least Squares (PLS) Regression, Elastic Net, Decision Tree, Random Forest, XGBoost, Support Vector Regression (SVR), and K-Nearest Neighbors were evaluated. This step provided a quick assessment of which models were most effective when applied to a completely unseen patient, helping to narrow down the most promising candidates.

Step 2: Model Selection and Hyperparameter Optimization

Once the promising models were identified, the next phase involved comprehensive hyperparameter optimization using the entire dataset from patients 2-16. Techniques such as RandomizedSearchCV combined with Leave-One-Group-Out Cross-Validation (LOGO CV) were employed to fine-tune models like PLS Regression and XGBoost. This step ensured that the selected models were optimized for maximum predictive accuracy, laying the groundwork for robust glucose level predictions.

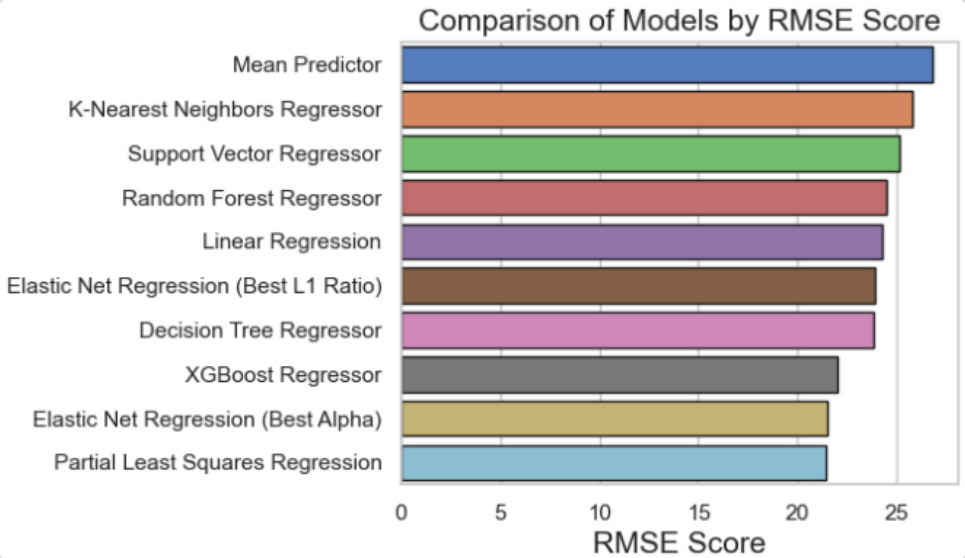
Step 3: Model Performance Testing on Patient 1

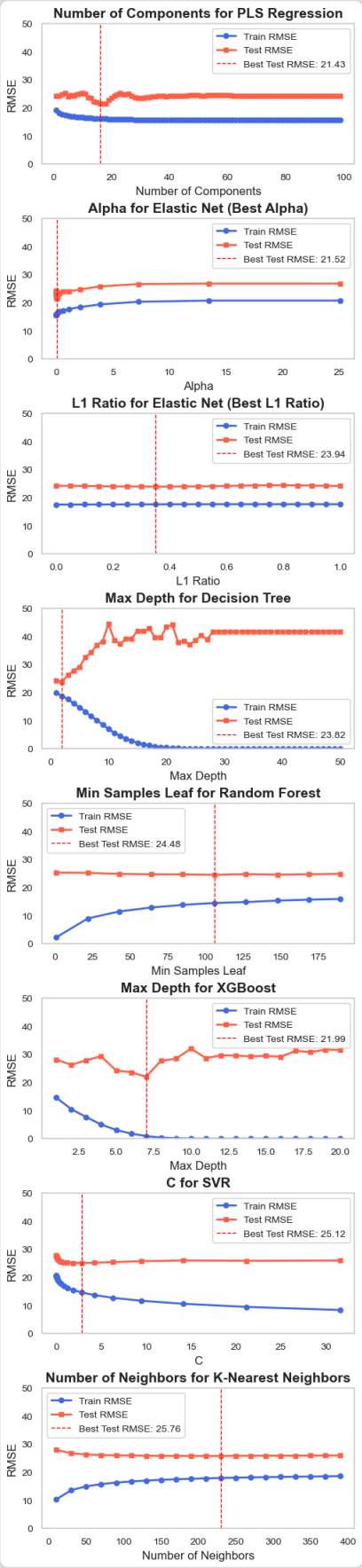
The final step involved testing the optimized models on patient 1, who had been completely excluded from all prior phases. This allowed for an unbiased evaluation of the models' ability to generalize to new, unseen data, providing a clear indication of their real-world performance.



Pipeline & Model Survey

The model pipeline simplifies data preprocessing, training, and evaluation. It starts by **extracting the first 25%** of data for each patient using the get_first_quarter_data() function, with **patients 3-16** used for training and **patient 2's data reserved for testing**. Preprocessing includes **imputing missing values with the mean** and **standardizing features** with StandardScaler. The evaluate_model() function trains the models, makes predictions, and calculates metrics such as **MAE, RMSE, R^2, and MAPE**. Visualization involves comparing true and predicted values with plot_best_results() and plotting RMSE against parameter values with plot_rmse_vs_parameter(). This setup **tests various hyperparameters** (e.g., max depth, number of neighbors, C) for each model respectively, offering a **quick evaluation** on a limited dataset before selecting models for full training.





Model Survey Discussion:

1. **Best Performance:** Models like Partial Least Squares Regression and Elastic Net Regression (Best Alpha) excelled in minimizing error, making them robust choices for handling complex data patterns.
2. **Tree-Based Approaches:** Decision Tree, Random Forest, and XGBoost models show strong performance, with XGBoost leading in capturing intricate relationships effectively.
3. **KNN, SVR, and Mean Predictor:** KNN and Support Vector Regressor (SVR) face challenges with this dataset, highlighting potential areas for further tuning or improvement.

Model Selection and Hyperparameter Optimization

In the final model optimization phase, both PLS Regression and XGBoost models were refined using RandomizedSearchCV with Leave-One-Group-Out Cross-Validation (LOGO CV) to identify the best hyperparameters for accurate glucose prediction. The process began by preparing the data, which included patients 2-16 while excluding patient 1. The pipeline integrated data imputation, scaling, and regression. Hyperparameters such as n_estimators, max_depth, and learning_rate were meticulously tuned to enhance model performance. The results indicated that PLS Regression achieved the best RMSE, suggesting superior generalization capabilities. Visualization efforts included plotting the impact of hyperparameters on RMSE and comparing the performance of various models, including the Mean Regressor, Linear Regression, XGBoost, and PLS Regression, with the final model optimization focusing on the full dataset from patients 2-16 to ensure robust predictive accuracy.

Hyperparameter Optimization:

Partial Least Squares (PLS):

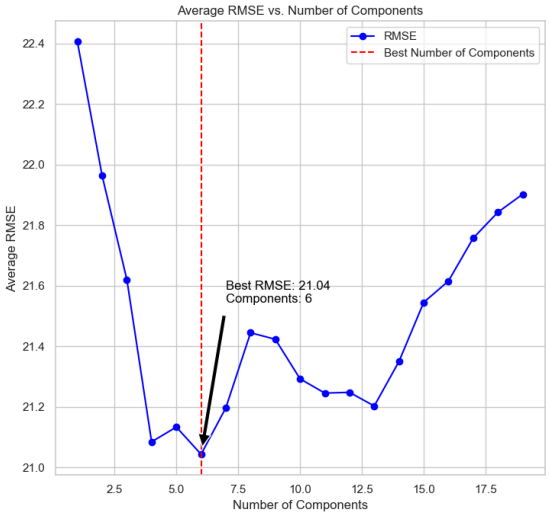
- **Optimized N_Components** = 6
- **RMSE = 21.04**

XGBoost:

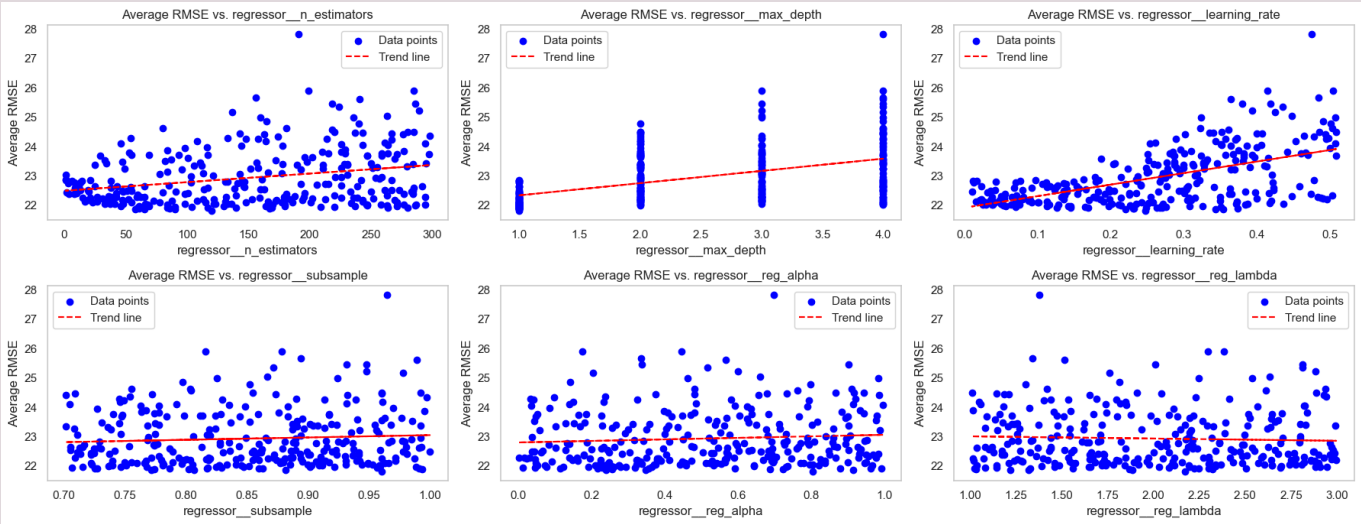
- **Optimized parameters:** learning rate of 0.247, max depth of 1, 120 estimators, reg_alpha of 0.806, reg_lambda of 2.171, and a subsample rate of 0.960.

RMSE = 21.80

PLS N_Components LOGO-CV



XGBoost LOGOCV Random Grid Search



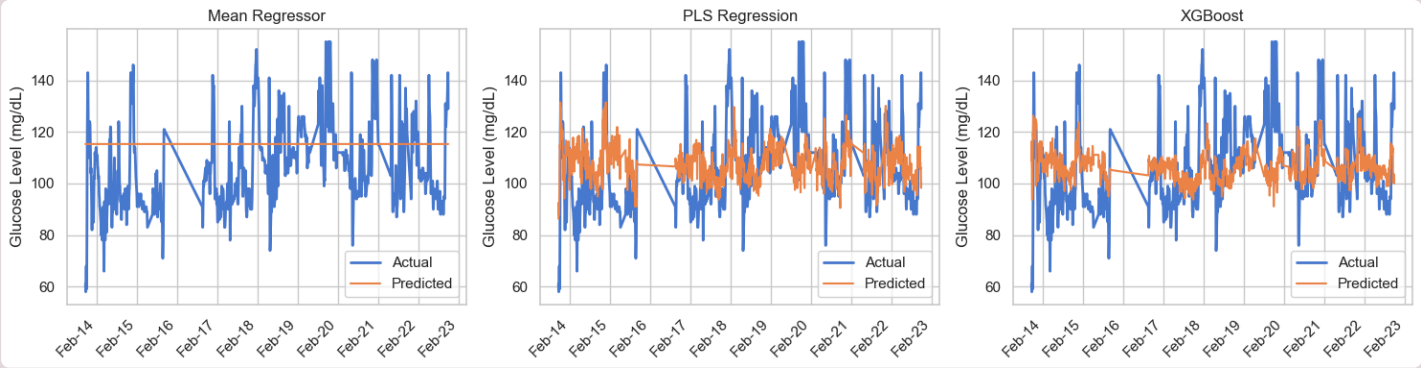
Final Model Performance

In the final stage **optimized models were tested on patient 1**, who had been completely excluded from the training and tuning phases. This crucial step provided an **unbiased assessment** of the models' ability to generalize to new, unseen data, offering a true reflection of their real-world performance. The optimization process had involved training on patients 2-16 using a comprehensive pipeline that included data imputation, scaling, and regression. RandomizedSearchCV with Leave-One-Group-Out Cross-Validation (LOGO CV) was used to fine-tune parameters such as n_estimators, max_depth, and learning_rate.

Final Model Performance:

1. **XGBoost performed best on patient1** for both RMSE and MAPE and PLS scored very similarly.
2. **Both XGBoost and PLS performed significantly better than Mean Regressor**, which is reassuring.

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

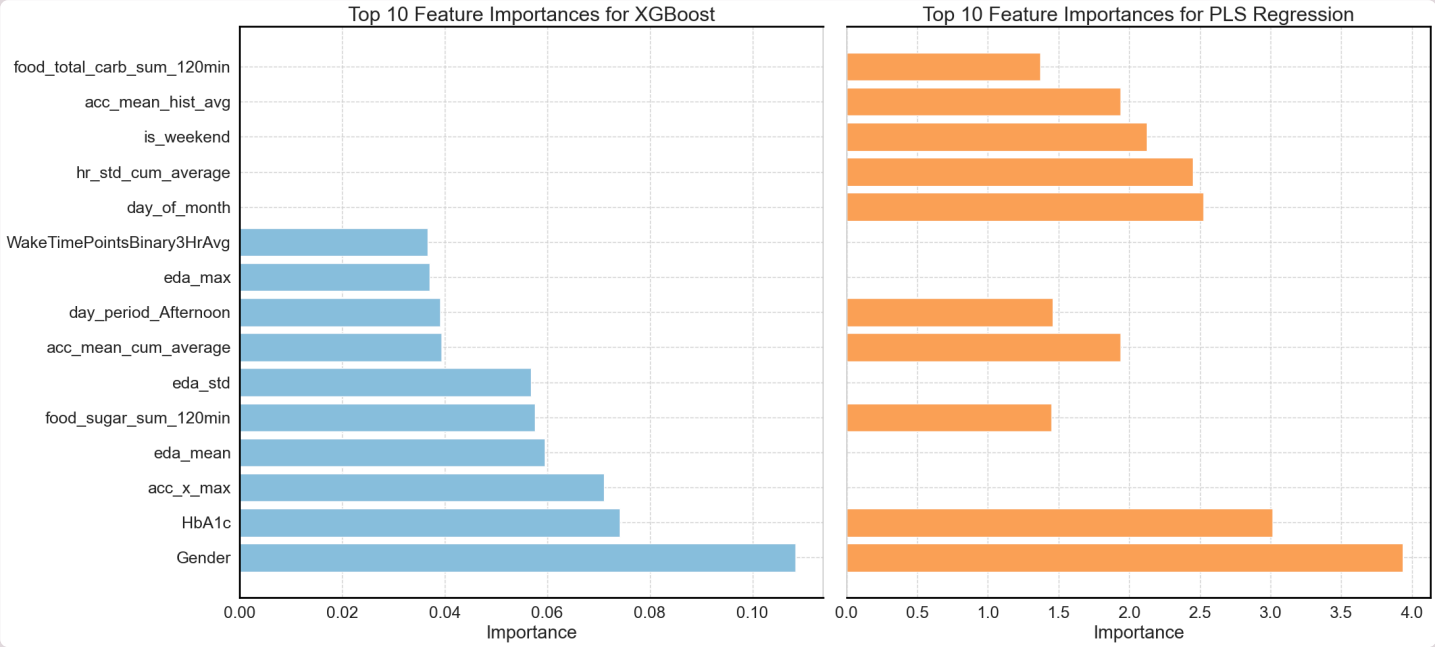


Performance Comentary

1. **Promising Performance:** XGBoost achieves an **RMSE of 15.84 mg/dL** and a **MAPE of 11.56%**, demonstrating its **capability to capture complex glucose patterns** and trends effectively.
2. **Areas for Improvement:** The model successfully **predicts several glucose spikes but misses most**, highlighting its strengths and the need for further refinement.
3. **Key Takeaway:** While the models can sometimes predict high and low glucose values, it's crucial to note that it **should not be used for diagnostic purposes** until further improvements are made.

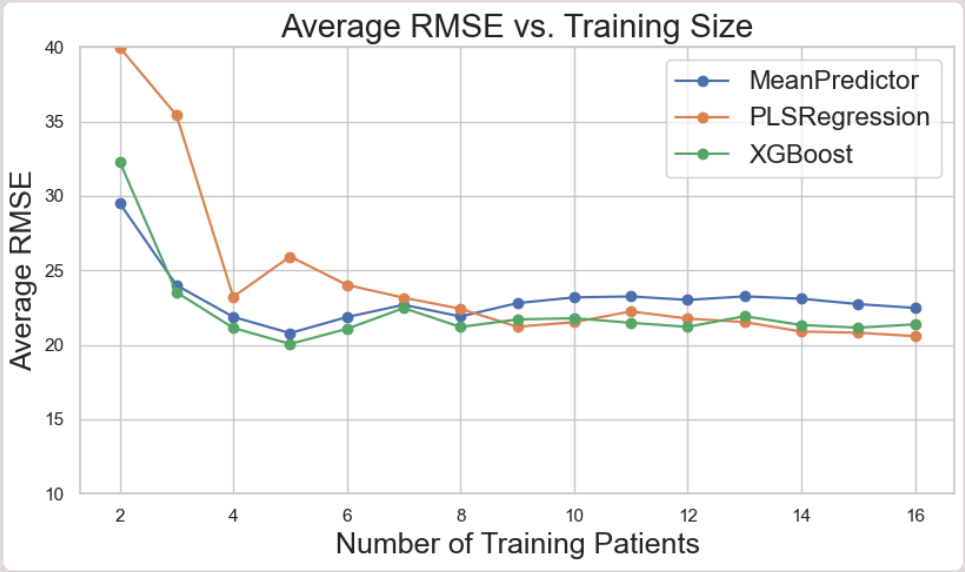
Feature Importance

- **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.



Training Set Size Impact on RMSE

The graph below illustrates how model performance improves as additional patients are included in the training data. Initially, with a small number of patients, the models struggle to make accurate predictions. However, performance significantly improves when the training set grows to include around 8 to 10 patients. Beyond this point, the gains in prediction accuracy become incremental, indicating diminishing returns as more patients are added.



Conclusions and Future Work

The glucose prediction system has demonstrated substantial improvements by employing sophisticated models such as **PLS Regression** and **XGBoost**, **outperforming the simpler mean regression** approach. The effectiveness of these models was notably enhanced through **thoughtful feature engineering**, as illustrated in the feature importance analysis. While adding more training patients initially boosts model accuracy, the benefits diminish beyond 8-10 patients, indicating that our models are efficiently utilizing available data, though additional data has a reduced impact.

Future Work:

1. **Expanded Dataset:** Incorporate data from a broader range of patients, say thousands, to enhance model generalization and accuracy.
2. **Improved Food Logs:** Ensure more accurate and consistent tracking of food intake to better correlate with glucose levels.
3. **Incorporate Stress, Sleep, and Objective Sleep Data:** Add self-reported and wearable data on stress and sleep, including data from devices like Apple Watch for sleep schedules and O2 measurements.
4. **Include Environmental Factors:** Consider variables such as temperature and humidity that may indicate responses to glucose levels.
5. **Track Hydration Levels:** Monitor and incorporate hydration data into the prediction model.
6. **Advanced Features:** Continue exploring additional features to capture more complex relationships to rapid changes in glucose levels.