**Next Basket Recommendation**

**Intro:**

On a visit to a store either offline or online, Customers purchase a list of products. Their subsequent visits to the stores is spaced over a period.

The store managers can review historical transaction of each user to analyze their purchase needs and subsequently give personalized recommendation on list of products that each user is likely to purchase in their next visit.

**Problem Description:**

Assist each customer with a shopping list for the next purchase according to their current needs We will predict the product list the user is most likely to have in the next basket

**Exploratory Data Analysis :**

**Approach**:

There are bunch of users (U) and we have transaction history for each user.

We refer each transaction as a basket (B) and items purchased in each transaction as products(P).

We use Temporal Annotated Recurring Sequence Based Predictor (it’s a hybrid approach that uses sequential and pattern-based recommender)

**Details**:

1) Capture purchasing habits – using TARS, which in turn uses FP Growth algorithm.   
 Derive the recurring base sequences and get the active sequences

2)Forecast the next basket – using the TARS based predictor

Select active sequences and compute the score for every item. Rank the item based on the score

Select top k items as the basket prediction for the customers

PFP growth Algorithm :

**Defining TARS - Temporally Annotated Recurrent Sequences :**

Sequences are generated from Periodic FP growth algorithm and each are characterized by an head and tail.

S = Milk Curd

Inter times = [a1,a2...an]

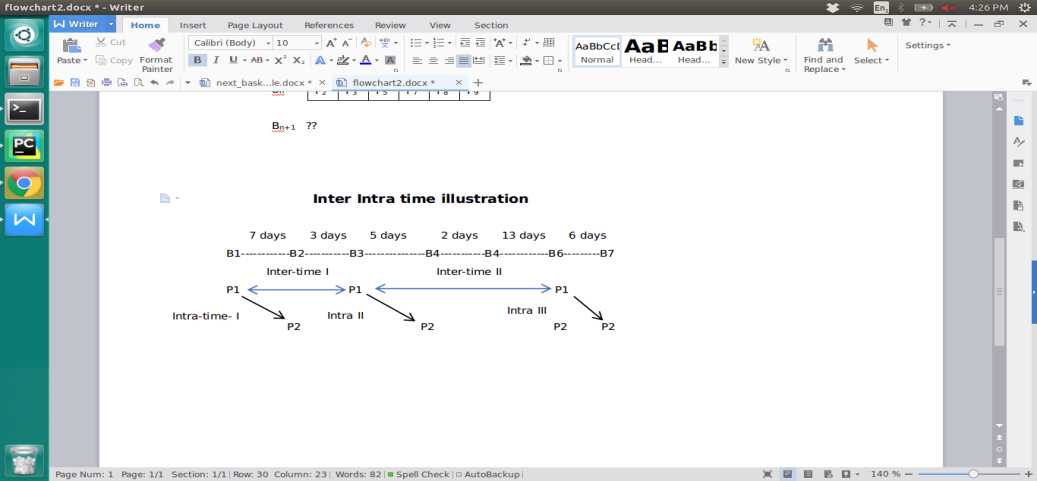
Intra times = [I1,I2,..In]

Qmed = Median number of occurrences

P = Number of periods.

**Defining Inter and Intra times :**

Let’s have a look of Products P1, P2 occurrences in B1..B7 of User Un



Given a sequence P1 P2,

Lets map the occurrences of P1 = [B1, B3,B6]

occurrences of P2 = [B2, B4,B6,B7]

Let’s map the occurrences of sequence P1 P2.

B1 B2 , B3 B4 , B6 B7

Here , we have to note that P1 occurred in B6 is not mapped with P2 in same basket because same basket items are not mapped

**Computing of Inter-times :**

Inter times are calculated based on difference of times between two consecutive heads of all occurrences.

In our case, we have three occurrences of sequence, hence we will obtain two inter times -

Intertime I = Headoccurence I - Headoccurence II = 3+ 5 = 10 days

Intertime II = Headoccurence II - Headoccurence III = 5+2+13 = 20 days

Inter time of sequence S1 = [Inter time I, Inter time II ] = [10,20]

**Computing of Intra-time :**

Intra time is the time taken between head and tail of a sequence occurrence.Lets take above occurrences itself

Intra time I = B1 - B2 = 7 days

Intra time II = B3- B4 = 5 days

Intra time III = B6 - B7 = 6 days

Intra time of sequence S2 = [Intra time I, Intra time II, Intra time III] = [7,5,6]

**Defining periodicity for a sequence :**

Periodicity is calculated by using values of Intertimemax and qmin . We will look at how to calculate them later. Assuming we have those values, period is a set of occurrences of ‘heads’ of which each and every inter-time is less than Intertimemax and number of occurrences is greater than qmin .

Let’s assume a sequence , P1 P2 and let the head P1 occur in transactions as below,

Transaction ids of head of Sequence occurrences = [1,3,6,7,8,9,15,17,18]

Inter-times for occurrences = [5,6,5,3,22,35,3,4]

Let’s assume Inter-timemax = 10 and qmin = 4

Now to calculate periods, let’s start at starting of inter time list, keep adding all inter times to first period until we reach an inter time which is greater than Inter-timemax . Now if the length of period is greater than qmin, we retain it. Otherwise discard it and move to next.

In the above case, [5,6,5,3] are added and we come across 22 which is greater than 10 (Inter-timemax). Now we look at size of first period in in terms of transactions which will be 4 + 1 = 5 as 5 transactions will give 4 inter times. Now, 5 is greater than qmin = 4, so this is period.

I Period in terms of inter times = [5,6,5,3]

I Period in in terms of transaction ids of head of sequence = [1,3,6,7,8]

Now, let’s proceed, 22 and 35 are greater than 10, so we discard them. Proceeding further we get [3,4] as next period before we end up with list. But length = 2, this means number of occurrences will be 3 which is still less than qmin which doesn’t meet the minimum requirements.

Periods - {[1,3,6,7,8]}

Now if we assume qmin = 3, so last period will also get added. Now periods for this sequence

Periods = {[1,3,6,7,8], [15,17,18]}

**Defining Recurrence for a sequence:**

Recurrence for a sequence is defined as number of periods it has occurred.

Rec(S) = |Periods|

In the above case, we have recurrence as 1,2. We define a minimum Rec(S) for a sequence to be called recurrent.

A sequence (S) is said to recurrent,

Rec(S) > Recmin

**Defining Predictor for TARS:**

We generate a list of k sequences S = {S1, S2, Sk} for every user. For all the sequences, tails are made as a list {t1, t2, ti.} and a predictor βi is initialized for each item in tails. Now each tail in list is picked and we iterate through S (sequences list). Whenever a tail tk is encountered in Sk , we find ‘increment’ value as follows

Increment = Number of occurrences of Sk in last period - Median number of occurrences in each period

βk = βk + increment

Support of tails among sequences:

At the end, we add support of each tail ti to predictor βi. In addition, we have added a modified parameter after extensive experimentation. The parameter is proportional to difference of days since last head occurrence and median intra time

**Top K items to be predicted:**

Determining the top k items was a challenge. K can be taken as mean number of items in all orders across user or median as well. But we did a regression fitting the number of items as target label and days since prior order and previous order’s quantity are the dependent variables. Regression analysis gave us better scores.

**Flow of the algorithm:**

1. Aggregate data of orders and products together
2. Pick a user and generate data for baskets with products
3. From the data of a user, generate initial set of sequences using Periodic FP Growth Algorithm
4. Eliminate sequences which has less frequency and periodic nature.
5. Map each sequence with inter time and intra time lists
6. Calculate qmin and inter-timemax based algorithm given.
7. Based on above values, determine periodicity
8. Based on periodicity, map the recurrence and filter out sequences which has less than Recmin
9. The final filtered sequences will be used for prediction. For each sequence, tail is taken and predictor is calculated.
10. A list of products with predictor value sorted in descending order is obtained.
11. Predict the ‘k’ items to be recommended from the regression analysis

**Intuition of our LGBM applied user-wise with TBP Models**

For each user , we create a data frame with a unique id which is formed from order Id and product id.

Unique\_id = 1000000\* order\_id + product\_id

This is calculated for every order\_id and distinct\_id combinations. Suppose there three orders as follows

|  |  |
| --- | --- |
| Order Id | Products |
| 103 | [1,2,3] |
| 104 | [1,3] |
| 105 | [2,3] |
| 106 | [1,2] |

Now, lets build the unique index as defined.

Distinct Products = [1,2,3]

Order Ids = [103,104,105,106]

Let’s simplify unique id formula as

Unique\_id = 1000\* order\_id + product\_id

Now index will be as follows

|  |
| --- |
| Unique ID |
| 103001 |
| 103002 |
| 103003 |
| 104001 |
| 104002 |
| 104003 |
| 105001 |
| 105002 |
| 105003 |
| 106001 |
| 106002 |
| 106003 |

Now let’s define target label, passing through each index, see that product is available in given order id. If it is, mark it as 1 , else 0, now the labels will look as follows.

|  |  |
| --- | --- |
| Unique ID | Target Label |
| 103001 | 1 |
| 103002 | 1 |
| 103003 | 1 |
| 104001 | 1 |
| 104002 | 0 |
| 104003 | 1 |
| 105001 | 0 |
| 105002 | 1 |
| 105003 | 1 |
| 106001 | 1 |
| 106002 | 1 |
| 106003 | 0 |

Now coming to features part,