Using Machine Learning to Evaluate the Impact of Food Feature Intake on Blood Glucose Response: mmol/L

Jon Golding
Toronto, ON, Canada
jgolding.contact@gmail.com

ABSTRACT

Diabetes mellitus is a highly heterogeneous condition where blood glucose levels are often observed in an elevated state potentially leading to long-term health impacts for the patient. For this reason, diabetes is a condition that requires carefully evaluated treatment plans that are patient-specific, in addition to regular patient monitoring. The rise of Machine Learning offers a promising avenue to evaluate the environmental factors that can influence blood glucose response at a personalized patient level.

In recent years, diabetes blood glucose monitoring has seen technological advancements to include continuous blood glucose monitoring (CGM) via sensor inserted under a user's skin, often placed on the back of the arm, or the stomach region.

While proprietary phone applications developed by CGM manufacturers have the ability to collect the CGM sensor data and subsequently map descriptive statistics over time such as average blood glucose in mmol/L, or percent time-in-range, there still does not exist a tool that allows diabetic CGM-users access to Machine Learning insights in order to evaluate how certain food choices may (or may not) influence directional changes in blood glucose levels at a patient-specific level.

The objective of this study is to determine if machine learning can be effectively used to evaluate different food features consumed as they relate to predicting change in blood glucose mmol/L levels over time.

9367 time-stamped instances of data were collected from the Author's personal CBM from March 10, 2023 through to July 2, 2023. Over 400 different food categories were initially logged using text-based notes via sensor-paired phone app for each timestamped instance where food intake was logged. The dataset food features were reduced to the top 20 features of most interest to keep the number of feature columns manageable and to allow space for further time-based feature engineering. The overall strategy of reducing the initial set of food features to those of most interest was done to reduce likelihood of high model training times occurring due to high cardinality.

BIASES, ASSUMPTIONS & MODEL LIMITATIONS :

1. As the author's own personal data was gathered and used for this study, the author's domain knowledge is informed and potentially influenced through his own observations and perceptions of how different food types influenced his glucose response.
2. During the data gathering process, food intake for each food category was not defined in terms of the amount consumed for each food feature due to the logistical challenges presented in accurately recording food amounts consumed. This presents a narrower scope of how food intake is defined, potentially leading to higher variability in model results. Food features were one-hot-encoded and represented by 1 for intake present at a given timestamp, and 0 for intake not-present at a given timestamp.
3. It should be noted that a significant challenge arises when using machine learning to work with time series data: machine learning strategies often shuffle data instances or draw inferences in ways that may not preserve or capture the temporal dependencies between ordered time steps. This can lead to feature leakage where past data points are trained on information extracted from future data points leading to inaccurate modeling and feature importance conclusions. The methods used in this study take careful measures to capture temporal dependencies between ordered timesteps.
4. Stratified sampling was used for the correlation funnel steps of the exploratory data analysis to focus on time windows of specific food features consumed. It was later decided not to pursue the stratified sampling outside of EDA during modeling processes due to:
    - Sample sizes that were effectively too small for use in reliable machine learning models.
    - Concerns about preserving the temporal order of timestamped data instances and not introducing significant gaps between data points.

ML MODELING STRATEGIES :

Two different machine learning models were applied to compare the utility of model-based feature importance:

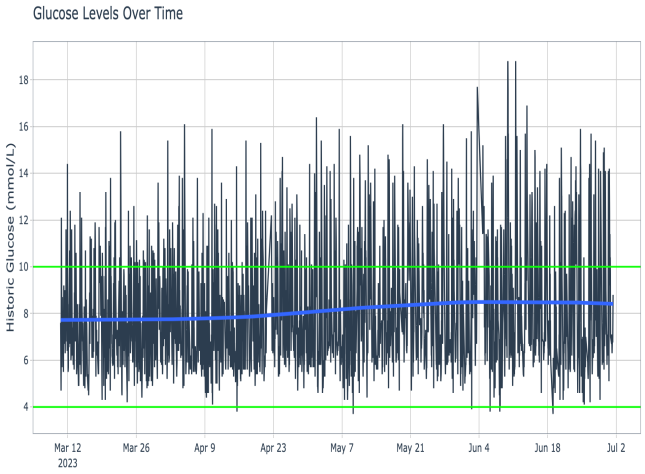
1. Linear regression was selected as the first strategy used to model the data in this study. The linear regression model was chosen given its inherent model coefficients which can be used to plot feature importance for non-agnostic model

explainability. While the utility of directional coefficients is highly desirable to suitable for the specific use-case of this study, one trade-off is that linear regression algorithm is limited in its ability to account for the variance caused by non-linear problems such as non-linearity caused by possible food feature interactions.

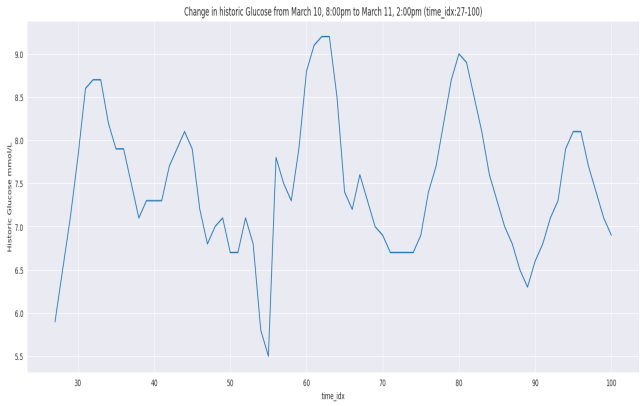
2. XGBOOST Regressor: [1] A decision-tree-based machine learning algorithm which can be well-adapted for the use-case of modelling multivariate time-series data. XGBOOST uses gradient boosted trees to make y-variable predictions through combining individual decision trees to form a strong learner.[3] The XGBOOST algorithm is well-suited at handling multivariate datasets, as well as non-linear problems. One disadvantage of XGBOOST is that is lacks the model coefficients inherent to the linear regression model which can be helpful in determining the linearly derived statistical significance level of each model regressor. For plotting global feature importance, the XGBOOST algorithm integrates well with SHAP feature explainability tools. SHAP (Shapley Additive Explanations) shows the directional impact that each feature contributes to a given model's overall predictions.

EXPLORATORY DATA ANALYSIS:

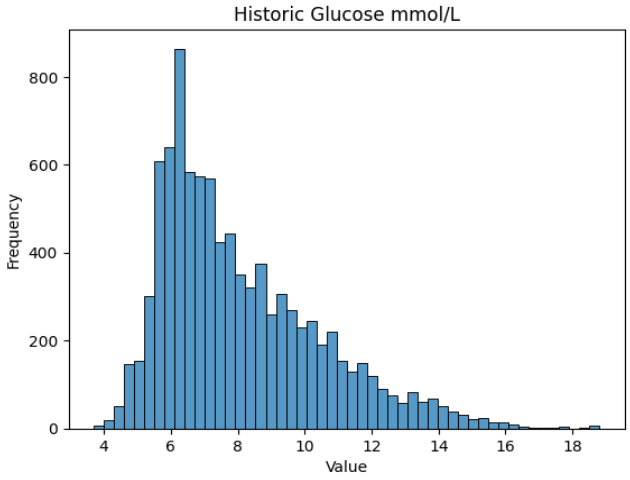
Historic blood glucose mmol/L is plotted as a function of time over a 3-month span showing high volatility. The upper and lower bounds of a normative glucose range were defined by the green y axis intercepts:



In order to capture further context of target variable change within a shorter time window, a 30-hour sample of recorded time stamp data was plotted below showing Historic Blood Glucose mmol/L as a function of time:



Shown below is a histogram of the target variable data distribution. Given the right skew, a log transform was later applied during data preprocessing to normalize the distribution:



CORRELATION FUNNEL FOR PRELIMINARY EDA:

To gain some intuition about the potential strength of the relationships between predictor variables (primarily food features) and the target variable 'Historic Glucose mmol/L', a [42]correlation funnel technique was used for the purposes of exploratory data analysis.

*It should be noted that the correlation funnel process used in this study, was carried out for the purposes of preliminary EDA as a means of gathering preliminary insights and was not used to inform final model-based results or model-based conclusions.

[1] Adapted from Grogan, M. (2021, September 8). *XGBoost for Time Series Forecasting: Don't Use it Blindly.* Towards Data Science.

[2]Adapted from Danso, M. Business-science. *Correlationfunnel.* Business-science.github.io/correlationfunnel/

[53]The correlation funnel technique, invented by Data Scientist Matt Danso, uses a binning process to convert a dataset's features into binary format so that each feature can be then correlated with binned ranges of the continuous target variable.

Note that binarize function found in the imported R library `correlationfunnel`, assigns bin ranges to continuous variables, and the resulting ranges are one-hot-encoded into binary values. In the case of the target variable, the following 4 bin ranges were created:

`Historic_Glucose_mmol_L__-Inf_6.3

'Historic_Glucose_mmol_L__6.3_7.4'

'Historic_Glucose_mmol_L__7.4_9.5'

'Historic_Glucose_mmol_L__9.5_Inf'

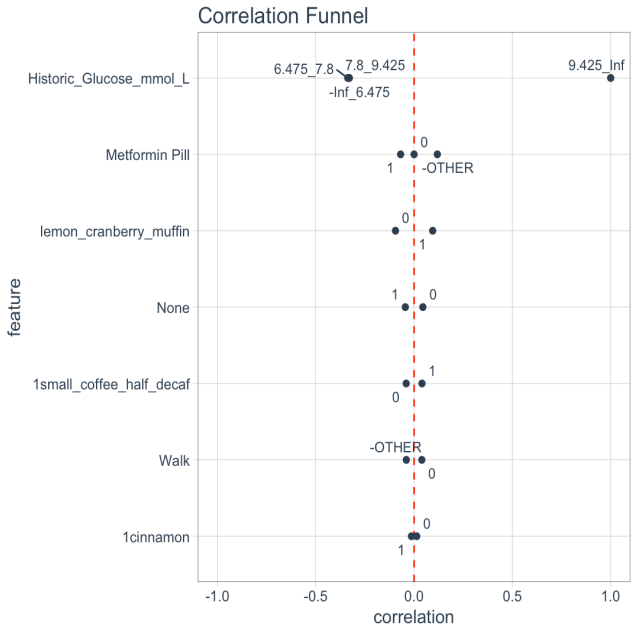
To illustrate the result of this process, shown below is a correlation funnel plot for the highest bin range of the target variable 'Historic_Glucose_mmol_L__9.5_Inf':

As shown, the correlation funnel for the high glucose bin of 9.5 to infinity shows no correlation for any food feature with the target variable range beyond random noise. This observation in is useful when thinking about static glucose readings at the time of food intake: Glucose response takes time to show an effect post food-intake. The addition of future-based lag feature columns (lead columns) could potentially be useful during the modeling process in capturing time-delayed patterns in blood glucose response. The lack of observed correlations between food features and the target variable may also be due to the sparseness of the data points for food features consumed in the data set. Taking the lemon cranberry muffin food feature as an example, the food intake is not constant and quite sparse as it was consumed 14 times across the dataset of 9367 instances.

[3]Adapted from Danso, M. Business-science. *Correlationfunnel.* Business-science.github.io/correlationfunnel/

Next, a strategy of taking a stratified sample containing 14 four-hour time windows post lemon cranberry muffin consumption was used.

Shown below is the resulting correlation funnel plot for the stratified muffin sample data-frame which was binarized and correlated to the high glucose bin range:



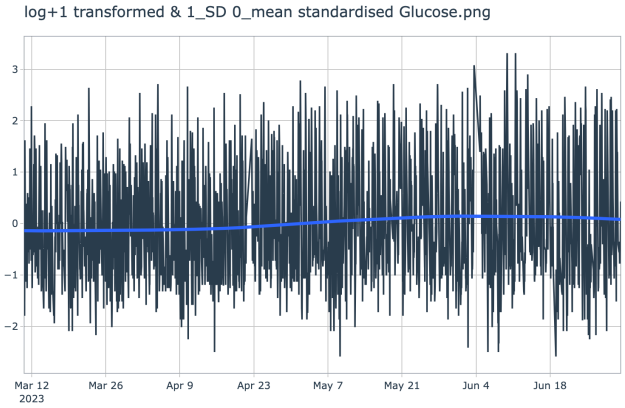
DATA PREPARATION:

- Creating a pre-processing pipeline for the linear regression model:

Given the non-linear nature of the dataset, it will be worth exploring feature engineering, and non-linear transformations to see if the data can modeled in a way that will allow for feature-based insights from the model coefficients.

A log transformation was applied to the target variable to reduce the effect of outliers and to reduce overall variance. [64]The benefit of applying a log transformation is that it can linearize the relationship between variables in a non-linear dataset, making it more suitable for linear regression. A log transformation will also stabilize the variance, reducing heteroscedasticity across residuals which is a key assumption that needs to be met for linear regression.

The resulting log transformed glucose target was then standardized using mean 0 and a standard deviation of 1 to ensure that the scale of the target variable was in alignment with the scale of the regressor variables:

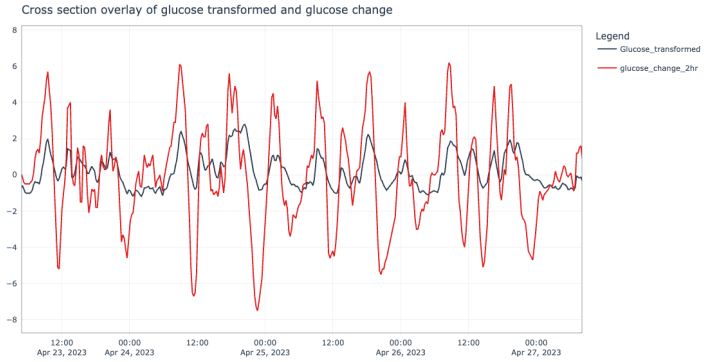


FEATURE ENGINEERING:

- Features representing glucose change windows:

4 features were next extracted from the original glucose target variable to be used as external regressors during the modeling process: *glucose_change_1hr*, *glucose_change_2hr*, *glucose_change_3hr*, and *glucose_change_4hr*. The use of these hourly window-based change features became crucial in developing models with higher variance explained.

As an example, shown below is a plot overlay of the log-transformed-standardized target glucose variable with the 2-hour glucose change column:



As evident from viewing the plot above, the glucose change predictor variables also required scale standardization and were subsequently standardized in the data preprocessing pipeline using mean 0 and standard deviation 1.

- Time-based feature engineering:

The following time-based features were extracted from the device timestamp column using the *tk_augment_timeseries* function from the *timetk* library in R:

[4]Adapted from (2024, September) *ChatGPT 4o.* chatgpt.com.

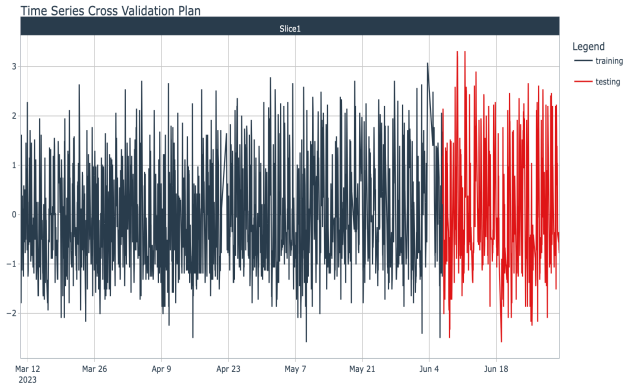
day, hour, hour12, wday, mday, qday,mweek,week, week2, week3, week4, mday7.

The time-based features were subsequently standardized in the preprocessing data pipeline using mean 0 and standard deviation 1.

- Lag-based feature engineering:

13 lag-based features of the target variable *Glucose_transformed* were created and spaced at 15 minute intervals across a 3 hour time-window. The lag features were engineered to help the model better account for the time-delay seen in blood glucose response post food intake, and to learn these resulting patterns during model training.

With the data preprocessing and feature engineering steps completed, a time series cross-validation plan was created and plotted using an 80-20 split where 80 percent of the collected glucose data was allocated for model training, and the remaining 20 percent was allocated as a validation set for testing model performance on unseen data. It should be noted that the temporal order of timestamped data instances was strictly maintained during the train-test split process:



MODEL EVALUATION:

Two iterations of the linear regression model were trained and fine-tuned using the `tidymodels` ecosystem (package) in R which uses the OLS (Ordinary Least Squares) optimization function to estimate the coefficients for the Linear Regression models.

The end goal was to find the best possible trade-off between r-squared (variance explained) and food-based feature explainability obtained from the linear model coefficients.

As observed for the first iteration of the linear model, the r-squared value of .975 for the validation set was slightly lower than the r-squared of .98 observed on the training data showing virtually no overfitting. Variance explained was robust for the linear model demonstrating the power of feature engineering derived from the time-based regressors, glucose-change-over-time regressors, glucose lag regressors, and food-based regressors.

*This variance explained result suggests that the model can effectively predict glucose changes on unseen data and is not specifically a measure of how reliably each feature contributes to glucose change. To gain better intuition about food feature impact on glucose response, a summary table of feature coefficients and their possible significant levels is provided below in addition to model-based feature importance plots.

Validation set R-squared:

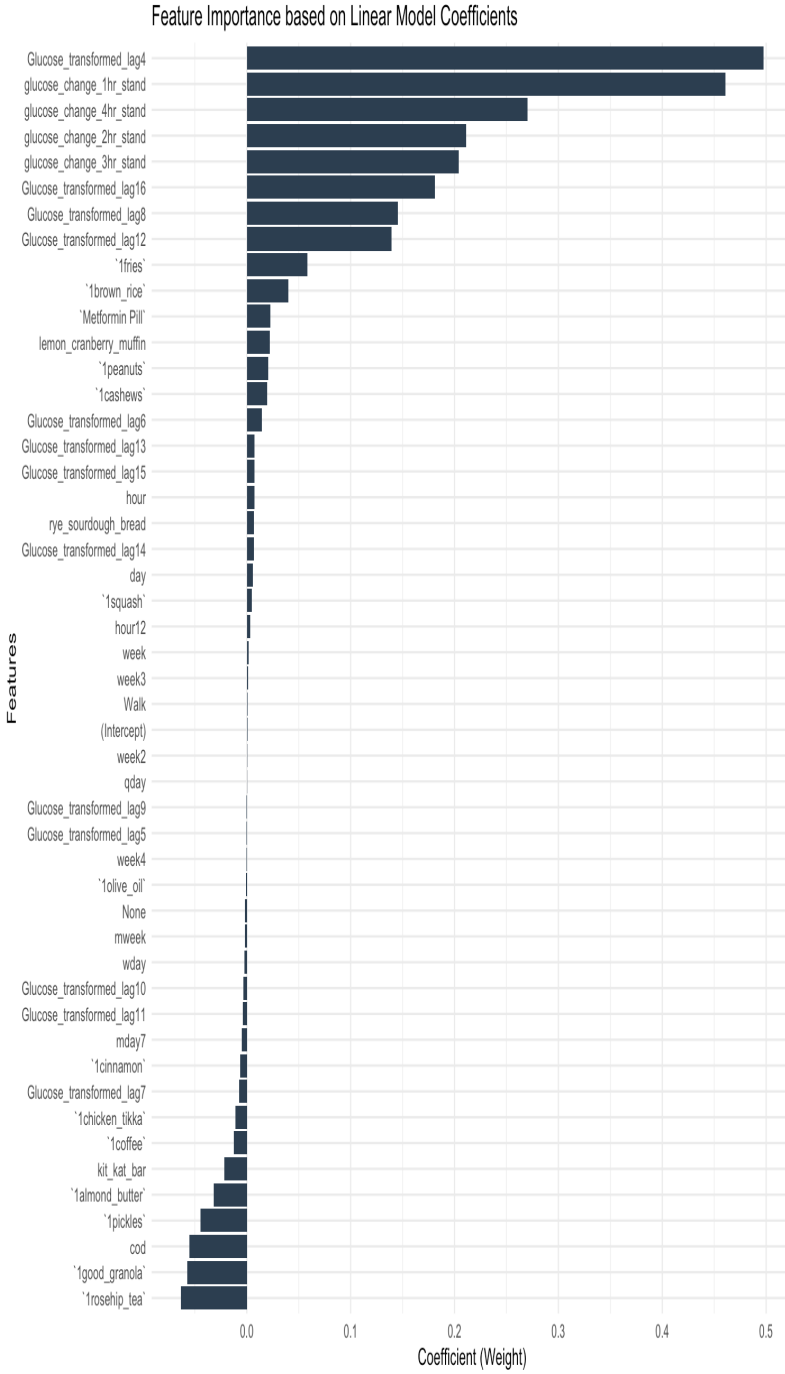
```
# A tibble: 1 x 9
  .model_id .model_desc .type   mae  mape  mase smape  rmse   rsq
      <int> <chr>       <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1         1 LM          Test  0.134  36.4 0.505  25.7 0.190 0.975
```

Training set model coefficients and significance levels:

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.76455 -0.06461  0.01355  0.07741  0.54232

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.0008525  0.0040739   0.209 0.834254
None                    -0.0016014  0.0044175  -0.363 0.716973
Walk                     0.0010027  0.0005847   1.715 0.086388 .
`1cinnamon`             -0.0065265  0.0164241  -0.397 0.691103
`Metformin Pill`         0.0224262  0.0120992   1.854 0.063845 .
`1rosehip_tea`          -0.0636436  0.0335326  -1.898 0.057740 .
`1almond_butter`        -0.0316311  0.0196775  -1.607 0.107992
`1good_granola`         -0.0574111  0.0473205  -1.213 0.225076
`1peanuts`               0.0205687  0.0293685   0.700 0.483721
`1squash`                0.0048971  0.0403394   0.121 0.903380
`1pickles`              -0.0445112  0.0386678  -1.151 0.249720
lemon_cranberry_muffin   0.0219205  0.0371794   0.590 0.555486
kit_kat_bar             -0.0215381  0.0545867  -0.395 0.693174
`1coffee`               -0.0125342  0.0325962  -0.385 0.700598
`1olive_oil`            -0.0008525  0.0179564  -0.047 0.962136
`1fries`                 0.0584114  0.0403461   1.448 0.147727
`1chicken_tikka`        -0.0111376  0.0545483  -0.204 0.838219
`1cashews`               0.0198397  0.0423278   0.469 0.639288
`1brown_rice`            0.0398823  0.0235483   1.694 0.090376 .
cod                     -0.0555365  0.0666989  -0.833 0.405072
rye_sourdough_bread      0.0070659  0.0403833   0.175 0.861107
```

Training set model coefficients and significance levels continued:

```
glucose_change_1hr_stand   0.4610304  0.0109657  42.043  < 2e-16 ***
glucose_change_2hr_stand   0.2109294  0.0166170  12.694  < 2e-16 ***
glucose_change_3hr_stand   0.2038008  0.0177209  11.501  < 2e-16 ***
glucose_change_4hr_stand   0.2700549  0.0158629  17.024  < 2e-16 ***
day                        0.0057831  0.0076070   0.760 0.447138
hour                       0.0073383  0.0019804   3.705 0.000213 ***
hour12                     0.0034155  0.0018574   1.839 0.065978 .
wday                      -0.0021831  0.0015661  -1.394 0.163355
qday                       0.0001937  0.0017784   0.109 0.913278
mweek                     -0.0016809  0.0024789  -0.678 0.497744
week                       0.0015813  0.0016349   0.967 0.333465
week2                      0.0003077  0.0020907   0.147 0.883016
week3                      0.0010175  0.0015596   0.652 0.514142
week4                     -0.0003854  0.0023653  -0.163 0.870582
mday7                     -0.0046791  0.0074704  -0.626 0.531101
Glucose_transformed_lag4   0.4976568  0.0125254  39.732  < 2e-16 ***
Glucose_transformed_lag5  -0.0002203  0.0110399  -0.020 0.984080
Glucose_transformed_lag6   0.0144749  0.0118598   1.220 0.222316
Glucose_transformed_lag7  -0.0075023  0.0119371  -0.628 0.529702
Glucose_transformed_lag8   0.1454925  0.0168531   8.633  < 2e-16 ***
Glucose_transformed_lag9  -0.0001782  0.0119527  -0.015 0.988108
Glucose_transformed_lag10 -0.0033718  0.0119361  -0.282 0.777581
Glucose_transformed_lag11 -0.0037441  0.0119366  -0.314 0.753782
Glucose_transformed_lag12  0.1395479  0.0168628   8.275  < 2e-16 ***
Glucose_transformed_lag13  0.0074645  0.0119436   0.625 0.532006
Glucose_transformed_lag14  0.0067738  0.0118584   0.571 0.567864
Glucose_transformed_lag15  0.0073410  0.0108728   0.675 0.499587
Glucose_transformed_lag16  0.1810577  0.0128308  14.111  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.133 on 7432 degrees of freedom
Multiple R-squared:  0.9805,    Adjusted R-squared:  0.9804
F-statistic:  7780 on 48 and 7432 DF,  p-value: < 2.2e-16
```



The feature importance plot for the first linear model iteration showed some initial insights about the possible directional contribution of different food features when explaining the target response variable ‘Glucose_transformed’. The target lag-based regressor ‘glucose_change_lag_4’ and ‘glucose_change_1hr_stand’ feature as shown were heavily weighted in terms of model contribution impact:

Interpretation of Initial Linear Regression Model Results:

The following observations were specific to the Author's own personal blood-glucose response and are subject to the previously mentioned model limitations and study biases:

The resulting feature plots and food feature coefficients yielded interesting results which were mostly congruent with the author's expectations from personal domain knowledge and observations over time. Some findings were also unexpected:

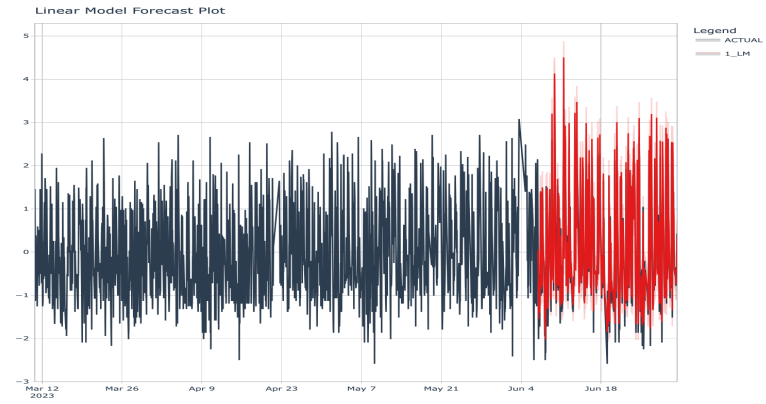
It was anticipated through personal observations that the `Fries` food feature would contribute to an upward movement in glucose response over a window of several hours after each point of food consumption for this feature. While not indicated as a statistically significant feature at this time based on its model coefficient, the model feature importance plot showed the `fries` variable as the most important food feature in terms of food feature model impact.

It was very surprising that the `kit_kat_bar` feature demonstrated a slight negative directional model contribution for glucose response (as seen, previous feature importance plot) given its obvious glucose content. Putting this observation into context with the author's domain knowledge, the `kit_kat_bar` feature was consumed as few as 8 times across the 3-month time series dataset and was usually consumed at times of low glucose levels. This could be a telling aspect of the possible limitations of the linear model in separating periods of extreme lows (or highs) in glucose levels from food-based feature impacts.

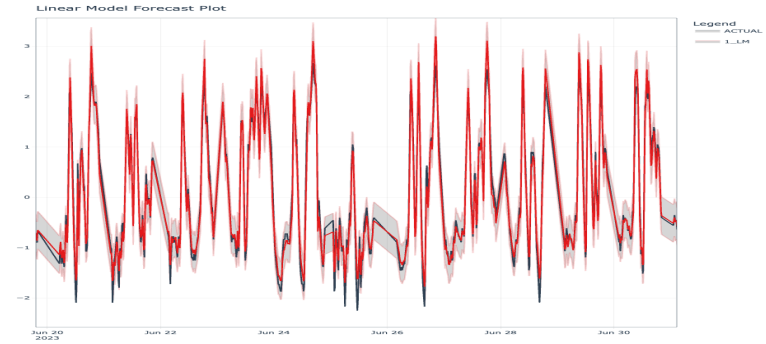
The `metformin_pill` feature impact and positive coefficient significance level was surprising for similar reasons. It was expected that the medication would demonstrate a model-based feature impact toward lower glucose response, but instead showed an impact in the opposite direction. Using domain knowledge context, the `metformin_pill` feature was almost always taken in the morning with a first daily meal (both the medication and first meal taken on an empty stomach after waking up) and was preceded by a large glucose spike each time.

Given the previously mentioned domain knowledge and the authors personal observations over time: `coffee`, `almond_butter`, `olive_oil`, `cod`, and `cinnamon` features were all expected to have neutral or slight downward feature importance contributions on glucose response. This was shown to be the case as shown in the model-based feature importance plot.

To gain a visualization of model forecast estimates, shown below is an overlay of predicted vs actual glucose levels for the first linear model iteration:



A zoomed view of the linear model forecast window:



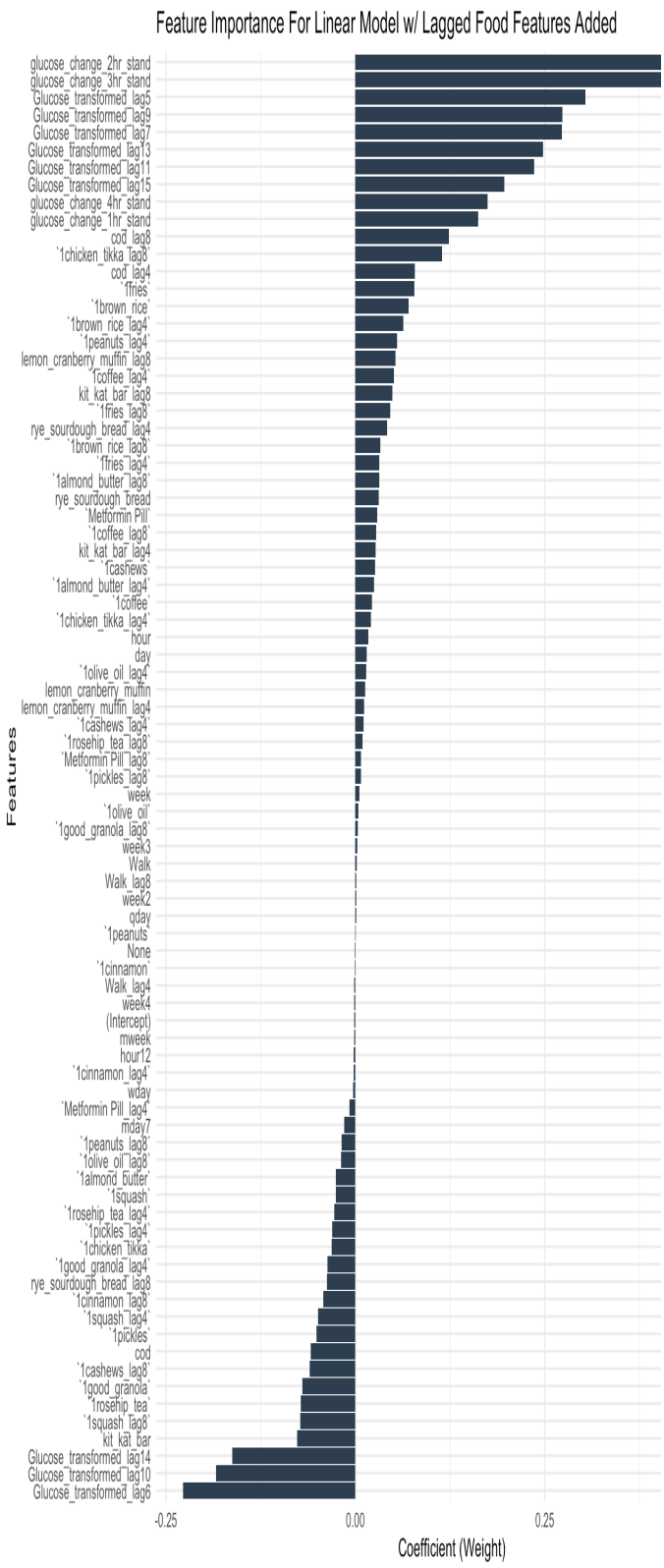
2nd linear regression model iteration:

For the second linear model iteration, lagged food features were engineered at 1 hour and 2 hour lagged time windows for each respective food feature. The goal of introducing the lagged food features was not to affect an already robust r-squared value but to provide more insights about the possible contribution of food features over time. Below, the validation set r-squared remains largely unchanged from the first model iteration at .962.

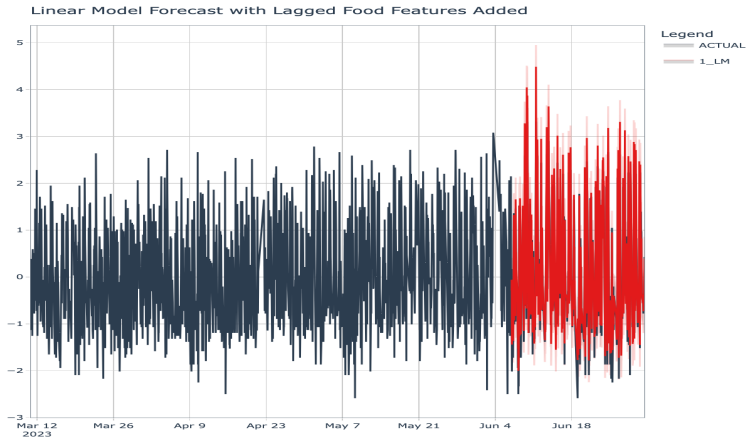
```
A tibble: 1 × 9
.model_id .model_desc .type   mae  mape  mase smape  rmse   rsq
    <int> <chr>       <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
        1 LM          Test  0.169  51.6 0.636  31.0 0.236 0.962
```

Training set model coefficients and significance levels, 2nd linear model iteration:

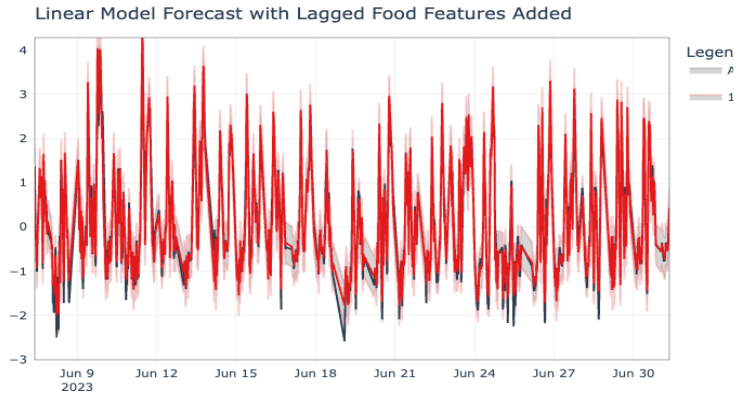
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.83854 -0.08951  0.01027  0.09815  1.15694

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -0.0017182  0.0050744  -0.339  0.73492
None                        -0.0011073  0.0054642  -0.203  0.83942
Walk                         0.0020552  0.0007165   2.868  0.00414 **
`1cinnamon`                 -0.0011480  0.0201766  -0.057  0.95463
`Metformin Pill`             0.0289316  0.0148840   1.944  0.05196 .
`1rosehip_tea`              -0.0721311  0.0418378  -1.724  0.08474 .
`1almond_butter`            -0.0259974  0.0241732  -1.075  0.28220
`1good_granola`             -0.0697367  0.0580078  -1.202  0.22933
`1peanuts`                  -0.0006042  0.0360883  -0.017  0.98664
`1squash`                   -0.0261725  0.0496822  -0.527  0.59835
`1pickles`                  -0.0516369  0.0474821  -1.088  0.27685
lemon_cranberry_muffin       0.0127353  0.0456347   0.279  0.78020
kit_kat_bar                 -0.0770931  0.0669320  -1.152  0.24944
`1coffee`                    0.0220060  0.0400040   0.550  0.58227
`1olive_oil`                 0.0036823  0.0222081   0.166  0.86831
`1fries`                     0.0774413  0.0494642   1.566  0.11748
`1chicken_tikka`            -0.0315763  0.0731857  -0.431  0.66615
`1cashews`                   0.0257696  0.0521460   0.494  0.62119
`1brown_rice`                0.0704415  0.0290368   2.426  0.01529 *
cod                         -0.0587823  0.0822638  -0.715  0.47490
rye_sourdough_bread          0.0309986  0.0495131   0.626  0.53129
glucose_change_1hr_stand     0.1620045  0.0066366  24.411  < 2e-16 ***
glucose_change_2hr_stand     0.4819341  0.0122069  39.480  < 2e-16 ***
glucose_change_3hr_stand     0.4604768  0.0128906  35.722  < 2e-16 ***
glucose_change_4hr_stand     0.1742104  0.0086556  20.127  < 2e-16 ***
day                          0.0151942  0.0093386   1.627  0.10377
hour                         0.0168559  0.0024307   6.934 4.42e-12 ***
hour12                      -0.0020964  0.0022893  -0.916  0.35982
wday                        -0.0029647  0.0019241  -1.541  0.12341
qday                         0.0008285  0.0021882   0.379  0.70497
mweek                       -0.0019218  0.0030434  -0.631  0.52776
week                         0.0051938  0.0020215   2.569  0.01021 *
week2                        0.0011823  0.0025697   0.460  0.64548
week3                        0.0024175  0.0019169   1.261  0.20730
week4                       -0.0015153  0.0029059  -0.521  0.60206
mday7                       -0.0148659  0.0091719  -1.621  0.10510
Glucose_transformed_lag5     0.3039734  0.0119569  25.422  < 2e-16 ***
Glucose_transformed_lag6    -0.2273695  0.0135149 -16.824  < 2e-16 ***
Glucose_transformed_lag7     0.2723879  0.0117469  23.188  < 2e-16 ***
Glucose_transformed_lag9     0.2729584  0.0118936  22.950  < 2e-16 ***
Glucose_transformed_lag10   -0.1839736  0.0135393 -13.588  < 2e-16 ***
Glucose_transformed_lag11    0.2359484  0.0117398  20.098  < 2e-16 ***
Glucose_transformed_lag13    0.2478383  0.0118962  20.833  < 2e-16 ***
Glucose_transformed_lag14   -0.1628049  0.0135040 -12.056  < 2e-16 ***
Glucose_transformed_lag15    0.1965436  0.0117309  16.754  < 2e-16 ***
`1fries_lag4`                0.0317104  0.0492501   0.644  0.51968
`1fries_lag8`                0.0459401  0.0492704   0.932  0.35116
kit_kat_bar_lag4             0.0263325  0.0667794   0.394  0.69336
kit_kat_bar_lag8             0.0489353  0.0669724   0.731  0.46500
lemon_cranberry_muffin_lag4  0.0113567  0.0453702   0.250  0.80235
lemon_cranberry_muffin_lag8  0.0530440  0.0453366   1.170  0.24204
`1peanuts_lag4`              0.0552976  0.0358201   1.544  0.12269
`1peanuts_lag8`             -0.0183965  0.0360364  -0.510  0.60972
`1cashews_lag4`              0.0105469  0.0517650   0.204  0.83856
`1cashews_lag8`             -0.0601349  0.0522523  -1.151  0.24983
`1cinnamon_lag4`            -0.0021615  0.0196650  -0.110  0.91248
`1cinnamon_lag8`            -0.0421160  0.0196620  -2.142  0.03223 *
cod_lag4                     0.0787979  0.0818170   0.963  0.33553
cod_lag8                     0.1230788  0.0816796   1.507  0.13189
`1good_granola_lag4`        -0.0365804  0.0578194  -0.633  0.52697
`1good_granola_lag8`         0.0032916  0.0577864   0.057  0.95458
`1rosehip_tea_lag4`         -0.0282561  0.0412574  -0.685  0.49345
`1rosehip_tea_lag8`          0.0093827  0.0409527   0.229  0.81879
`Metformin Pill_lag4`       -0.0075712  0.0149706  -0.506  0.61306
`Metformin Pill_lag8`        0.0070694  0.0150144   0.471  0.63777
`1almond_butter_lag4`        0.0246033  0.0237096   1.038  0.29945
`1almond_butter_lag8`        0.0316482  0.0237030   1.335  0.18185
`1chicken_tikka_lag4`        0.0205713  0.0731207   0.281  0.77846
`1chicken_tikka_lag8`        0.1147135  0.0670457   1.711  0.08713 .
`1olive_oil_lag4`            0.0143632  0.0217386   0.661  0.50881
`1olive_oil_lag8`           -0.0187180  0.0215983  -0.867  0.38617
`1pickles_lag4`             -0.0307671  0.0473305  -0.650  0.51568
`1pickles_lag8`              0.0070490  0.0474329   0.149  0.88187
rye_sourdough_bread_lag4     0.0420333  0.0493028   0.853  0.39393
rye_sourdough_bread_lag8    -0.0375720  0.0493464  -0.761  0.44645
Walk_lag4                   -0.0014190  0.0007179  -1.977  0.04811 *
Walk_lag8                    0.0014263  0.0007181   1.986  0.04705 *
`1squash_lag4`              -0.0490897  0.0498578  -0.985  0.32486
`1squash_lag8`              -0.0731525  0.0499560  -1.464  0.14314
`1brown_rice_lag4`           0.0631259  0.0288065   2.191  0.02846 *
`1brown_rice_lag8`           0.0329319  0.0287108   1.147  0.25141
`1coffee_lag4`               0.0507808  0.0397386   1.278  0.20133
`1coffee_lag8`               0.0276141  0.0397269   0.695  0.48702
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.163 on 7390 degrees of freedom
Multiple R-squared:  0.9708,    Adjusted R-squared:  0.9705
F-statistic:  2997 on 82 and 7390 DF,  p-value: < 2.2e-16
```



A plot overlay of predicted vs actual glucose levels, 2nd linear model iteration:



A zoomed view of the forecast horizon, 2nd LM iteration:



Interpretation of Linear Regression Model Results, 2nd iteration:

The addition and use of lagged food features in the second linear model iteration helped to provide some additional time-based context about the possible lagged effects of food feature intake on glucose response and offered more granularity in how feature importance can contribute to model results over time.

While the `cod` and `cod_lag` features were not shown as being statistically significant in the resulting model coefficient table, it was surprising that the `cod_lag` features were ranked higher (feature importance plot) in terms of overall model contribution in contrast to many of the other food features. It is possible that this discrepancy could be explained by more complex non-linear feature interactions that help to inform model predictions in ways that are not immediately obvious from observing the linear model coefficients. With domain knowledge applied, it can be noted that the intake of the `cod` food feature took place during meal time (evening) when consumed along with carbohydrates such as rice. It was also interesting to note that `cod` and `cod_lag` features showed opposite direction model impacts on glucose response. The `cod_lag` features showed a contribution toward a positive glucose response (rising) after a time delay of 1 to 2 hours. The

overall context of this observation though, is that this delayed change in glucose response happened during time windows when other food features were also consumed.

It should also be noted that the overall model impact of food-based features and their lags is much smaller than that of the glucose-lag regressors and the glucose change regressors. Although the overall observed model impact of the food based regressors was smaller than that of the glucose-change and glucose-lag regressors, their importance should not be discounted, particularly in the context of the time delayed windows that occur post food intake.

The `cinnamon_lag8` (2-hour-lag) feature showed statistical significance (alpha of .05) with a negative directional model coefficient. This observation was consistent with the author's domain observations that cinnamon could have a slight lowering impact on glucose response after a time delay window of 1 to 2 hours. It should also be noted that cinnamon was always consumed with at least one other food item such as coffee, squash, or oatmeal.

XGBOOST Model Parameters:

Three model iterations were next completed using the XGBOOST algorithm adhering to the following parameters:

- The XGBOOST models were validated using the same 80-20 train-test split as the linear regression models: ordered timestamp instances were strictly preserved and not shuffled. The last 20% of dataset instances were used to validate the trained XGBOOST models.
- Hyperparameters were initially chosen and mostly remained constant for all 3 model iterations. It was decided not to optimize model hyperparameters due to time-constraints and the tine-intensive involvement of model grid search. Instead, feature engineering was used to add features between model iterations which resulted in significant improvements to variance explained (r-squared evaluation metric).
- Regularization was increased for the 3rd model iteration due to observed overfitting.
- Initially chosen hyperparameters:

  Objective_function = MSE (regression squared error)

  learning_rate = 0.05,

  subsample = 0.9,

  colsample_bynode = 1,

  reg_lambda = 2,

  max_depth = 5

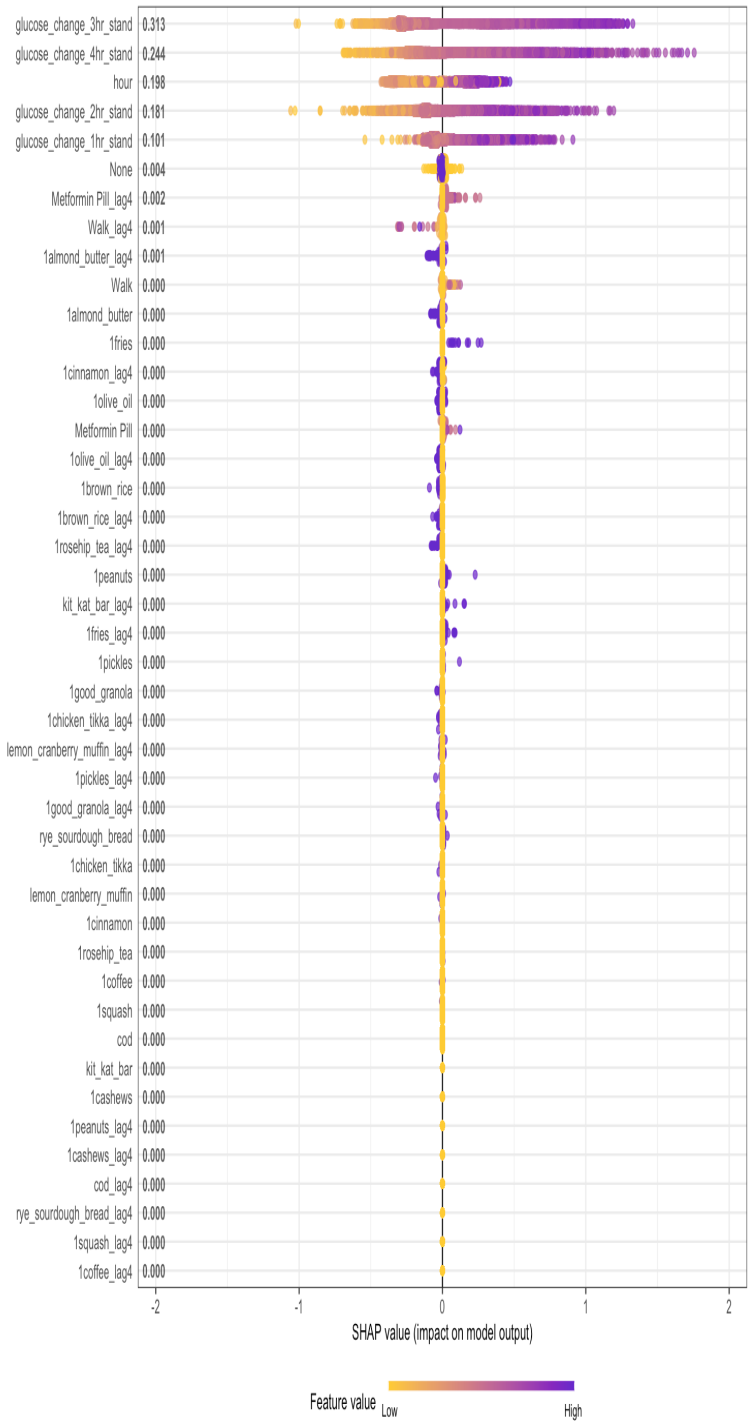
XGBOOST Model Evaluation, 1st Model Iteration:

The first XGBOOST model resulted in an r-squared of .7738 and used the same (previously modeled) food features, food feature lags, time-based regressors, and glucose change regressors. Glucose lag features were omitted from the first model iteration.

XGBOOST evaluation metrics, 1st model iteration:

R_squared	MAE	MAPE	RMSE	MASE
0.7738	0.4036	128.5618	0.5592	1.508

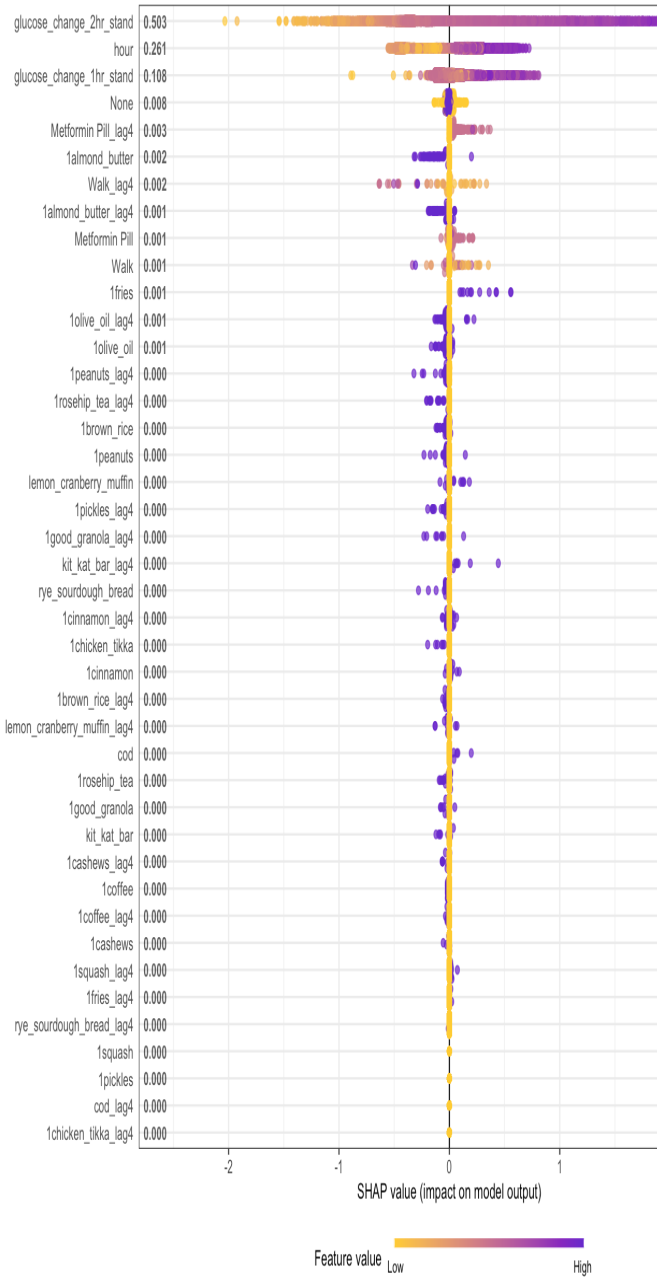
SHAP Global Feature Importance Plot, 1st XGBOOST Model Iteration:



XGBOOST Model Evaluation, 2nd Model Iteration:

The second model iteration with `glucose_change_1hr` and `glucose_change_3hr` features dropped shows more model dependence on food feature contributions but with a decreased r-squared value of .6528 (validation set):

R_squared	MAE	MAPE	RMSE	MASE
0.6528	0.5106	139.5744	0.6929	1.9078



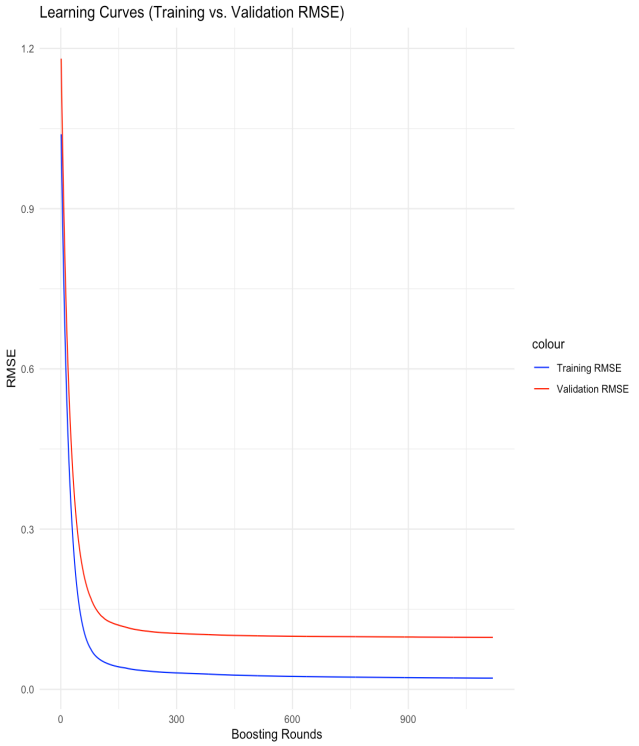
XGBOOST Model Evaluation, 3rd Model Iteration:

The third model iteration with glucose lag features added. R-squared for the validation set shows a robust .9958 for predicting glucose change. Food features are seen to have little to no model impact:

R_squared	MAE	MAPE	RMSE	MASE
0.9958	0.0329	8.4789	0.0762	0.1234

Overfitting:

Given the evaluation metric scores for the 3rd XGBOOST model iteration, overfitting was suspected. This was somewhat surprising given that the lambda regularization term was already being applied. Learning curves were plotted to confirm the overfitting of the validation data vs. train data:



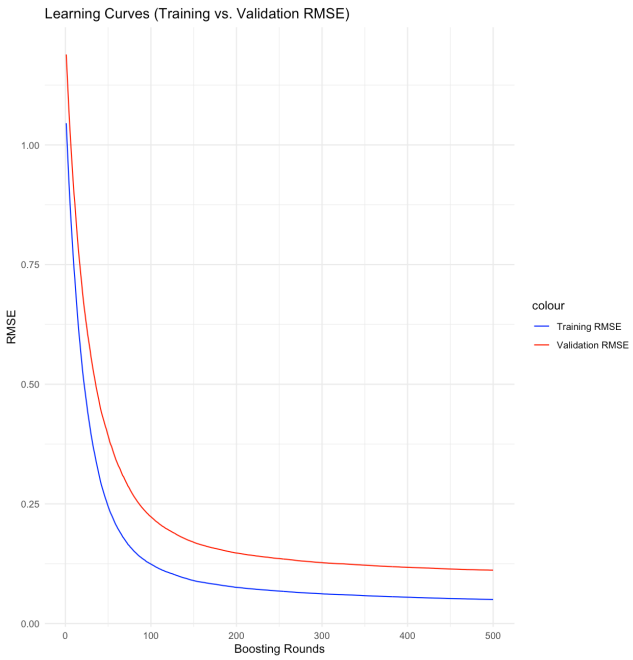
L1 regularization was accordingly increased and L2 regularization was also applied:

Reg_lambda: 20

Reg_alpha: 1.0

Evaluation metrics and learning curves post-regularization tuning, 3rd model iteration:

R_squared	MAE	MAPE	RMSE	MASE
0.991	0.0626	15.4788	0.1116	0.2339



The resulting learning curves post-regularization adjustments showed some improvement but still showed a moderate level of overfitting. The 3rd model iteration had less to offer in terms of food-feature insights and so further methods to reduce overfitting were not pursued.

XGBOOST Model Results:

As previously stated, the XGBOOST algorithm has the advantage of being able to model non-linear problems that can arise from feature interactions. XGBOOST, however, lacks the linear model coefficients used for model explainability that are inherent to a linear regression algorithm. When using the XGBOOST models, global feature importance plots were visualized using SHAP (SHapley Additive exPlanations) which was used as a non-model-agnostic approach to visualizing a feature's impact on the model's overall (global) target variable output. This approach to using SHAP is designed to work with the internal structure of the XGBOOST algorithm when making determinations about the importance of each feature.

For all three model iterations of the XGBOOST model, the food-based features were seen to have little overall impact on model output (global feature contribution). The `glucose_change` features and the time-based `hour feature` were the most important features in accounting for the variance explained in the first two model iterations. The third model iteration largely relied on the `glucose_change` features along with the `glucose_lag` features. The glucose_lag features were introduced only for the 3rd model iteration of the XGBOOST model.

It should be noted that there was a trade-off observed when all `glucose_change` and `glucose_lag` features were present as seen in the 3rd model iteration. This resulted in a very high r-squared value at the expense of observing food-based feature impact for model explainability. The 2nd model iteration provided the most insight for possible model impact for the food-based features, but with a much lower r-squared value of .65.

Although the resulting food feature model impact values (SHAP plots) were low in comparison to the `glucose_change` features, the results from the first two model iterations had considerable value when thinking about the directional impact of individual data points as plotted for the food-based features.

The following observations were specific to the Author's own personal blood-glucose response and are subject to the previously mentioned model limitations and study biases:

Kit Kat Bar: The SHAP values for Kit Kat Bar without considering lags shows overall negative contributions which can likely be attributed to the time of consumption during low glucose events. The lagged Kit Kat feature clearly showed positive SHAP values, indicating a delayed glucose spike that wasn't immediately obvious. This showed that Kit Kat Bar likely contributed to higher glucose levels, but with a time delay.

Fries: This feature exhibited positive SHAP values, indicating a likely role in raising glucose levels during the time windows that occurred post-consumption for each `Fries` feature instance. This is unsurprising given the carbohydrate content of this processed and deep-fried food.

Lemon Cranberry Muffin: While not frequently consumed, this feature also showed positive SHAP values. This result is unsurprising and consistent with domain knowledge about the carbohydrate and sugar content of this food feature.

Cinnamon and Cinnamon lag: Both features were shown to have an overall neutral impact on the SHAP feature importance plots.

Metformin Pill: This feature was shown to have positive SHAP model impact. This is likely explained by the timing of consumption, as the medication was normally taken at the time of each day's first meal when glucose levels typically form a large spike.

The Coffee and Coffee lag feature both showed neutral to slightly negative SHAP values, indicating a neutral to slightly negative directional contribution on glucose response.

Almond butter and Almond butter lag: both features showed negative SHAP values showing negative contributions on glucose levels. It should be noted that this food feature offers a combination of plant protein and healthy fat content.

RECOMMENDATIONS FOR NEXT STEPS:

a) Consider expanding on feature engineering to create interaction features to gain further insights about how different food combinations could contribute to glucose response.

b) Consider aggregating median glucose response across post-consumption time windows for each food feature and cross-reference the findings with this study's machine learning model results.

c) Investigate re-framing the problem objective as a classification task where separate classifiers are used to predict a given food feature's impact on pre-defined (binned) ranges of glucose change over time.

   Examples of possible binned ranges of the target (y) could include: low-change positive (<+2 mmol/L), moderate-change positive (+2 to + 5 mmol/L), high-change positive (> +5 mmol/L). Negative directional change could also be modelled using possible target (y) classes: low-change negative (> -2 mmol/L), moderate-change negative (-2 to -5 mmol/L), high-change negative (< -5mmol/L).

   The process of re-framing the problem into discrete variable classification tasks could align well with the end-goal of capturing the impact of different foods consumed on glucose change.

   Challenges seen in this study's regression models could be potentially addressed such as the noise from measurement variability caused by smaller sample sizes in food features consumed across a sparse dataset. By grouping outcomes into discrete ranges, the task of classification could reduce the sensitivity to noise in glucose measurement.

   Changing the problem framing to that of discrete variable range classification tasks could also allow for the use of deep learning model architectures such as the use of a temporal fusion transformer model. Currently, existing TFT model libraries contain non-agnostic feature importance plots that are unidirectional in nature. In using target classes that already reflect directional change themselves, unidirectional feature importance plots can then be suitably applied to address the goals of this problem set.

d) Begin working on model scalability strategies which could use this study's model framework(s) in ways that can be adapted to individualized health care models with the end-goal of allowing CGM users to have access to their own personalized food feature analysis relative to individual glucose response.

CONCLUSION:

Feature importance in the domain of machine learning algorithms is not commonly found as a primary task or end-goal in itself; rather the emphasis is often on the goal of producing model predictions. Built-in model packages for feature importance or custom applied feature importance algorithms are then applied after model predictions are made in order to plot feature importance rankings. The task of considering feature importance as a primary goal when looking at target variable change, does shift how modeling use-cases and problem definitions are defined. More work in this area of machine learning could prove useful for future research applications across multiple domains, including that of personalized food category feature importance in diabetes research.

ACKNOWLEDGMENTS

The use of the linear regression model architecture (R programming language) found in this study, was specifically adapted from use-cases found in Matt Danso's [5]Business Science course on Time Series Forecasting. This adapted model architecture was essential to the author's completion of many of the tasks and subsequent conclusions found in this study.

The XGBOOST model coding architecture in R was generated and iteratively adapted through a process of prompt engineering and user de-bugging using OpenAI's Chat GPT-4o large language model.

REFERENCES

[1] Danso, M. Business Science. *Correlationfunnel*. Business-science.github.io/correlationfunnel/

[2] Danso, M. Business-science. *High Performance Time Series Forecasting.* https://university.business-science.io/courses/enrolled/1032915

[3] Grogan, M. (2021, September 8). *XGBoost for Time Series Forecasting: Don't Use it Blindly.* Towards Data Science. https://towardsdatascience.com/xgboost-for-time-series-forecasting-dont-use-it-blindly-9ac24dc5dfa9

[3] OpenAI. (2024). Chat GPT-4. [Large Language Model] https://chat.openai.com/?model=gpt-4

[5] Danso, M. Business-science. *High Performance Time Series Forecasting.* https://university.business-science.io/courses/enrolled/1032915