Sample System Structure and Architecture Design

|  |  |  |
| --- | --- | --- |
| Version | Date | Changes |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Catalogue

[1 Introduction / Background 3](#_Toc46827533)

[2 Environment 3](#_Toc46827534)

[3 System Requirement 3](#_Toc46827535)

[4 Process 3](#_Toc46827536)

[4.1 Upload data into data management platform 4](#_Toc46827537)

[4.2 Generate raw master table 4](#_Toc46827538)

[4.3 Generate Category forecast 4](#_Toc46827539)

[4.4 Feature Engineering 4](#_Toc46827540)

[4.5 Model training and generate forecast 4](#_Toc46827541)

[4.6 Save related useful files 4](#_Toc46827542)

[5 Model Design 4](#_Toc46827543)

[5.1 Model Architecture 4](#_Toc46827544)

[5.2 Algorithm Explanation 4](#_Toc46827545)

[5.3 Data Source and Feature list 5](#_Toc46827546)

[5.4 Post Processing 10](#_Toc46827547)

[5.5 Future Improvement 10](#_Toc46827548)

# Introduction / Background

For Danone ELN International Label department team, it will take them a lot of time to calculate and generate demand plan.

The current plan is created manually. It takes a lot of time and efforts.

In order to save people’s staffing and improve the accuracy of forecast, we are going to use some data science methodology develop a forecast model to replace manual work.

# Environment

Python 3.7 with some open source libraries like pandas, numpy etc. The details can be checked in requirements.txt

After industrialization, the model will be deployed into Danone servers linking to Data lake. The model will be running monthly.

# System Requirement

An automatic process is going to be developed. This process can be used to generate IL total offtake by brand by country on monthly bases. In order to run this project, the system architecture is setup as below.

图片包含 截图, 游戏机

描述已自动生成

# Process

The goal of the use case is to generate monthly IL total offtake forecast by brand by country. This use case can be split into below uses.

### Upload data into data management platform

### Generate raw master table

### Generate Category forecast

### Feature Engineering

### Model training and generate forecast

### Save related useful files

# Model Design

### Model Architecture

We have developed a brand model for **Karicare ANZ, Nutrilon NL, Aptamil UK, C&G UK**. And a sub brand model for **Aptamil ANZ PN, Aptamil ANZ PF, Aptamil DE PN, Aptamil DE PF**.

### Algorithm Explanation

#### Algorithm explanation

It is called “an enhanced gradient boosting library” that makes use of a gradient boosting framework. Neural Network performs well when it comes to prediction problems that involve unstructured data like images and text.

But, decision tree-based algorithms are considered to be good performers when it comes to small to medium structured data or tabular data. XGboost is commonly used for supervised learning in machine learning. It was created by PhD student Tianqi Chen, University of Washington.

Let’s us understand the reason behind the good performance of XGboost -

Regularization: This is considered to be as a dominant factor of the algorithm. Regularization is a technique that is used to get rid of overfitting of the model.

Cross-Validation: We use cross-validation by importing the function from sklearn but XGboost is enabled with inbuilt CV function.

Missing Value: It is designed in such a way that it can handle missing values. It finds out the trends in the missing values and apprehends them.

Flexibility: It gives the support to objective functions. They are the function used to evaluate the performance of the model and also it can handle the user-defined validation metrics.

Save and load: It gives the power to save the data matrix and reload afterwards that saves the resources and time.

It carries out the gradient boosting decision tree algorithm. It has several different names like gradient boosting, gradient boosting machine, etc.

Boosting is nothing but ensemble techniques where previous model errors are resolved in the new models. These models are added straight until no other improvement is seen. One of the best examples of such an algorithm is the AdaBoost algorithm.

Gradient boosting is a method where the new models are created that computes the error in the previous model and then leftover is added to make the final prediction.

It uses a gradient descent algorithm that is the reason it is called a “Gradient Boosting Algorithm”. Weather classification or regression methods are supported for both types of predictive modelling problems.

#### Parameters Explanation

1. **learning\_rate** is the step size shrinkage and is used to prevent overfitting. This number ranges from 0 to 1.
2. **max\_depth** specifies the maximum depth of the tree. Increasing this number makes the model complex and increases the possibility of overfitting. The default is 6.
3. **alpha** is the L1 regularization on weights. Increasing this number makes the model more conservative.
4. **n\_estimators** is the number of boosted trees to fit
5. **subsample** is the proportion of sub samples used to train the model to account for the whole set of samples. If set to 0.5, it means that the XGBoost will randomly flush the entire set of random samples of the 50% sub samples to build a tree model, which can prevent over fitting.
6. **Gamma.** Loss reduction minimum required to make a further partition on a leaf node of the tree. the larger the more conservative the algorithm will be. Range: [0, 2]

### Data Source and Feature list

#### Brand Model

|  |  |  |
| --- | --- | --- |
|  | Brand Model | Description |
| Offtake | label\_country\_brand\_offtake\_sales1 | offtake group by label by country by brand lagging 1 month |
| Offtake | label\_country\_brand\_offtake\_sales2 | offtake group by label by country by brand lagging 2 month |
| Offtake | label\_country\_brand\_offtake\_sales3 | offtake group by label by country by brand lagging 3 month |
| Offtake | label\_country\_brand\_offtake\_sales4 | offtake group by label by country by brand lagging 4 month |
| Offtake | label\_country\_brand\_offtake\_sales5 | offtake group by label by country by brand lagging 5 month |
| Offtake | label\_country\_brand\_offtake\_sales6 | offtake group by label by country by brand lagging 6 month |
| Offtake | label\_country\_brand\_offtake\_sales7 | offtake group by label by country by brand lagging 7 month |
| Offtake | label\_country\_brand\_offtake\_sales8 | offtake group by label by country by brand lagging 8 month |
| Offtake | label\_country\_brand\_offtake\_sales9 | offtake group by label by country by brand lagging 9 month |
| Offtake | label\_country\_brand\_offtake\_sales10 | offtake group by label by country by brand lagging 10 month |
| Offtake | label\_country\_brand\_offtake\_sales11 | offtake group by label by country by brand lagging 11 month |
| Offtake | label\_country\_brand\_offtake\_sales12 | offtake group by label by country by brand lagging 12 month |
| Offtake | label\_country\_offtake\_sales1 | offtake group by label by country lagging 1 month |
| Offtake | label\_country\_offtake\_sales2 | offtake group by label by country lagging 2 month |
| Offtake | label\_country\_offtake\_sales3 | offtake group by label by country lagging 3 month |
| Offtake | label\_country\_offtake\_sales4 | offtake group by label by country lagging 4 month |
| Offtake | label\_country\_offtake\_sales5 | offtake group by label by country lagging 5 month |
| Offtake | label\_country\_offtake\_sales6 | offtake group by label by country lagging 6 month |
| Offtake | label\_country\_offtake\_sales7 | offtake group by label by country lagging 7 month |
| Offtake | label\_country\_offtake\_sales8 | offtake group by label by country lagging 8 month |
| Offtake | label\_country\_offtake\_sales9 | offtake group by label by country lagging 9 month |
| Offtake | label\_country\_offtake\_sales10 | offtake group by label by country lagging 10 month |
| Offtake | label\_country\_offtake\_sales11 | offtake group by label by country lagging 11 month |
| Offtake | label\_country\_offtake\_sales12 | offtake group by label by country lagging 12 month |
| Offtake | label\_brand\_offtake\_sales1 | offtake group by label by brand lagging 1 month |
| Offtake | label\_brand\_offtake\_sales2 | offtake group by label by brand lagging 2 month |
| Offtake | label\_brand\_offtake\_sales3 | offtake group by label by brand lagging 3 month |
| Offtake | label\_brand\_offtake\_sales4 | offtake group by label by brand lagging 4 month |
| Offtake | label\_brand\_offtake\_sales5 | offtake group by label by brand lagging 5 month |
| Offtake | label\_brand\_offtake\_sales6 | offtake group by label by brand lagging 6 month |
| Offtake | label\_brand\_offtake\_sales7 | offtake group by label by brand lagging 7 month |
| Offtake | label\_brand\_offtake\_sales8 | offtake group by label by brand lagging 8 month |
| Offtake | label\_brand\_offtake\_sales9 | offtake group by label by brand lagging 9 month |
| Offtake | label\_brand\_offtake\_sales10 | offtake group by label by brand lagging 10 month |
| Offtake | label\_brand\_offtake\_sales11 | offtake group by label by brand lagging 11 month |
| Offtake | label\_brand\_offtake\_sales12 | offtake group by label by brand lagging 12 month |
| Sellin | label\_country\_brand\_sellin\_01 | sellin group by label by country by brand lagging 1 month |
| Sellin | label\_country\_brand\_sellin\_02 | sellin group by label by country by brand lagging 2 month |
| Sellin | label\_country\_brand\_sellin\_03 | sellin group by label by country by brand lagging 3 month |
| Sellin | label\_country\_brand\_sellin\_04 | sellin group by label by country by brand lagging 4 month |
| Sellin | label\_country\_brand\_sellin\_05 | sellin group by label by country by brand lagging 5 month |
| Sellin | label\_country\_brand\_sellin\_06 | sellin group by label by country by brand lagging 6 month |
| Sellin | label\_country\_brand\_sellin\_07 | sellin group by label by country by brand lagging 7 month |
| Sellin | label\_country\_brand\_sellin\_08 | sellin group by label by country by brand lagging 8 month |
| Sellin | label\_country\_brand\_sellin\_09 | sellin group by label by country by brand lagging 9 month |
| Sellin | label\_country\_brand\_sellin\_10 | sellin group by label by country by brand lagging 10 month |
| Sellin | label\_country\_brand\_sellin\_11 | sellin group by label by country by brand lagging 11 month |
| Sellin | label\_country\_brand\_sellin\_12 | sellin group by label by country lagging 12 month |
| Sellout | label\_country\_brand\_sellout\_1 | sellout group by label by country by brand lagging 1 month |
| Sellout | label\_country\_brand\_sellout\_2 | sellout group by label by country by brand lagging 2 month |
| Sellout | label\_country\_brand\_sellout\_3 | sellout group by label by country by brand lagging 3 month |
| Sellout | label\_country\_brand\_sellout\_4 | sellout group by label by country by brand lagging 4 month |
| Sellout | label\_country\_brand\_sellout\_5 | sellout group by label by country by brand lagging 5 month |
| Sellout | label\_country\_brand\_sellout\_6 | sellout group by label by country by brand lagging 6 month |
| Sellout | label\_country\_brand\_sellout\_7 | sellout group by label by country by brand lagging 7 month |
| Sellout | label\_country\_brand\_sellout\_8 | sellout group by label by country by brand lagging 8 month |
| Sellout | label\_country\_brand\_sellout\_9 | sellout group by label by country by brand lagging 9 month |
| Sellout | label\_country\_brand\_sellout\_10 | sellout group by label by country by brand lagging 10 month |
| Sellout | label\_country\_brand\_sellout\_11 | sellout group by label by country by brand lagging 11 month |
| Sellout | label\_country\_brand\_sellout\_12 | sellout group by label by country by brand lagging 12 month |
| Category | mainstream\_sales\_1 | Mainstream Catetory sales lagging 1 month |
| Category | mainstream\_sales\_2 | Mainstream Catetory sales lagging 2 month |
| Category | mainstream\_sales\_3 | Mainstream Catetory sales lagging 3 month |
| Category | upre\_sales\_1 | Upre Catetory sales lagging 1 month |
| Category | upre\_sales\_2 | Upre Catetory sales lagging 2 month |
| Category | upre\_sales\_3 | Upre Catetory sales lagging 3 month |
| Category | spre\_sales\_1 | Spre Catetory sales lagging 1 month |
| Category | spre\_sales\_2 | Spre Catetory sales lagging 2 month |
| Category | spre\_sales\_3 | Spre Catetory sales lagging 1 month |
| Category | mainstream\_fcst | Mainstream Category sales forecast |
| Category | upre\_fcst | Upre Category sales forecast |
| Category | spre\_fcst | Spre Category sales forecast |
| Population | 0to6\_month\_population\_mean\_3M | 0 to 6 month baby population in the past 3 month |
| Population | 6to12\_month\_population\_mean\_3M | 6 to 12 month baby population in the past 3 month |
| Population | 12to36\_month\_population\_mean\_3M | 12 to 36 month baby population in the past 3 month |
| Population | 0to6\_month\_population\_mean\_6M | 0 to 6 month baby population in the past 6 month |
| Population | 6to12\_month\_population\_mean\_6M | 6 to 12 month baby population in the past 6 month |
| Population | 12to36\_month\_population\_mean\_6M | 12 to 36 month baby population in the past 6 month |
| Population | 0to6\_month\_population\_mean\_9M | 0 to 6 month baby population in the past 9 month |
| Population | 6to12\_month\_population\_mean\_9M | 6 to 12 month baby population in the past 9 month |
| Population | 12to36\_month\_population\_mean\_9M | 12 to 36 month baby population in the past 9 month |
| Population | 0to6\_month\_population\_mean\_12M | 0 to 6 month baby population in the past 12 month |
| Population | 6to12\_month\_population\_mean\_12M | 6 to 12 month baby population in the past 12 month |
| Population | 12to36\_month\_population\_mean\_12M | 12 to 36 month baby population in the past 12 month |
| One hot Encoding | ANZ\_APT | country\_brand one hot encoding |
| One hot Encoding | ANZ\_KC | country\_brand one hot encoding |
| One hot Encoding | DE\_APT | country\_brand one hot encoding |
| One hot Encoding | UK\_APT | country\_brand one hot encoding |
| One hot Encoding | UK\_C&G | country\_brand one hot encoding |
| One hot Encoding | NL\_NC | country\_brand one hot encoding |
| One hot Encoding | il | label one hot encoding |
| One hot Encoding | eib | label one hot encoding |
| One hot Encoding | di | label one hot encoding |
| Position Encoding | sin\_month | month position encoding |
| Position Encoding | cos\_month | month position encoding |
| Position Encoding | month | month position encoding |

#### *Brand Model*

|  |  |  |
| --- | --- | --- |
|  | Sub Brand Model | Description |
| Offtake | sub\_brand\_offtake\_il\_mean\_3M | Average of il total offtake by sub brand in the past 9 months |
| Offtake | sub\_brand\_offtake\_il\_mean\_6M | Average of il total offtake by sub brand in the past 6 months |
| Offtake | sub\_brand\_offtake\_il\_mean\_9M | Average of il total offtake by sub brand in the past 9 months |
| Offtake | sub\_brand\_offtake\_il\_mean\_12M | Average of il total offtake by sub brand in the past 12 months |
| Offtake | sub\_brand\_offtake\_il\_mean\_24M | Average of il total offtake by sub brand in the past 24 months |
| Offtake | sub\_brand\_offtake\_di\_mean\_3M | Average of di total offtake by sub brand in the past 3 months |
| Offtake | sub\_brand\_offtake\_di\_mean\_6M | Average of di total offtake by sub brand in the past 6 months |
| Offtake | sub\_brand\_offtake\_di\_mean\_9M | Average of di total offtake by sub brand in the past 9 months |
| Offtake | sub\_brand\_offtake\_di\_mean\_12M | Average of di total offtake by sub brand in the past 12 months |
| Offtake | sub\_brand\_offtake\_di\_mean\_24M | Average of di total offtake by sub brand in the past 24 months |
| Sellin | sub\_brand\_sellin\_il\_mean\_3M | Average of il total offtake by sub brand in the past 3 months |
| Sellin | sub\_brand\_sellin\_il\_mean\_6M | Average of il total offtake by sub brand in the past 6 months |
| Sellin | sub\_brand\_sellin\_il\_mean\_9M | Average of il total offtake by sub brand in the past 9 months |
| Sellin | sub\_brand\_sellin\_il\_mean\_12M | Average of il total offtake by sub brand in the past 12 months |
| Sellin | sub\_brand\_sellin\_il\_mean\_24M | Average of il total offtake by sub brand in the past 24 months |
| Category | price\_tier\_cat\_mean\_3M | Average of price tier in the past 3 months |
| Category | price\_tier\_cat\_mean\_6M | Average of price tier in the past 6 months |
| Category | price\_tier\_cat\_mean\_9M | Average of price tier in the past 9 months |
| Category | price\_tier\_cat\_mean\_12M | Average of price tier in the past 12 months |
| Category | price\_tier\_cat\_mean\_24M | Average of price tier in the past 24 months |
| Population | 0to6\_month\_population\_mean\_3M | Average of 0 to 6 month baby population in the past 3 month |
| Population | 0to6\_month\_population\_mean\_6M | Average of 0 to 6 month baby population in the past 6 month |
| Population | 0to6\_month\_population\_mean\_9M | Average of 0 to 6 month baby population in the past 9 month |
| Population | 0to6\_month\_population\_mean\_12M | Average of 0 to 6 month baby population in the past 12 month |
| Population | 12to36\_month\_population\_mean\_3M | Average of 12 to 36 month baby population in the past 3 month |
| Population | 12to36\_month\_population\_mean\_6M | Average of 12 to 36 month baby population in the past 6 month |
| Population | 12to36\_month\_population\_mean\_9M | Average of 12 to 36 month baby population in the past 9 month |
| Population | 12to36\_month\_population\_mean\_12M | Average of 12 to 36 month baby population in the past 12 month |
| Population | 6to12\_month\_population\_mean\_3M | Average of 6 to 12 month baby population in the past 3 month |
| Population | 6to12\_month\_population\_mean\_6M | Average of 6 to 12 month baby population in the past 6 month |
| Population | 6to12\_month\_population\_mean\_9M | Average of 6 to 12 month baby population in the past 9 month |
| Population | 6to12\_month\_population\_mean\_12M | Average of 6 to 12 month baby population in the past 12 month |
| One hot Encoding | ANZ\_APT | country\_brand one hot encoding |
| One hot Encoding | ANZ\_KC | country\_brand one hot encoding |
| One hot Encoding | DE\_APT | country\_brand one hot encoding |
| One hot Encoding | UK\_APT | country\_brand one hot encoding |
| One hot Encoding | UK\_C&G | country\_brand one hot encoding |
| One hot Encoding | NL\_NC | country\_brand one hot encoding |
| One hot Encoding | il | label one hot encoding |
| One hot Encoding | eib | label one hot encoding |
| One hot Encoding | di | label one hot encoding |
| One hot Encoding | PN | sub brand one hot encoding |
| One hot Encoding | PF | sub brand one hot encoding |
| Position Encoding | sin\_month | month position encoding |
| Position Encoding | cos\_month | month position encoding |
| Position Encoding | month | month position encoding |

### Post Processing

For 618, 11.11 and Chinese New year, there will appear some abnormal sales. And it’s a bit hard for our model to capture the abnormal sales. Such being the case, we have decided to do some post processing after we get the offtake forecast by machine learning model.

For 618 and Chinese New Year，we are using past four-year weighted average to get the uplift ratio for each brand, and then apply it into the output of AF model to get the forecast.

For 11.11, we are using past two-year weighted average to get the uplift ratio for each brand, and then apply it into the output of AF model to get the forecast.

### Future Improvement

Some data which is not added into our model because of quality issue may have an impact on the offtake forecast like A&P investment, Alibaba search index, social listening volume index etc.