Project 7 – Algorithm Prototype Development

GAM Model – Exercise recommendation model

Introduction

A generalised addictive model (GAM) is a generalised linear model that allows to learn non-linear features as the linear models in real world are not perfectly linear. It explains on the linear response variables on unknown smooth functions of some predictor variables and focuses on inference about these smooth functions.

Key features in GAM are that output can be modelled by a sum of arbitrary functions of each feature to apply on models that are showing non-linear relationship between predictors and target to utilise GAM technique in predicting the best fit model.

1. GAM modelling team

Team members:

Hyun Dong (Chris) Kim – Team Leader
Shashvat Joshi – Algorithm Code Leader
Jack

2. Projects

a. Developing a general GAM model – Completed

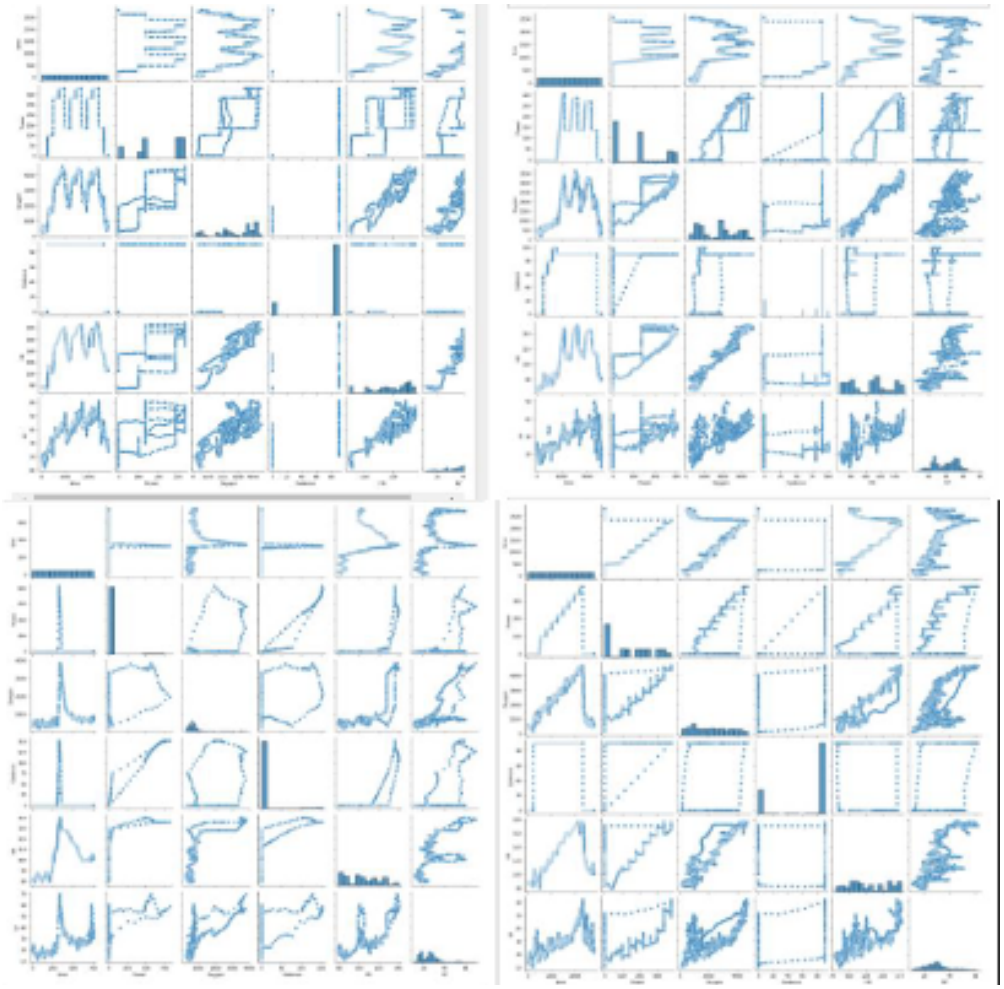
For the first part of the project, our project code leader Shashvat has come up with in-depth research on utilising GAM technique using an open source Kaggle Cycling VO2 dataset. The dataset comprises of a few variables including oxygen (O), power output (P), respiratory frequency (Rf), heart rate (HR) and cadence (W).

Understanding about the dataset with visualising using pairplot

- To visualise the correlation between trial 1, trial 2, wingate and incremental training datasets

Dataset visualisation

Pairplot Visualisation



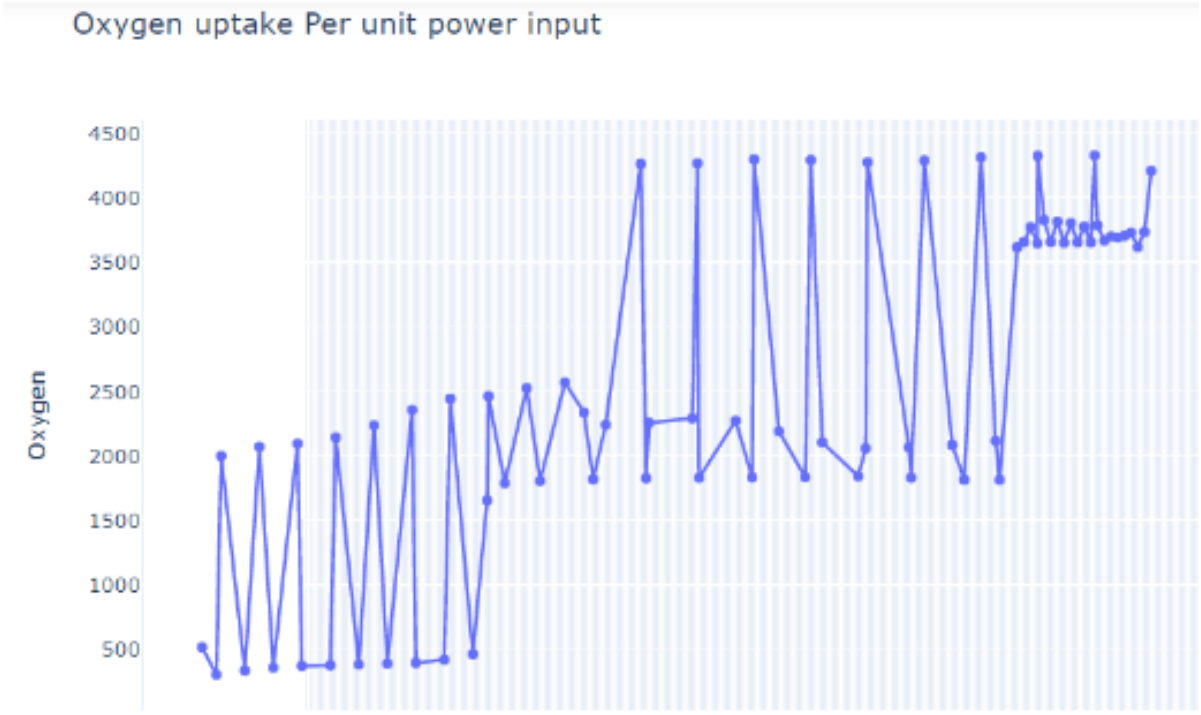
Correlation between different datasets clearly shows non-linear relationship with target variable “oxygen”

In linear regression, the relationship between target and predictors must be a simple weighted sum. GAM overcomes this limitation by replacing Beta coefficients with a flexible function where we can use higher order coefficients for non-linear relationships interpretation.

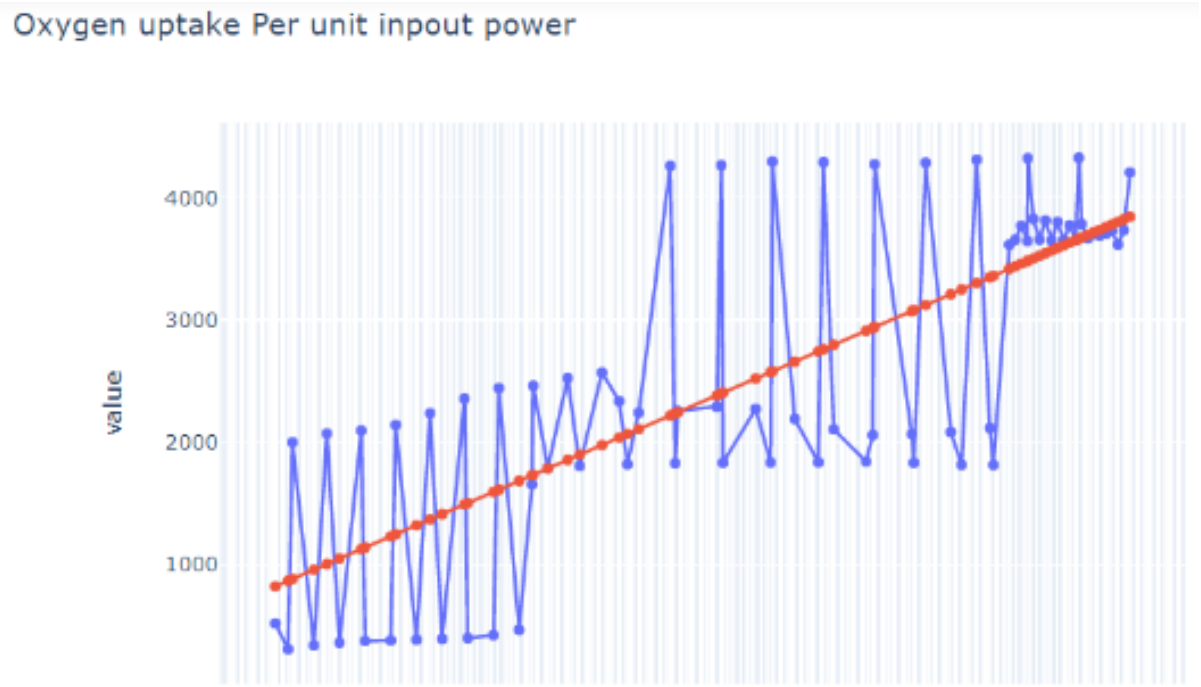
This flexible function is called a spline. They are complex functions to allow us to model non-linear relationships on each feature in which a sum of many splines forms a GAM.

Experiment - Linear Regression Model for Power & Oxygen

Fitting an Oxygen vs Power graph



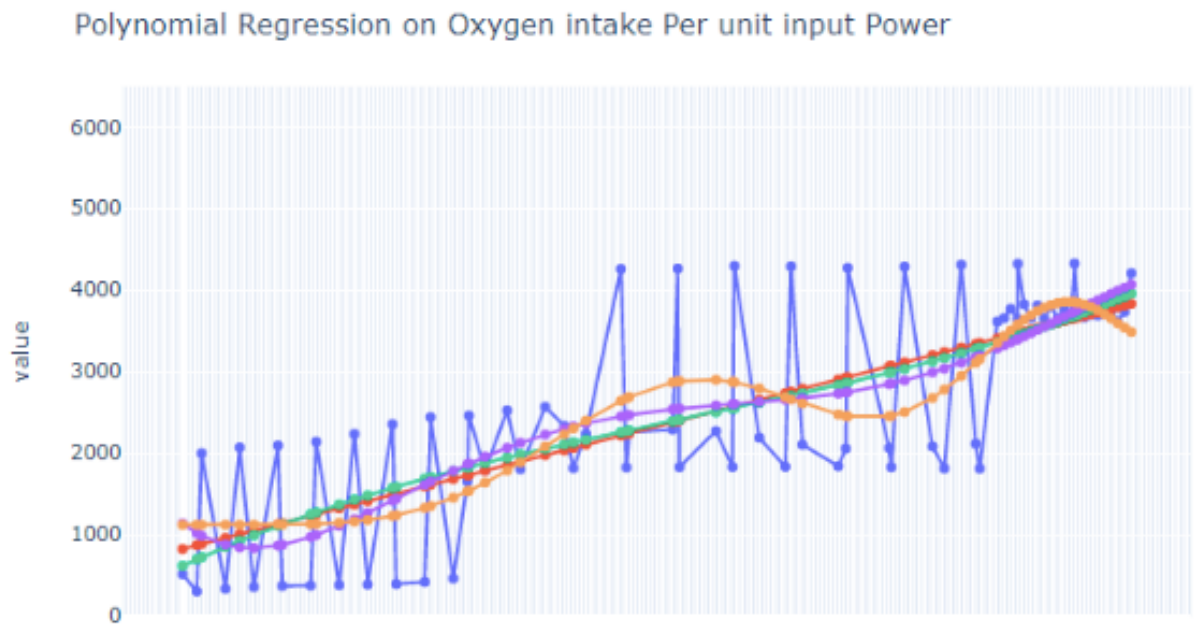
Fitting a linear regression line



Interpretation

- The line best fit does not seem to be a good solution
- Does not explain relationship between Oxygen and Power well

Polynomial features



- Building a non-linear model which will represent the relationship better
- Use polynomial features with high powers fit explain better about the relationship between power and oxygen

Hence, polynomial feature is implemented for GAM model development.

GAM Implementation

```
gam.summary()
```

```
LinearGAM
=============================================== ==========================================================
Distribution:                        NormalDist Effective DoF:                                     2.0013
Link Function:                     IdentityLink Log Likelihood:                                -1036.0691
Number of Samples:                           72 AIC:                                           2078.1408
                                                AICc:                                          2078.4941
                                                GCV:                                         745717.0844
                                                Scale:                                       708503.1252
                                                Pseudo R-Squared:                                 0.5626
=============================================== ==========================================================
Feature Function                  Lambda               Rank         EDoF         P > x        Sig. Code
================================= ==================== ============ ============ ============ ============
s(0)                              [100000.]            12           2.0          1.20e-09     ***
intercept                                              1            0.0          1.11e-16     ***
=============================================== ==========================================================
Significance codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

WARNING: Fitting splines and a linear function to a feature introduces a model identifiability problem
         which can cause p-values to appear significant when they are not.

WARNING: p-values calculated in this manner behave correctly for un-penalized models or models with
         known smoothing parameters, but when smoothing parameters have been estimated, the p-values
         are typically lower than they should be, meaning that the tests reject the null too readily.
```

In GAM, target variable can be computed by using a linear combination of multiple cariables in a model by simply using non-linear features, ie) polynomial, denoted by s, for ‘smooth function’

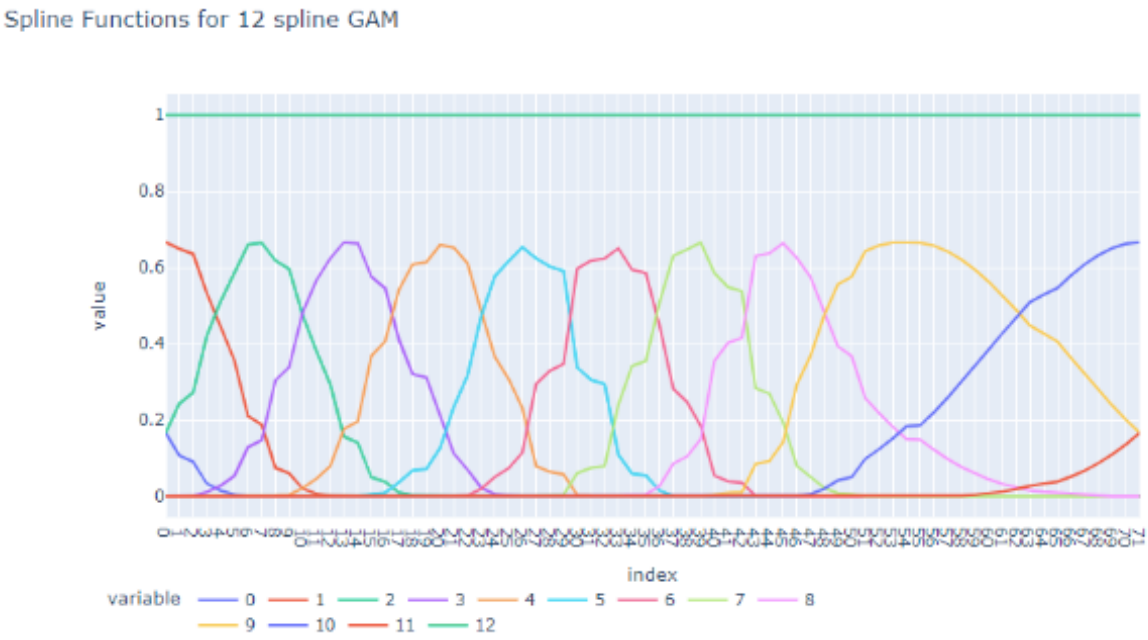
$$Z = s_0 x_0 + s_1 x_1 + \cdots + s_n x_n$$

This equation can be further explained as below:

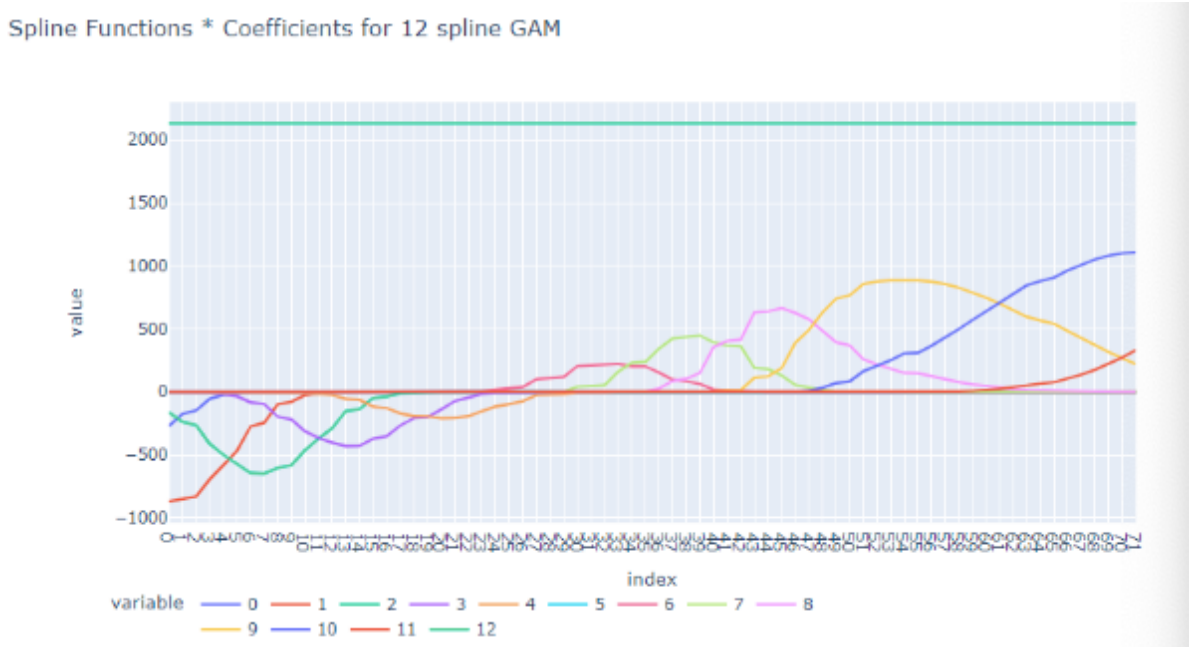
$$s(x) = \sum_{k=1}^{k} \beta_k b_k(x)$$

In this equation, Beta represents a weight, b represents a basis expansion (polynomials with high powers – multi-dimensional).

Smooth function - Spine



This looks odd since each spline function carries its own weight. Simplifying the output by the coefficients to understand better about the model.



This looks more likely to be a better visual representation as to how splines play in a role when it comes to GAM implementation.

Project outcome

- GAM factors in non-linear relationship by using polynomial features
- GAM is interpretable

Findings

Splines

- High number of splines and use cross-validation of lambda values to find the best fitting model that generalises the most

Wiggliness

- Measuring how wiggly the lines are as the number of splines increases, the lines get wigglier with respect to the feature
- Leading to overfitting problems

Overfitting Prevention

- Lambda is a parameter that penalises the splines
- The higher the lambda is, the less wiggly the lines are until it reaches a straight line

Link Functions

- Link functions can be used for different distributions
- Logit function for classification problems or Log for a log transformation

Distributions

- Various range of distributions including Poisson, Binomial, and Normal

Tensor Products

- Program interactions into GAM model, known as a tensor
- It helps modelling as it explains how variables interact with each other, rather than just considering each variable in isolation

b. Collaborative research with game development team – Completed

Web development & Game Development teams are targeting at creating a Redback coin system for the users to utilise coins in a Redback marketplace.

Keeping the users motivated, our team has focused on developing a recommendation model. Since the coins are based on number of steps, level of difficulties and other various factors, these features are considered in our model development procedures.

Factors considered include physical efforts, level of exercises, running steps, frequency, and safe limit. Safe limit is where GAM model fits perfectly in recommending exercise modules for users as the target variable, safe limit, varies on multiple variables as introduced above.

$$Z = s_0x_0 + s_1x_1 + \cdots + s_nx_n$$

According to the formula above, our GAM model is capable of factoring in multiple non-linear features to improve the accuracy of the model for producing recommendations (output) based on features that are tailored to individual users. Put it simply, as we gather more customer data, the model prediction will get better and better tailored to individual's exercise goals while allowing them to grab as many coins as possible.

We see our modelling to be a great fit for both game development and web development team as this model can be implemented in both web-based and mobile app-based applications.

However, there are certain areas of concern that require further research and collaboration with such teams including Cybersecurity, IoTs and marketing.

c. Further research to be carried on – To be continued

The goodness of fit or our exercise recommendation model is well on track. However, there are areas of concerns that require further collaboration and research to be conducted.

Areas include:
Cybersecurity – to get user consent in order to provide recommendation model. Some users would not prefer Redback operations to collect user-related data without consent and this could significantly impact on the viability of our model being utilised.

IoTs – in order for us to collect sufficient data which are tailored to individual user, collaboration with IoTs team is essential as they will be able to help us collect the "right" information in order for us to improve our modelling techniques.

Marketing – exercise recommendation model is based on providing the most effective exercise module for individual to not only maximise their training goals but also to gain as

much Redback coins as possible. To enable this from happening, collaboration with marketing team is highly recommended.

Continuous research and collaborative efforts are required to complete exercise recommendation modelling. This activity will be continued by the following students in next trimester.