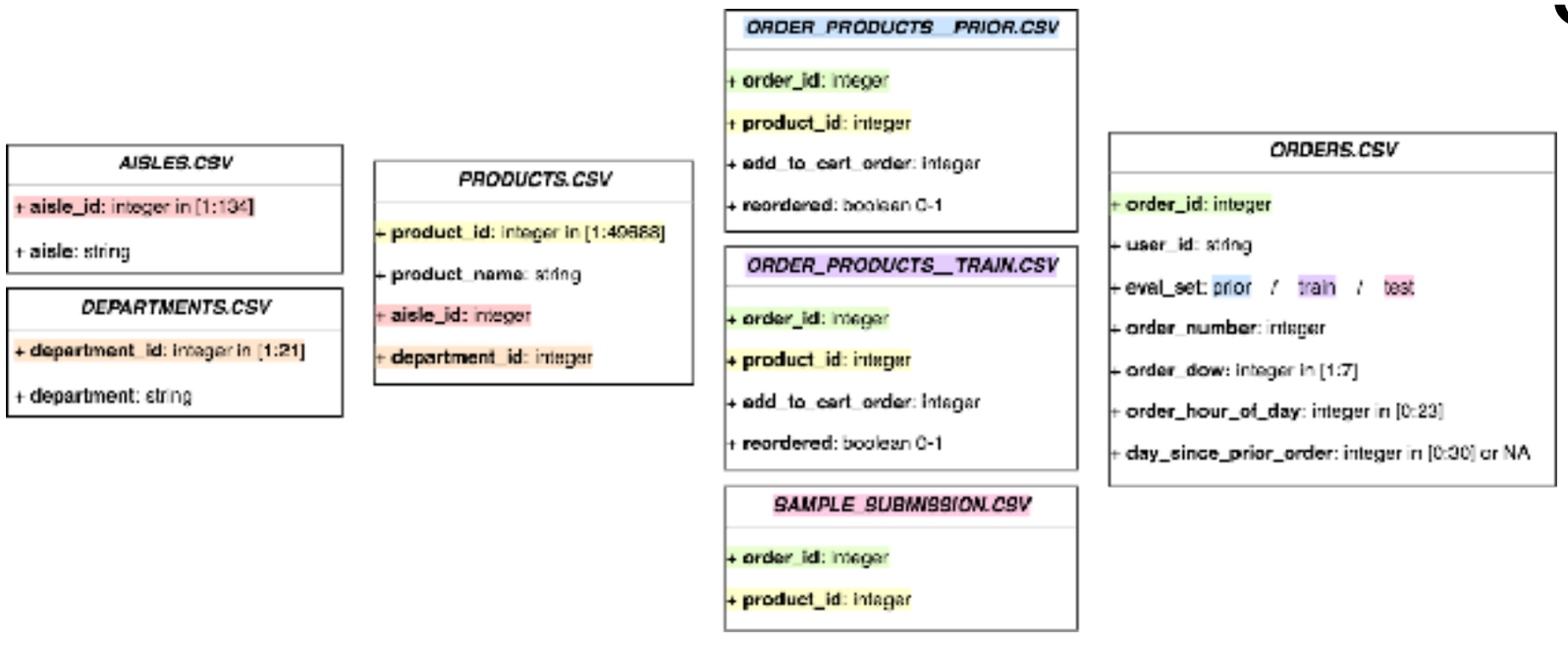
My Instacart Kaggle competition

Kuan Zhou

Datasets understanding



1), The datasets are like database. **It also have some NA - days_since_prior_order of each user first order;**

2), Prior, train, test for orders. **Prior is used to generate features for both train/test**, train is used to train;

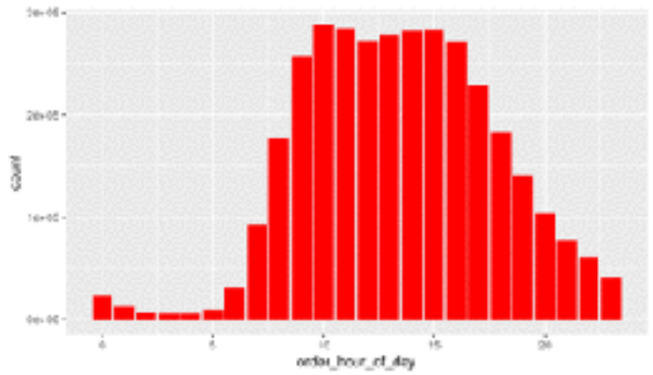
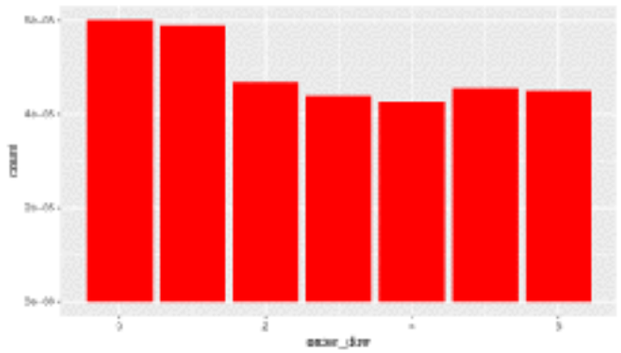
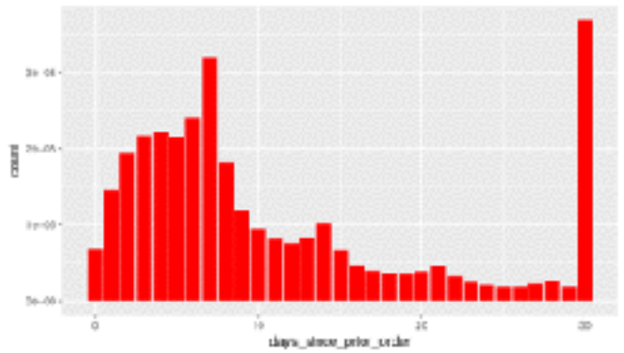
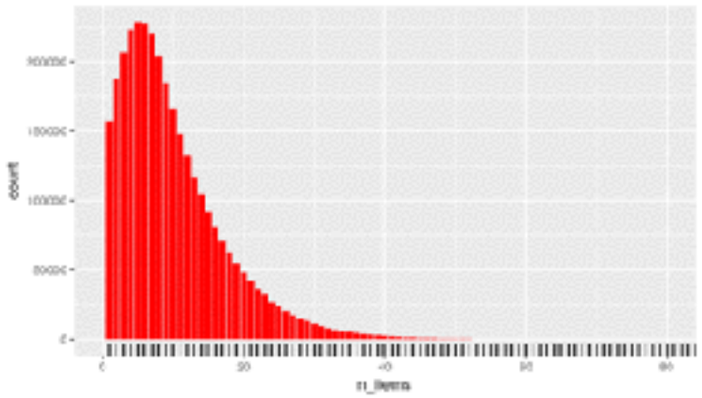
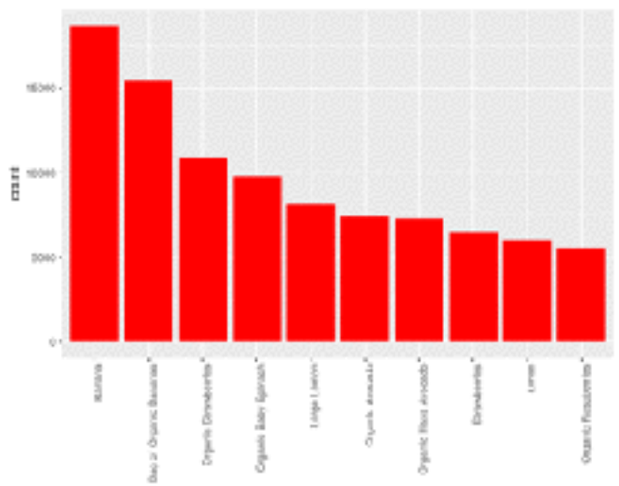
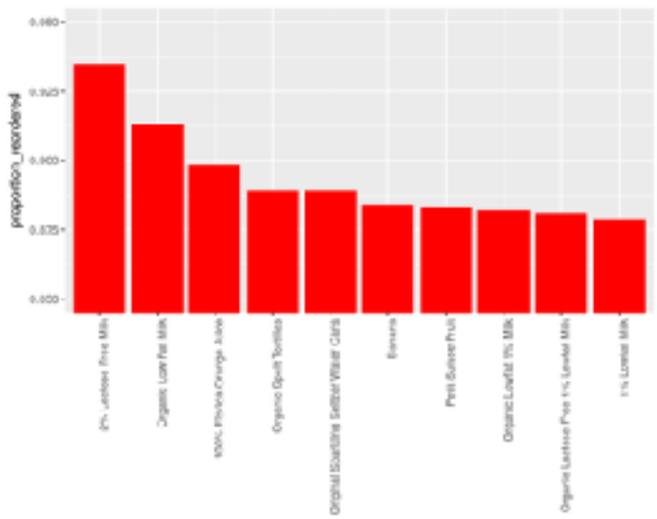
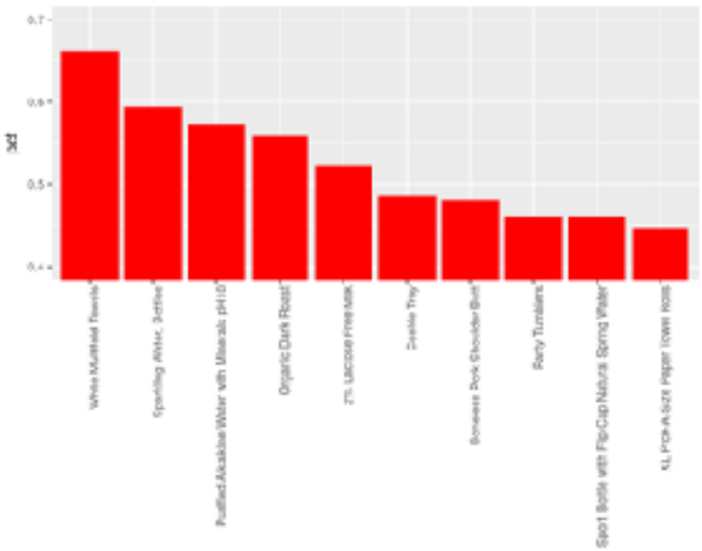
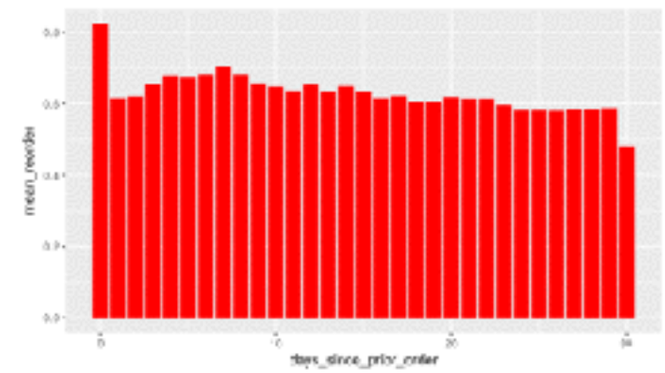
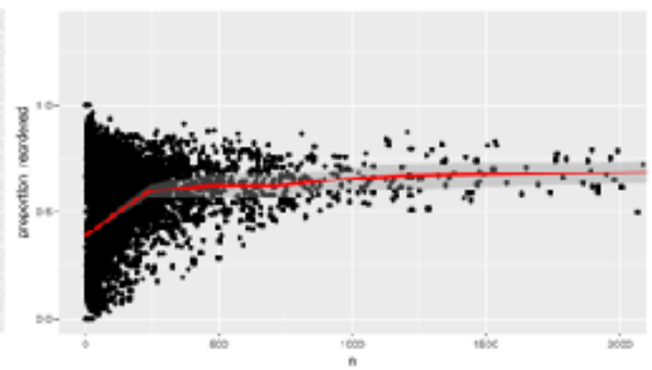
3) order_products_train with **add_to_cart_order and label;**

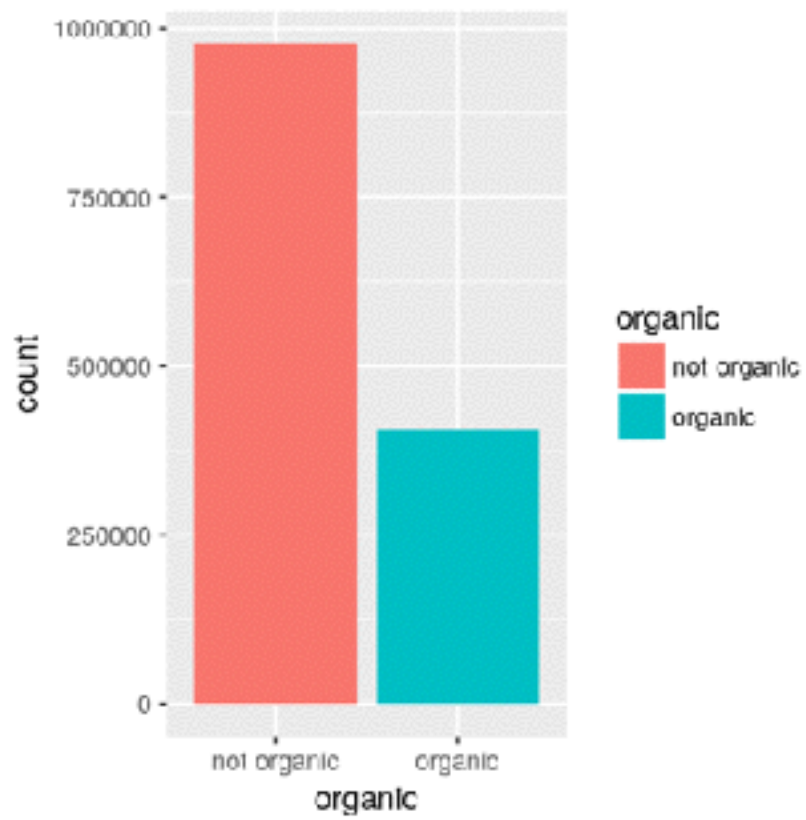
4) order_products_prior with add_to_cart_order and label;

5) product_id to products, aisles and departments.

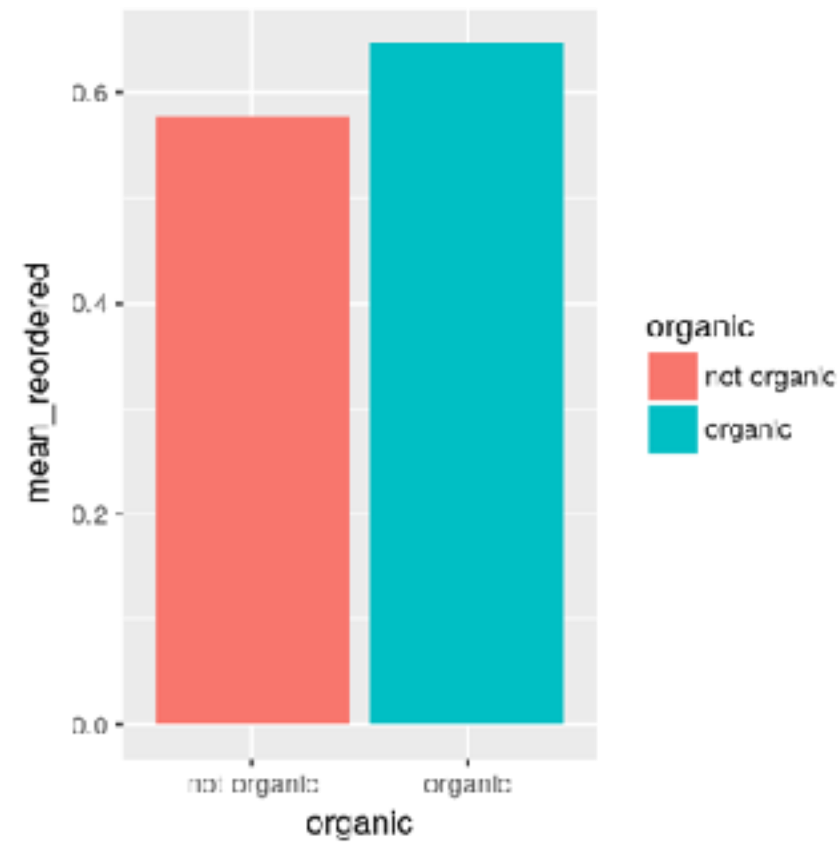
EDA

- Kernel: https://www.kaggle.com/philippsp/exploratory-analysis-instacart



1000000
750000
500000
250000
0
count
not organic
organic
organic
organic
not organic
organic



0.6
0.4
0.2
0.0
mean_reordered
not organic
organic
organic
organic
not organic
organic

produce
snacks
dairy eggs
beverages
frozen
personal care
pantry
bakery
deli
babies
international
canned goods
dry goods pasta
missing
household
meat seafood
breakfast
alcohol
pets
other
bulk

personal care
pantry
dairy eggs
household
beverages
canned goods
dry goods pasta
produce
bakery
missing
babies
alcohol
snacks
frozen
deli
international
breakfast
pets
meat seafood
other

Feature engineering

- **Users:**_user_total_orders, _user_sum_days_since_prior_order, _user_mean_days_since_prior_order, _user_reorder_ratio, _user_total_products, _user_distinct_products, _user_average_basket;

- **Prds:**_prod_tot_cnts, _reorder_tot_cnts_of_this_prod,_prod_order_once, _prod_order_more_than_once, _prod_reorder_prob, _prod_reorder_ratio, _prod_reorder_times;

- **Up:** _up_order_count, _up_first_order_number, _up_last_order_number, _up_average_cart_position, _up_order_rate, _up_order_since_last_order, _up_order_rate_since_first_order.

- Add **up** features to train datasets.

- Work with group by() and dictionary, lambda function to generate.

Tried features

- Median, mean, std, max, aisle_reordered, dep_reordered, weekend_ratio, time_or_day

- user_prd_matching

Other features

- Word2vev using gensim;
- Word embeddings;
- order_streak.

Word2vec

- Kernel: https://www.kaggle.com/omarito/word2vec-for-products-analysis-0-01-lb

- 1) Package: gensim, PCA applied to word vectors

- 2) SkipGram model, embedding_size=32, num_negative_sampled=64

order_streak

- How many times each prds has been ordered or not ordered in last 5 orders of users.

- The sign indicates the type of the streak (ordered vs not ordered). So for example "-3" means: product has not been ordered the last 3 previous orders and there is either no 4th previous order or the product has been ordered back then.

- Kernel: https://www.kaggle.com/mmueller/order-streaks-feature/code

Techniques

- CV
- F1-max
- Parameters tuning - gaussian process optimizer
- Feature importance
- Ensemble - median bagging
- Threshold(using F1-max)

CV

- order_id or user_id

- Because we are predicting what will appears in each user's next order: thus user_id is better.

F1 max

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

Ensemble

- bagging with LB;
- Lgbm and XGB with different parameters;
- median of 15 submissions.

I did not try

- Predict order size first.

- Advanced models.