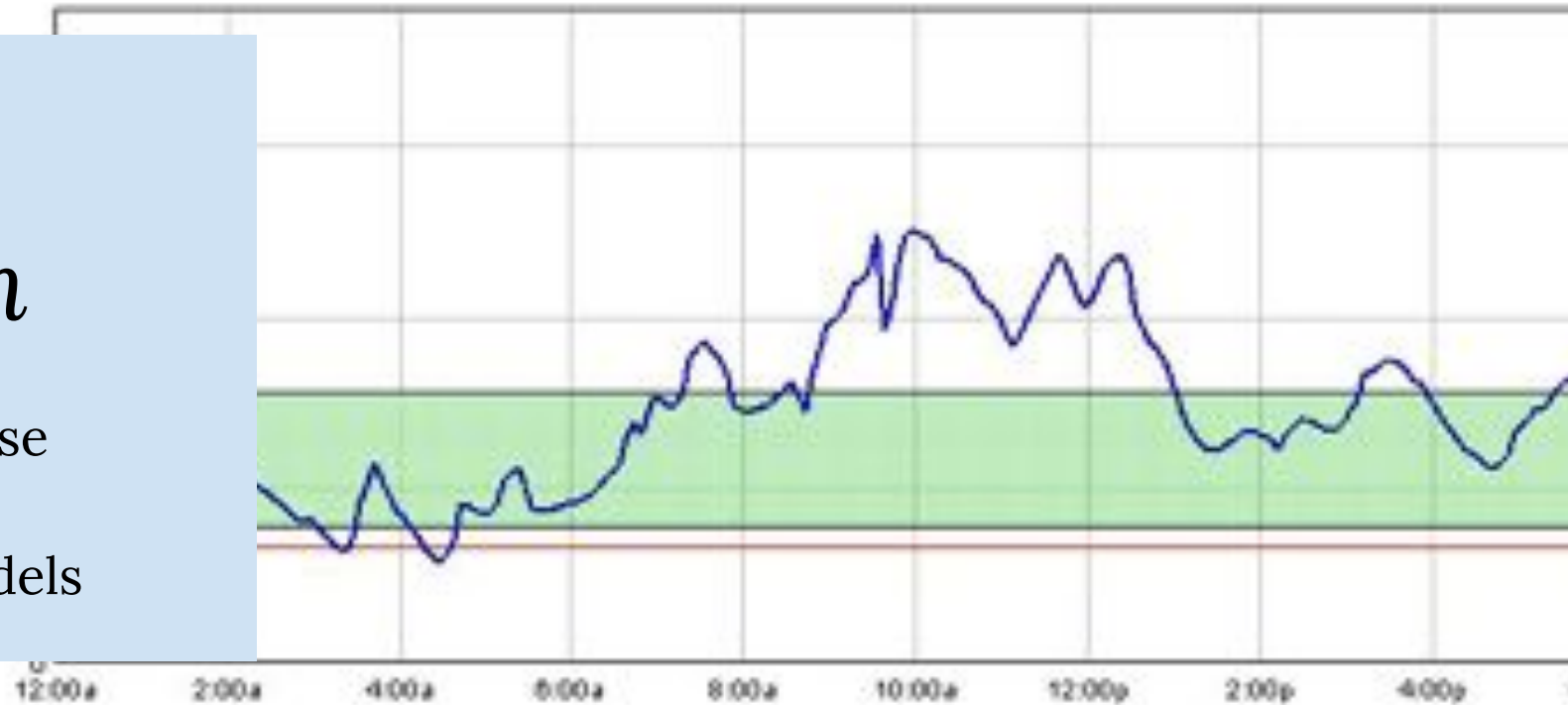


Blood Glucose Level Prediction

for Diabetics, based on Glucose
Level Logs from CGM, using
Personalized Time Series Models



Agenda

1. *Business Insights*

Understanding the problem statement, addressing the issue and building solution, risk assessment

2. *Data Science*

Data insights, Data cleaning, Model building and evaluation

3. *Further Development*

Enhancing the current results, Ideas on creating other supportive models

Problem Statement

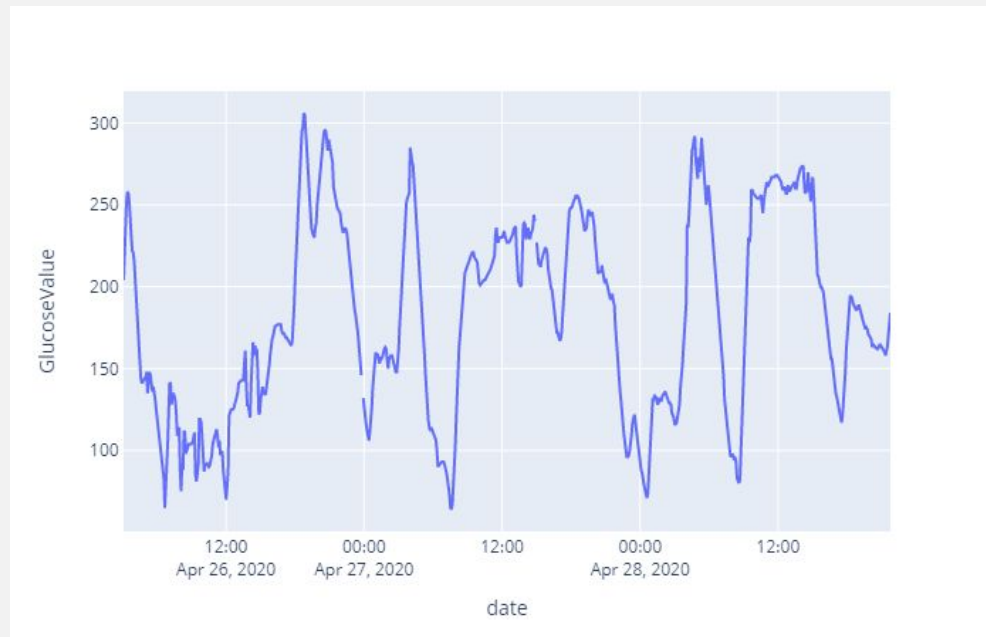
To develop a prediction algorithm that can help the patients to control blood sugar level and prevent hypoglycemia

Overview - Defining Goal

The goal of this study is to understand the data from the CGM and develop a personalized prediction model to predict the patient's BG level accurately and at regular intervals.

Understanding Data

- Data of patients who participated in **“A Randomized Trial Comparing Continuous Glucose Monitoring With and Without Routine Blood Glucose Monitoring in Adults with Type 1 Diabetes”**
- Blood Glucose Levels were recorded through Dexcom devices. A patient can have up to 288 readings per day as each recording is taken automatically, every 5 minutes
- The data contains the relative day, time and glucose level of each user over a period of ~6 months



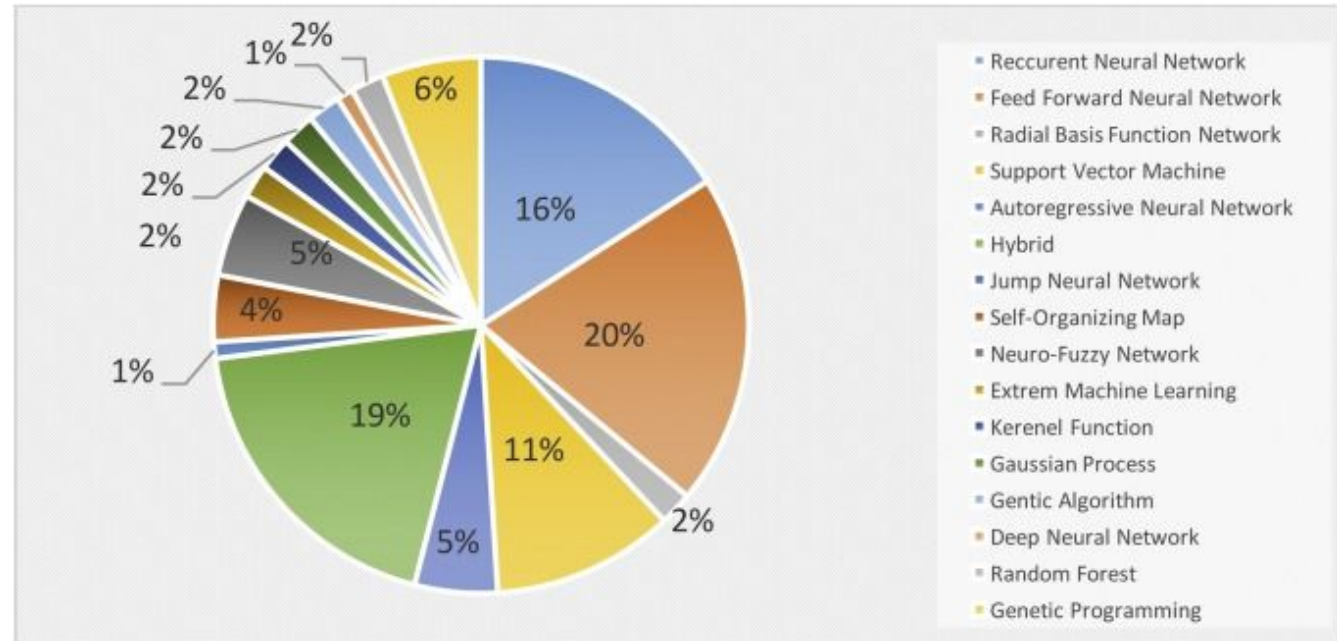
Researching Literature

Researching Literature - A lot of papers are present already on the subject matter and it is important to invest some time to understand the methodologies used and the caveats behind each.

A lot of methodologies including KNN, SVM, DT, RF, RNN were tested. Only very few articles were on time-series. Linear regression with implied regularization is another model that was widely explored.

The most important metrics conveyed in the evaluation of the models were:

1. RMSE
2. R-Squared
3. Clarke Error Grid



Defining the Model

Model

- **Time Series Model** - As the key influencing factor here is time and because of time constraints a time Series model was chosen for initial implementation
- **Personalized for each user** - Since there is no data on the user to classify/stratify them, the model will be personalized to fit each user
- **AutoRegressive Model** - Based on the ACF and PACF, AR modelling is chosen. This suggests that the output variable depends linearly on its own previous values
- **No Moving Average Detected**

Data Preparation

- A Single user was chosen at random and inspected
- For the single user, data check was performed to understand their glucose levels and ranges

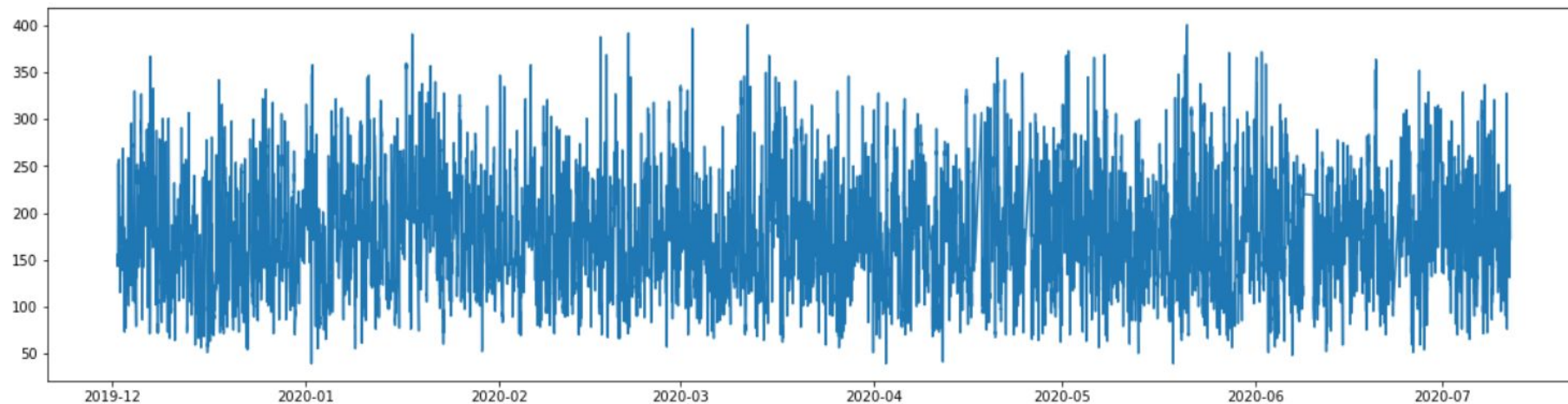
Pt Id: 293

	DexInternalDtTmDaysFromEnroll	DexInternalTm	GlucoseValue
13978688	189.0	08:19:10	280.0
13978689	189.0	08:14:10	280.0
13978690	189.0	08:09:10	279.0
13978691	189.0	08:04:10	279.0
13978692	189.0	07:59:10	280.0

Data Preparation

- **Combining Date and Time:** The dummy date '2020-01-01' is assigned to the DexInternalDtTmDaysFromEnroll column to combine the time with date

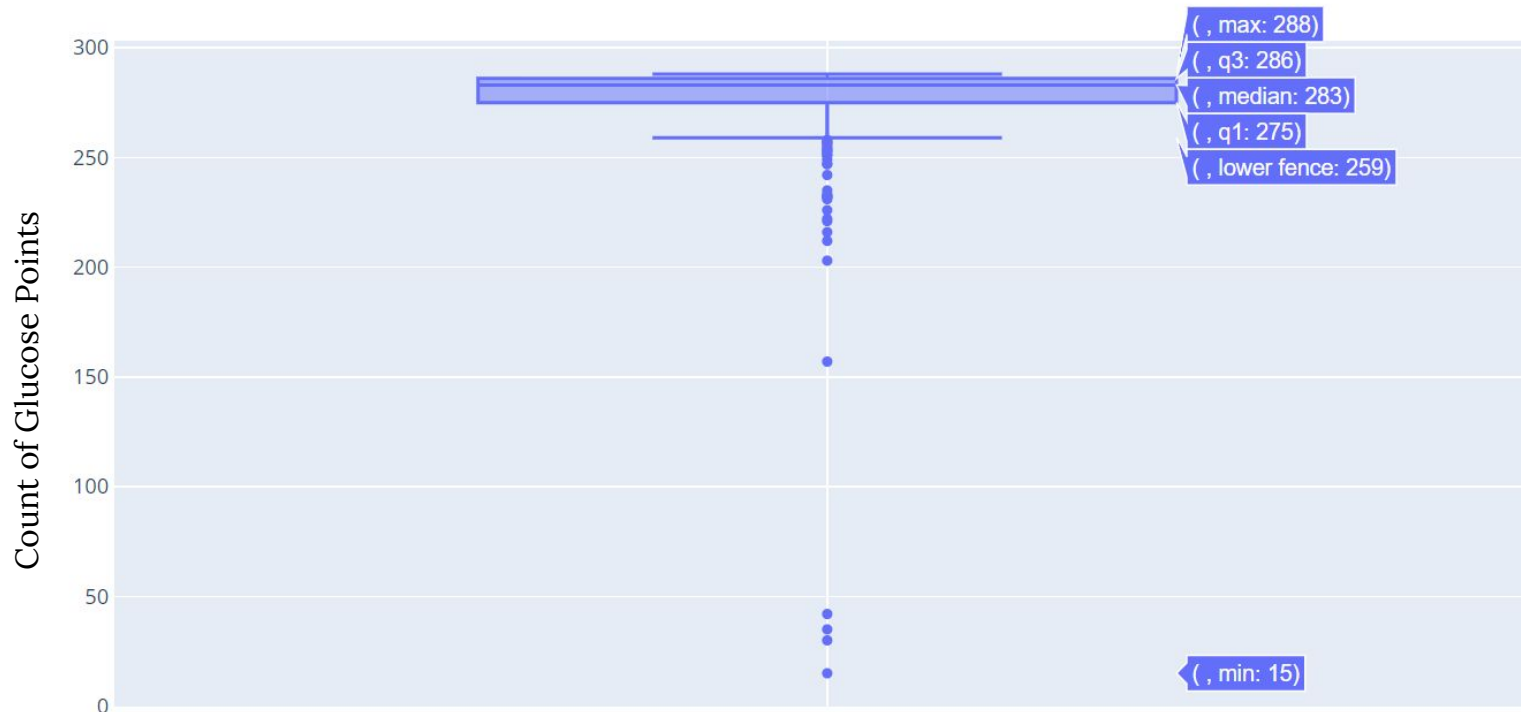
Pt Id: 293



Data Preparation

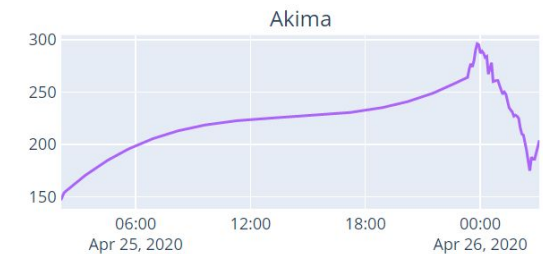
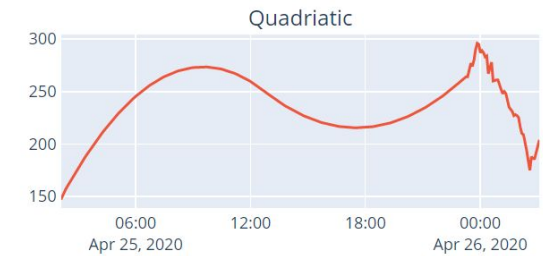
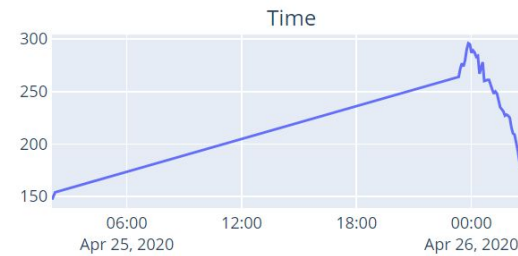
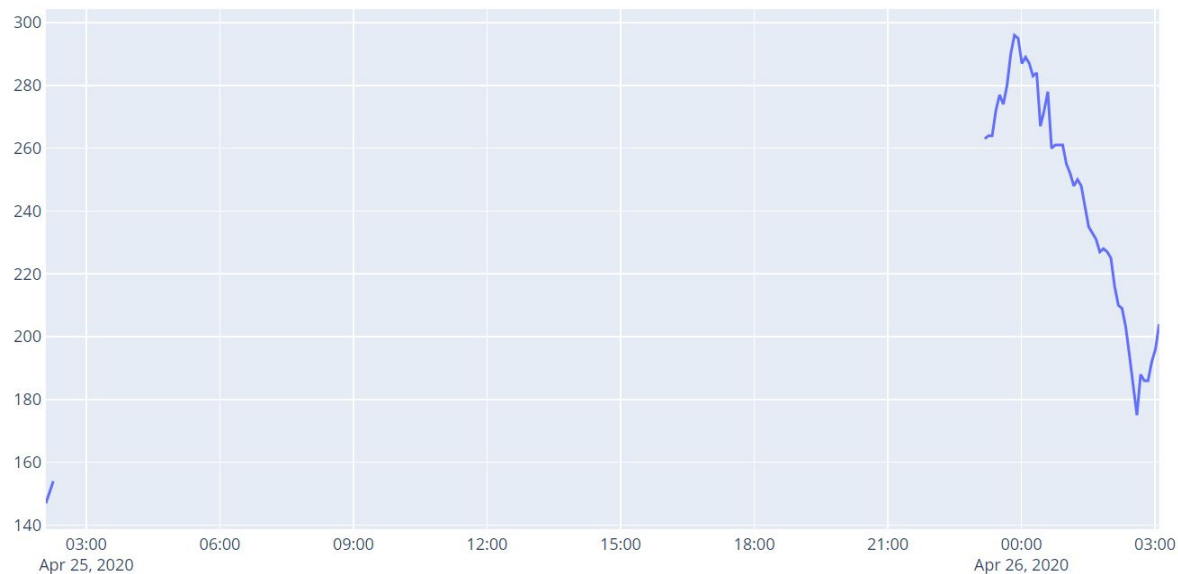
- **Dates with missing data:** In a time series model, it is important to have continuous flow of data. From CGM data, we know a day contains 288 records.

Pt Id: 293



Data Preparation

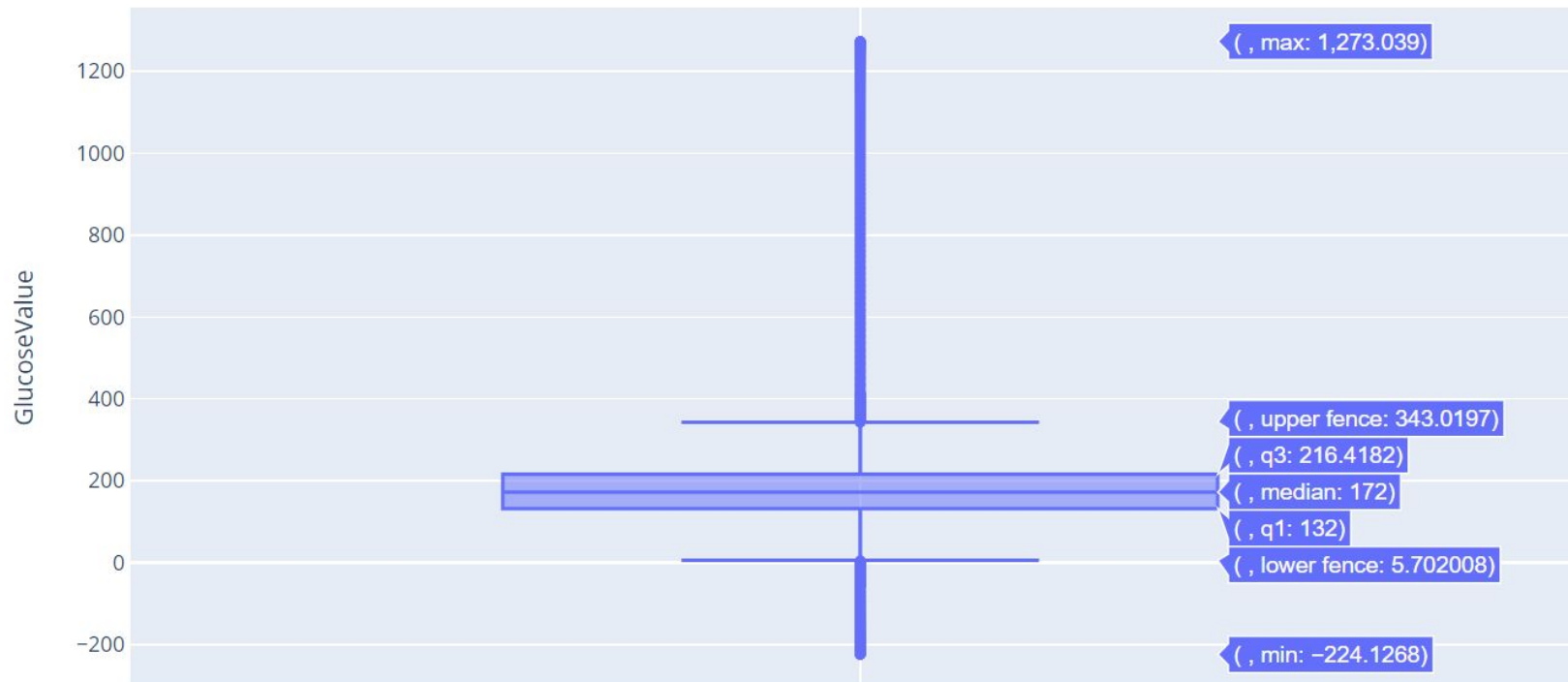
- **Handling Missing Data:** Found the missing data by creating a continuous time dataframe and assigning the patient's data to the right time slot. This showcased the missing data in the time series. It is then Interpolated to fill the data



Data Preparation

- **Handling Outliers:** This patient has some really bad days with glucose. The extreme outliers were grounded to 400 and 0 based on their direction

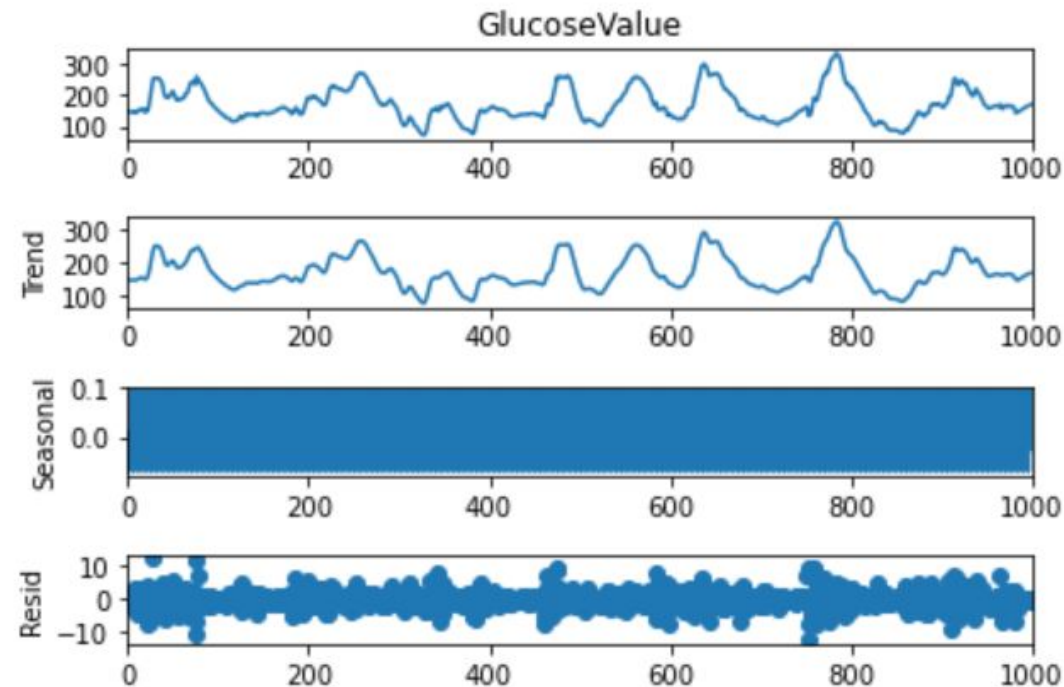
Pt Id: 293



Model Generation

- **Trend** and **Seasonality** were assessed for the data - No trend, as the data is always between bounds. Overall seasonality is not observed in the patient

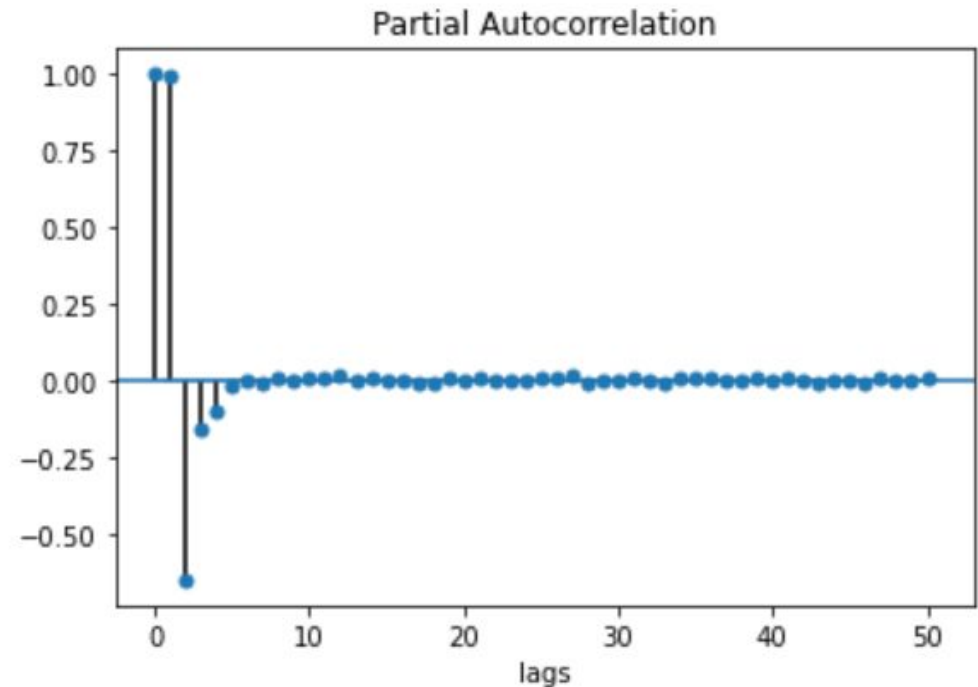
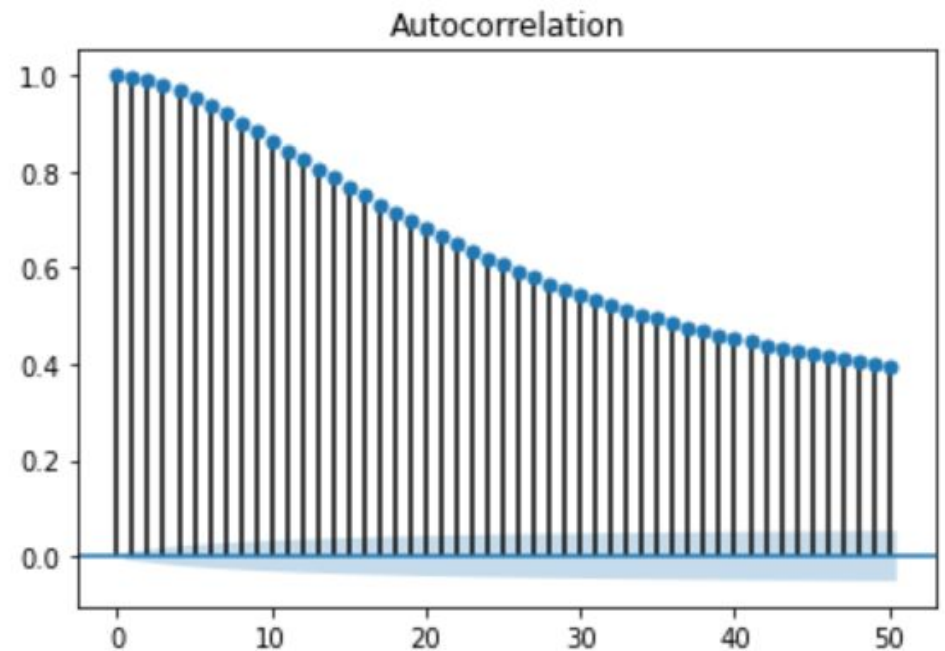
Pt Id: 293



Model Generation

Looking at ACF and PACF to determine Lags:

- No Moving Average as the autocorrelation has a geometric decay
- Partial autocorrelation shows significant peak at 4 and also peaks are present until lag 50



AR (50) was chosen: One-Step Forecast and Rolling Predictions for up to 35 minutes were analyzed

Model Evaluation

ARMA(4,0)

- Test MSE: 24.323
- R_Squared: 0.533
- Residuals normally distributed around 0

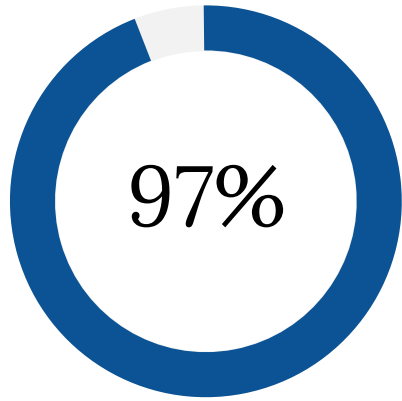
AR(50)

- Test MSE: 18.040
- R_Squared: 0.654
- Residuals normally distributed around 0

MSE - This is the average squared difference between the predictions and the actuals. It can be taken as a measure of the quality of the estimator. The MSE has to be closer to zero

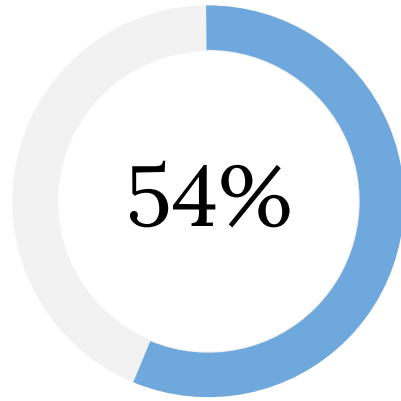
R-Squared - Also called the coefficient of determination, it is the proportion of the variance in the dependent variable that is predictable from the independent variables. It can be taken as a measure of how well the predictions reflect the actual values. Closer to 1 shows effective model

Model Evaluation – R_Sqaured



One Step
Forecast

The model showcased strong predictive power when predicting the next point in 5 minutes



Rolling
Predictions

The prediction dropped by about 32% when the prediction was for the following 30 minutes

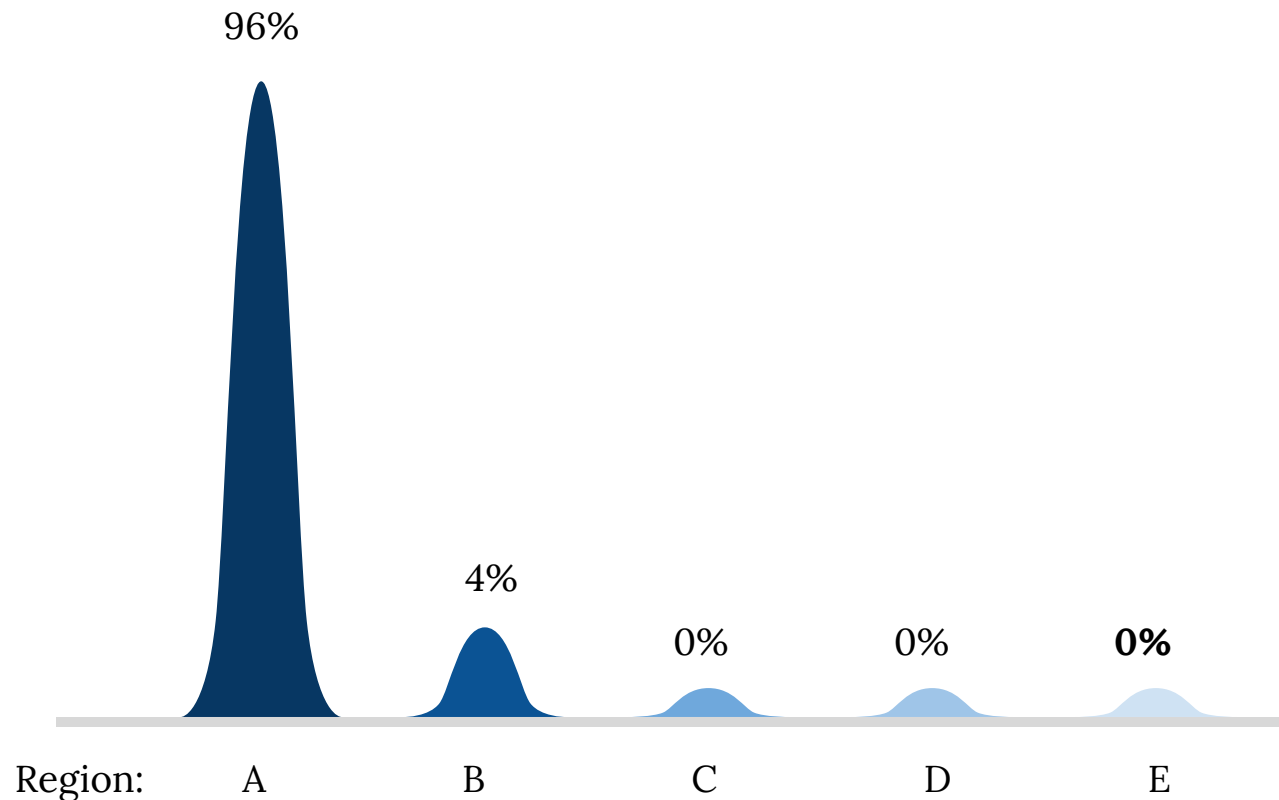
Risk Assessment:

The model accuracy was inspected across various scenarios:

- Fluctuations in MSE was observed in extreme curves (glucose level > 300 mg/dl) and ≤ 30mg/dl)
- Hypoglycemia was not handled well
- Same was observed in very hyper state
- The overall accuracy was maintained for about 10 mins and hence a **10 minute forward prediction** might be considered for Time Series Modelling
 - R_Sqaured: 92%

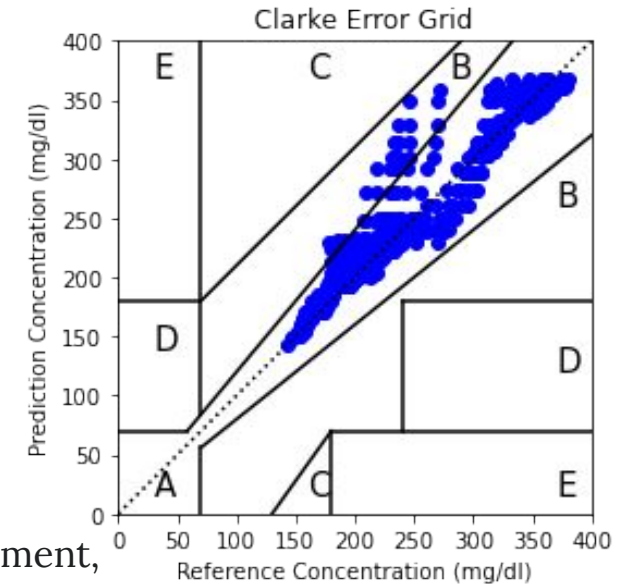
Link to Interactive Plot:
file:///C:/Users/skyli/Downloads/Predict%20glucose
%20for%20next%2030%20minutes%20(1).html

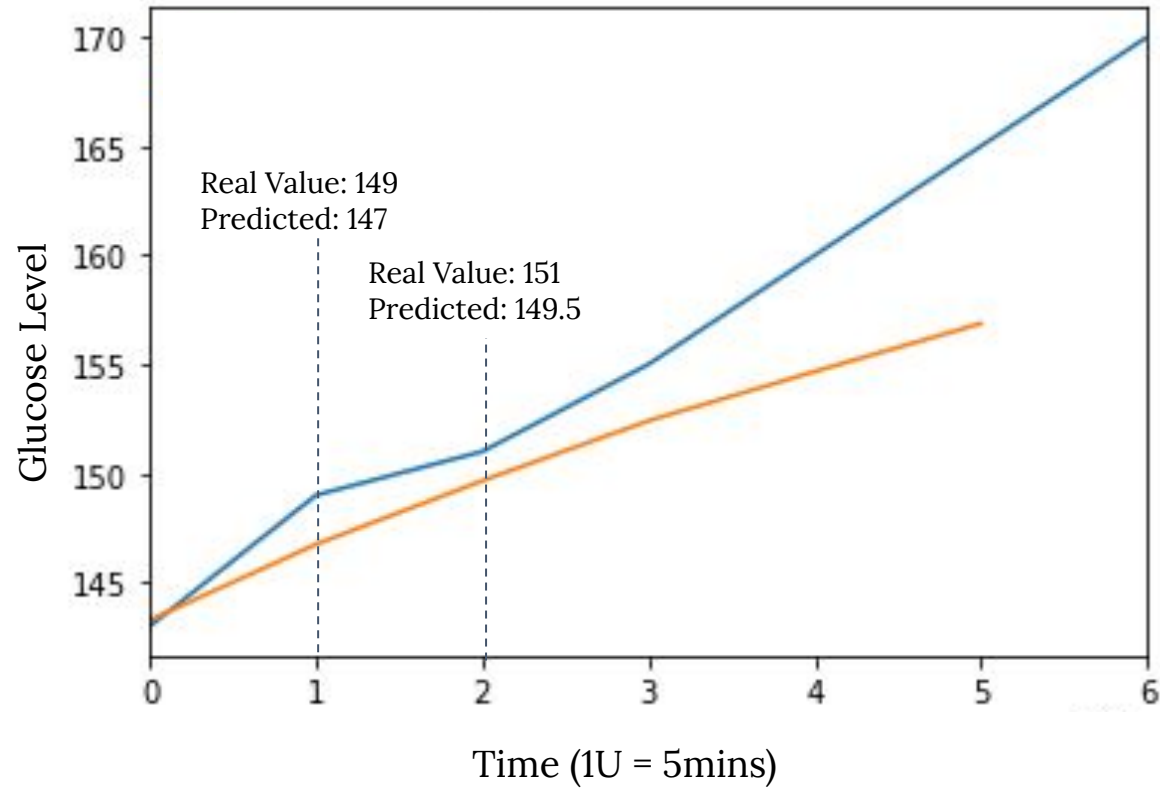
Model Evaluation - Clarke Error Grid



- Region A are those values within 20% of the reference sensor,
- Region B contains points that are outside of 20% but would not lead to inappropriate treatment,
- Region C are those points leading to unnecessary treatment,
- Region D are those points indicating a potentially dangerous failure to detect hypoglycemia or [hyperglycemia](#), and
- Region E are those points that would confuse treatment of hypoglycemia for hyperglycemia and vice versa.

The Clarke Error Grid Analysis (EGA) was developed in 1987 to quantify clinical accuracy of patient estimates of their current blood glucose as compared to the blood glucose value obtained in their meter.






Current Results

The experiment results demonstrate that the personalized time series model can give:

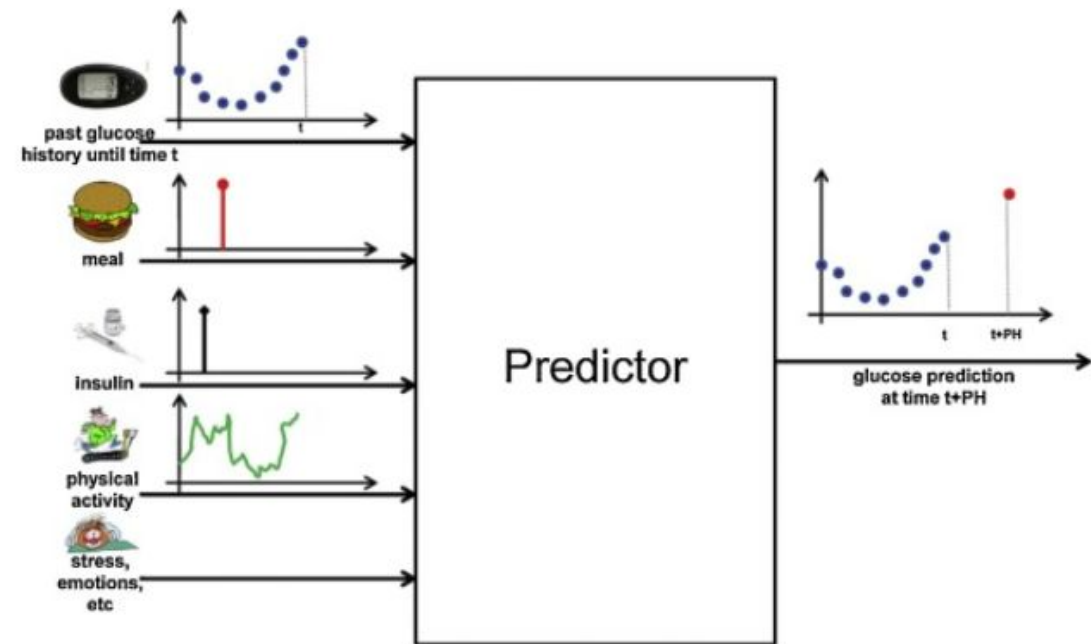
- Strong predictions for **10 mins** prediction window.
MeanAbsoluteDeviation - 2.5 mg/dl
- Do not handle the crests and troughs of the wavy glucose pattern present
- Further research and incorporation of features to understand the impact of carbs, insulin, workout and stress may lead to a robust solution.

Improving Time Series Model

- 
1. Test was repeated for another patient to confirm the findings. The AR(50) model worked well for them as well.
 2. To reduce the prediction error, a few improvements can be suggested:
 - a. Smoothing the curves
 - b. Increasing the granularity to 15 minute windows

Further development

1. Adding Additional predictors
2. Continue researching the published papers
3. Stratifying users to generalize the model
4. Exploring other Models



Thank You