**Part-2**

It is continuation of part 1. So, to understand the problem better you can start reading with first blog.

<https://kuldeepsangwan.medium.com/instacart-market-basket-analysis-6cb3511f6664>

The things I would be discussing in this part:

1. ML formulation of business problem
2. EDA
3. Feature Engineering
4. ML models
5. Introducing Autoencoders in the problem, are they any useful.
6. Trying some approaches from the Kaggle winners
7. Future works
8. References

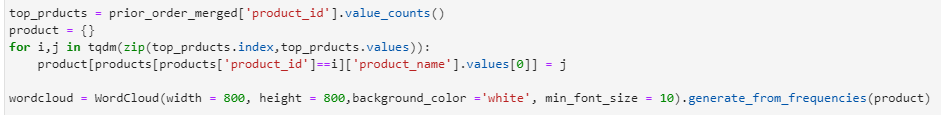
**ML formulation of business problem**

For the user to get product recommendations based on his past N orders, we need to observe patterns and generate rules which will give recommendations with high probability. Since we have over 3 Million data points, we need to automate this learning process and using Machine Learning we can achieve this to give probabilistic prediction. Machine Learning works better on large sets of data and generates rules from patterns learned from features. Other Alternative would be a rule-based system, which works best when we know the rules. But it’s very difficult to generate rules by going over all data samples manually and make sense of the patterns. This can’t guarantee in high predictive power

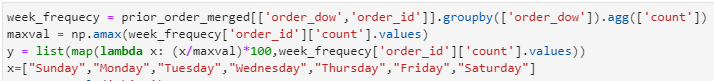
**EDA**

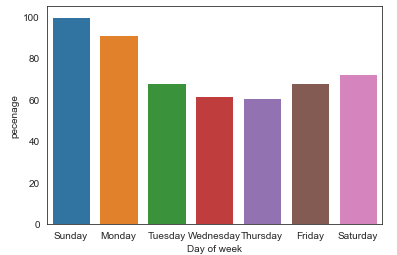
so, we understand something better by asking questions related to that. for in-depth EDA you can follow my code on gitub.

**Most ordered products**



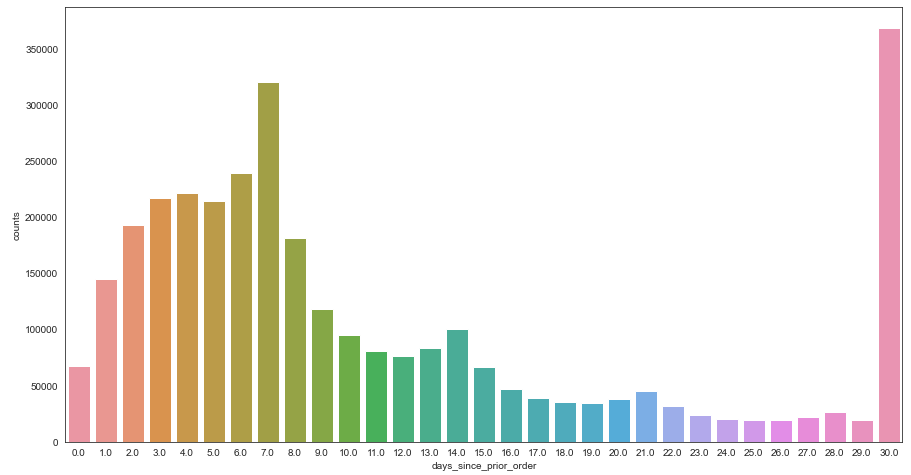
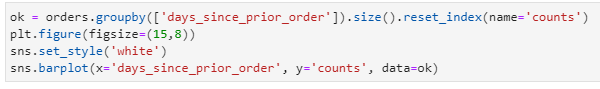


**what day of the week people order**

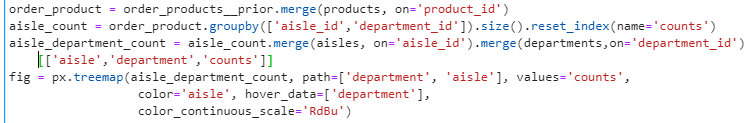


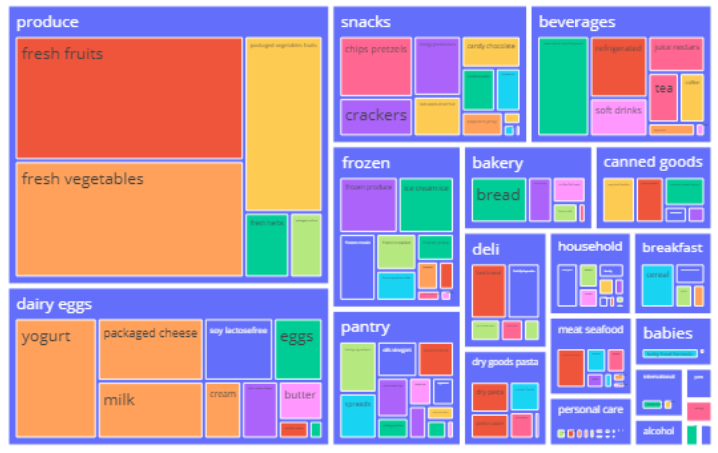
Mostly the Products are ordered on Sunday and Monday

**After how many days people order again**



**Number of Orders from Department/Aisle**





**Feature Engineering**

The idea for feature engineering is we can use every feature that we have in different tables as order feature or product related feature to get different other features.

**Note** – for getting the features for training and testing, we need to use our prior data as to get the features So we would get features likes say user related or product related then we take all these features and merge with train orders

**Different types combination of original data to get features:**

* user\_id , product\_id and days\_since\_prior\_order - to get features like days before a particular product is ordered
* user\_id and Department, user\_id and aisle – to get features related to department or aisle the product belongs to.
* product\_id and users to get features related to a specific user and a product.

The idea behind the feature creation **-**

**User features**

* **How often the user reordered items**
* **Time between orders**
* **Time of day the user visits**
* **Whether the user ordered organic, gluten-free, or Asian items in the past**
* **Features based on order sizes**
* **How many of the user’s orders contained no previously purchased items**

**Item features**

* How often the item is purchased
* Position in the cart
* How many users buy it as “one shot” item
* Stats on the number of items that co-occur with this item
* Stats on the order streak
* Probability of being reordered within N orders
* Distribution of the day of week it is ordered
* Probability it is reordered after the first order
* Statistics around the time between orders

**User x Item features**

* Number of orders in which the user purchases the item
* Days since the user last purchased the item
* Streak (number of orders in a row the user has purchased the item)
* Position in the cart
* Whether the user already ordered the item today
* Co-occurrence statistics
* Replacement items

**ML models**

ML models that I tried –

* Logistic Regression
* Naive Bayes
* Decision Trees
* Random Forest
* Gradient Boosting with XGBoost

|  |  |
| --- | --- |
| Model | CV F1 score |
| Logistic Regression | 0.305 |
| Naive Bayes | 0.322 |
| Decision Trees | 0.265 |
| Random Forest | 0.217 |
| Gradient Boosting with XGBoost | 0.431 |
| Autoencoders with GB | 0.327 |

So, in all models Gradient Boosting model performed best

**Introducing Autoencoders in the problem, are they any useful**

Autoencoders are a specific type of feedforward neural networks where the input is the same as the output. They compress the input into a lower-dimensional code and then reconstruct the output from this representation.

We want to reduce the dimensions for features so that we can train our model fast and autoencoder also help in reducing the noise of data but leaving the theory aside they don’t help much in this scenario as I got the CV f1 score as 0.327 that’s way less then our gradient Boosting model. So, this was kind of a failed attempt.

**Trying some approaches from the Kaggle winners**

So, to increase my F1 score I tried few of the Kaggle winner approaches like –

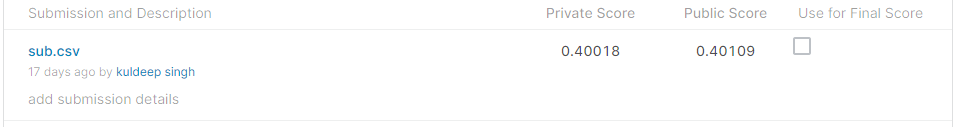
* Trying predicting None with other Products

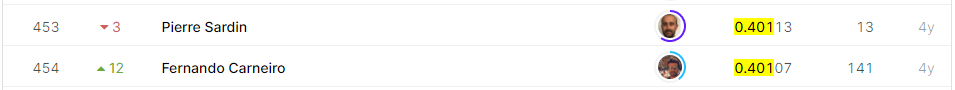
As we can see from the output file that has been published on Kaggle Competition page. From that we can see, if a user doesn’t order anything then None should be predicted. So to follow that we gonne create two different models one that we have already created that predicts products and the other one is to predict None (we are gonne consider none as product and gonne predict it with other products)

* F1 optimization approach to improve f1 score

This approach I have already discussed in part-1

**Private and Public score on Kaggle-**





**Future works**

To further improve our F1 score. We can try to predict the basket size for an order. So, if we know the basket size then we can pick the top products from our other model predictions.

I think this can improve the f1 score significantly.

**References**

1. <https://www.kaggle.com/c/instacart-market-basket-analysis>
2. <https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/35716>
3. <https://arxiv.org/abs/1206.4625>

<https://www.kaggle.com/mmueller/f1-score-expectation-maximization-in-o-n/>

1. <https://www.kaggle.com/kruegger/approximate-caclulation-of-ef1-need-o-n>
2. <https://medium.com/kaggle-blog/instacart-market-basket-analysis-feda2700cded>

**Project code -** <https://github.com/KuldeepSangwan/InstacartAnalysis>

**Contact me -** <https://www.linkedin.com/in/kuldeep881/>