Instacart Market Basket Analysis

01.

Introduction and EDA

02.

1. Featurization and Modeling

03.

1. Conclusion and Recommendation

01.

Introduction and EDA

Introduction

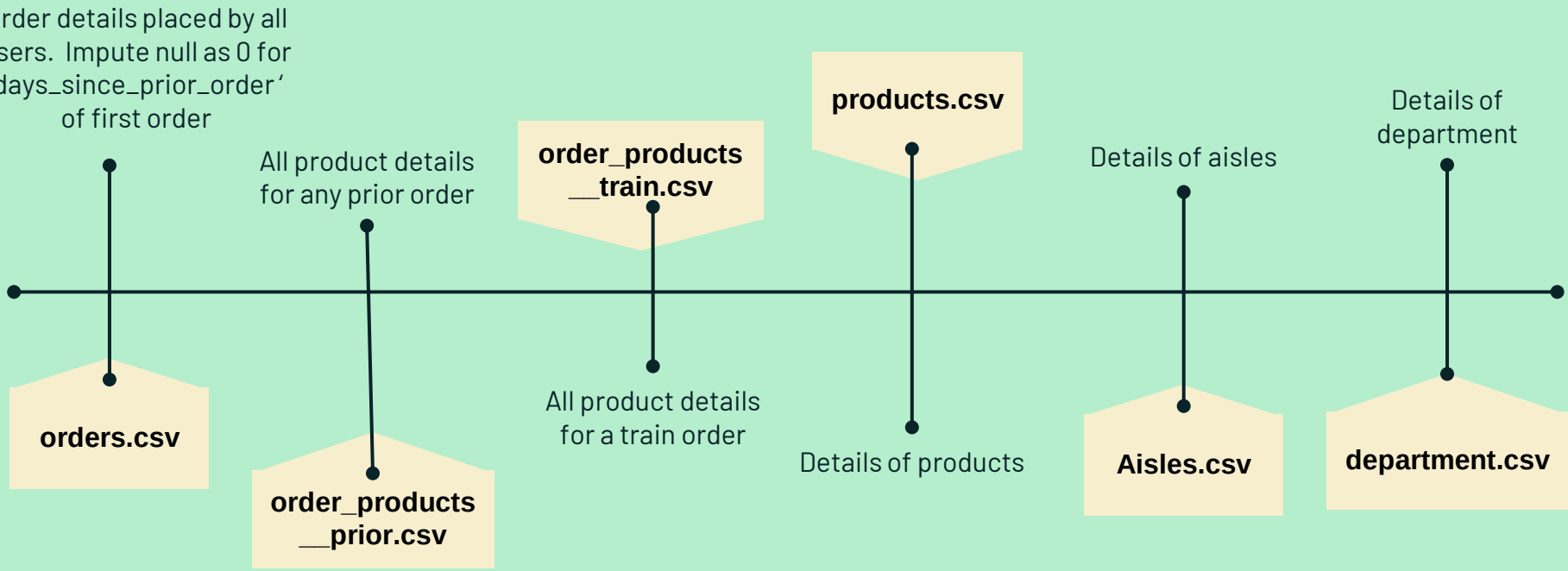
Instacart is an American company that operates a **grocery delivery and pick-up service** in the United States and Canada.

Instacart would like to maintain **customer retention rate** and promote customer **shopping experience**.

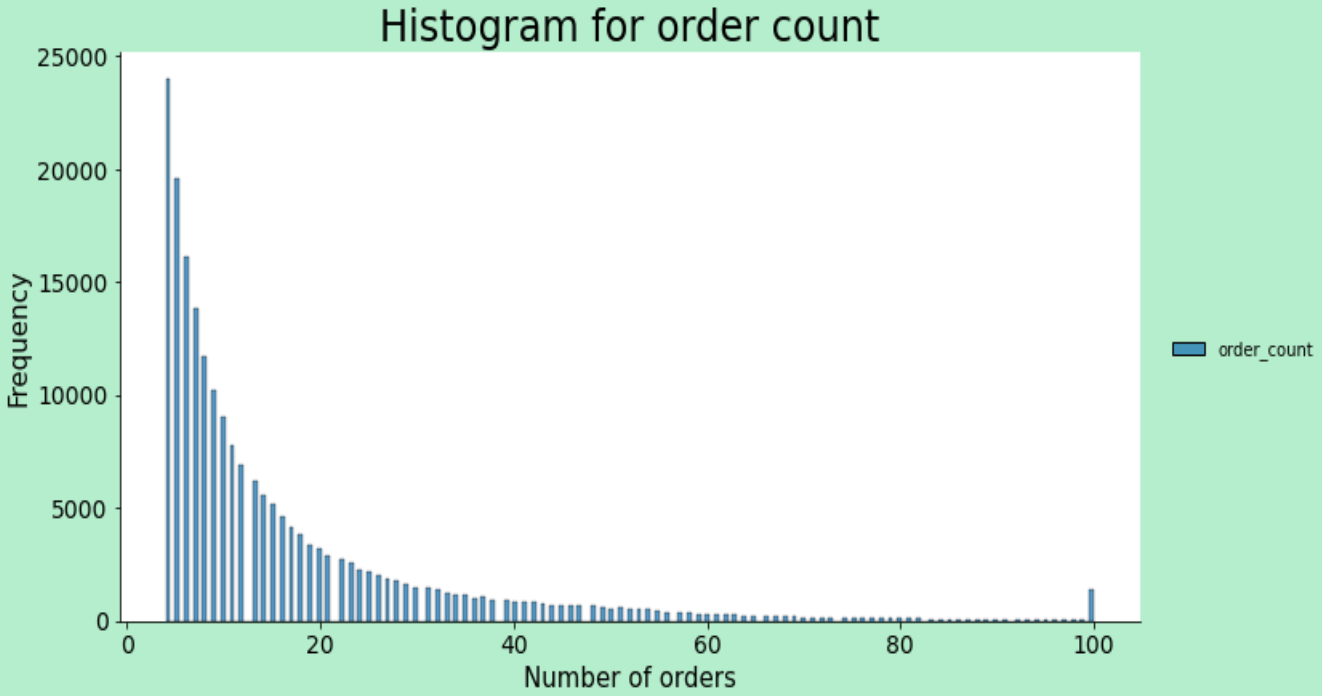
Objective

Apply machine learn to **predict** whether customers will purchase the products **repeatedly** in their next order.

Dataset Description

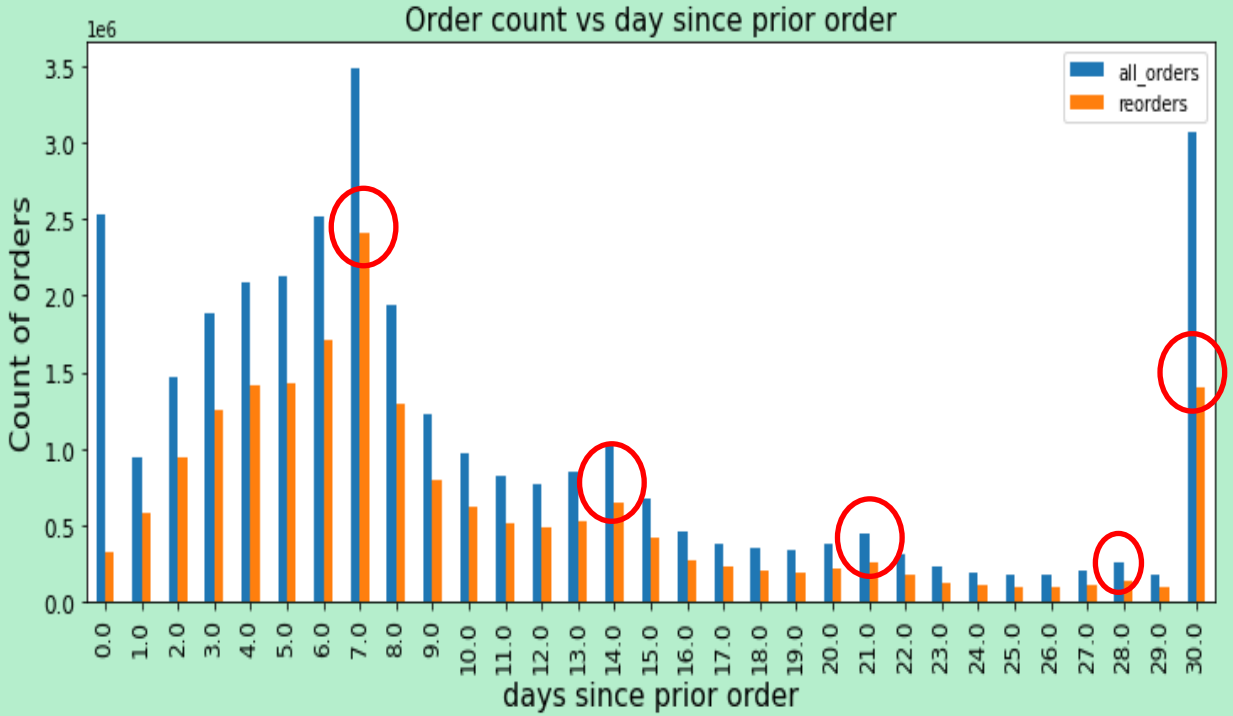


Order count histogram



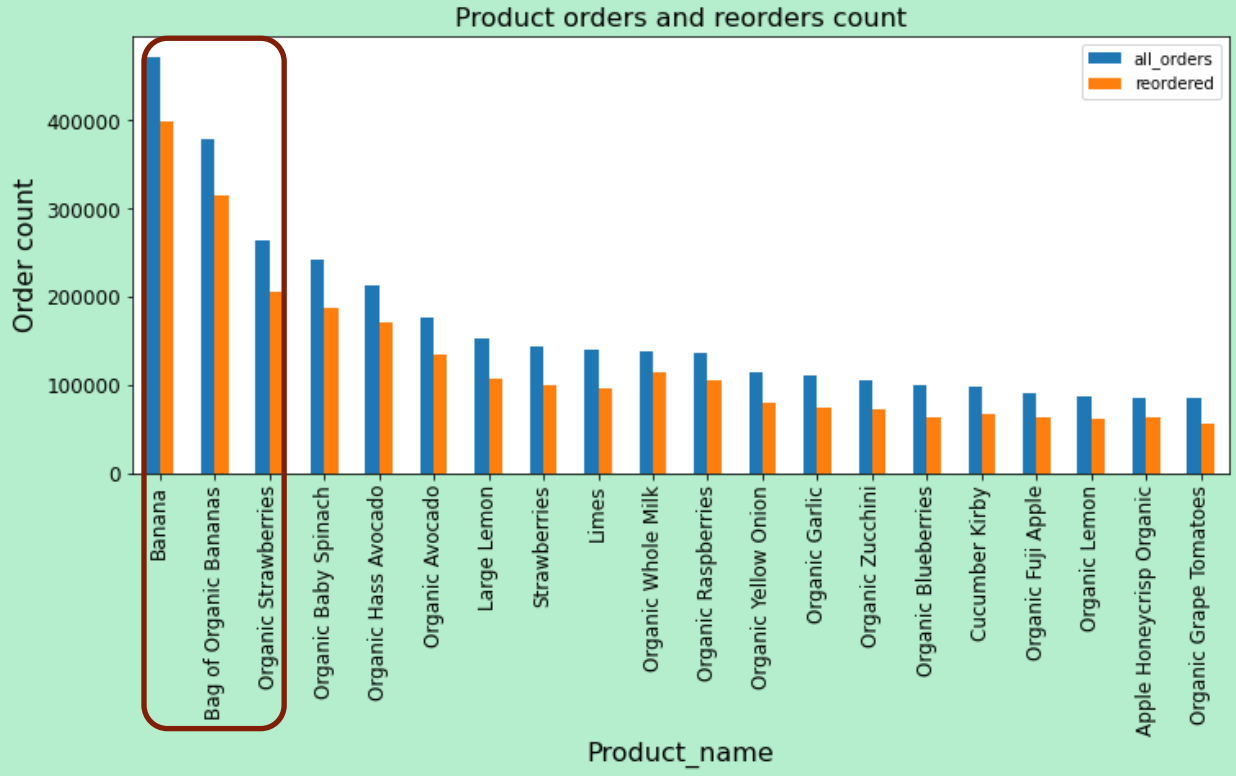
- Order count between 4-100
- Most users made few orders

Order count vs. day since prior order



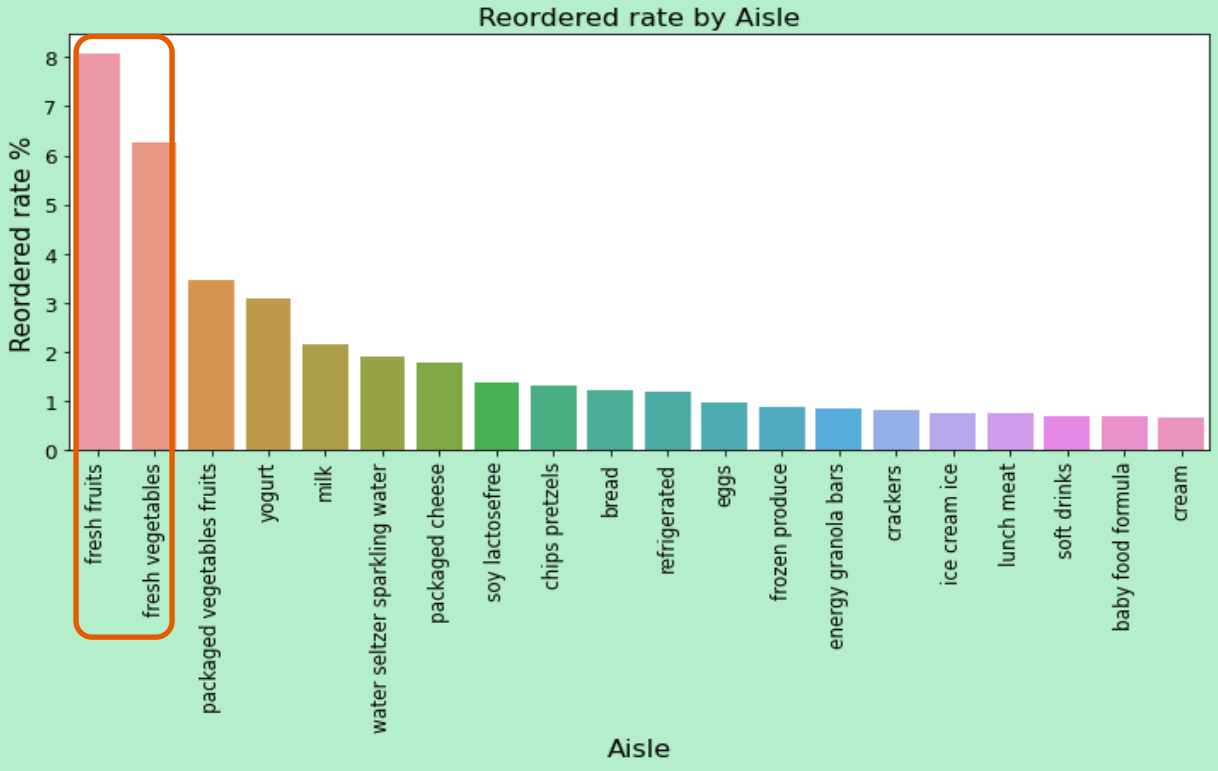
- Made orders weekly (every 7 days) and monthly (every 30 days).
- Similar observation for reorder count

Product orders and reorders count (Top20)



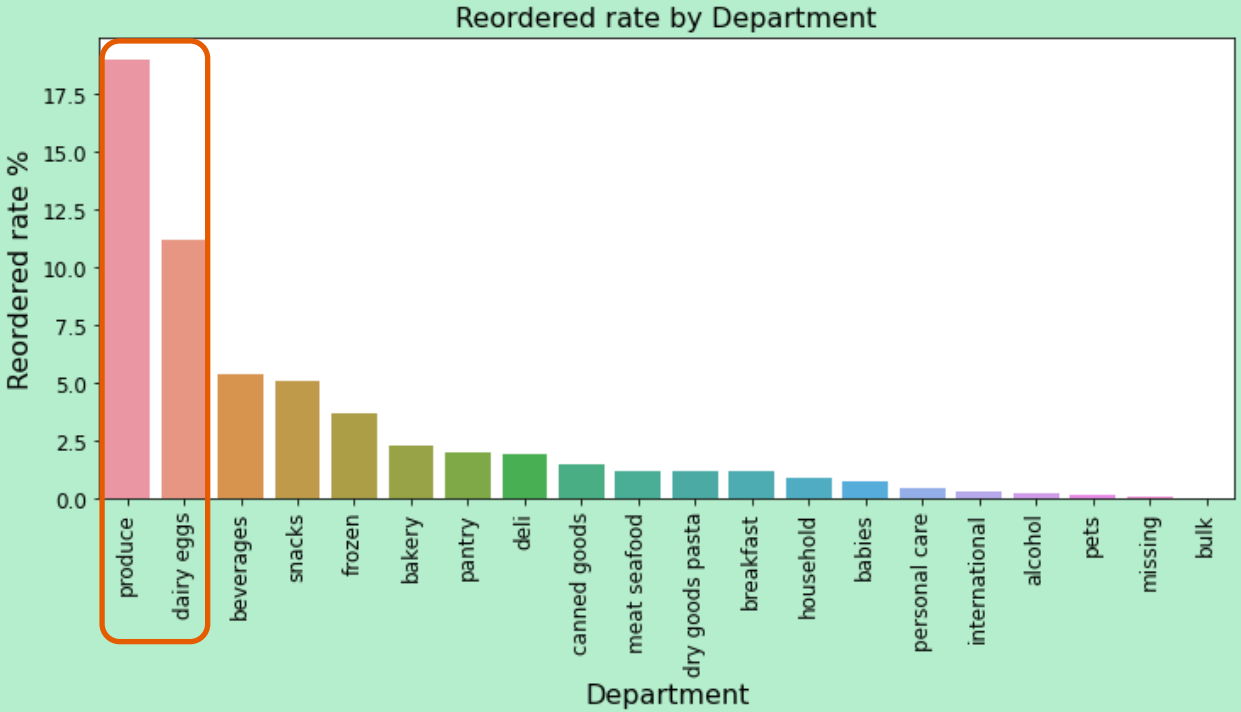
- **Banana** is the most popular product
- 15 out of top20 **organic** products
- Similar pattern for reorder count

Reordered rate by aisle (Top20)



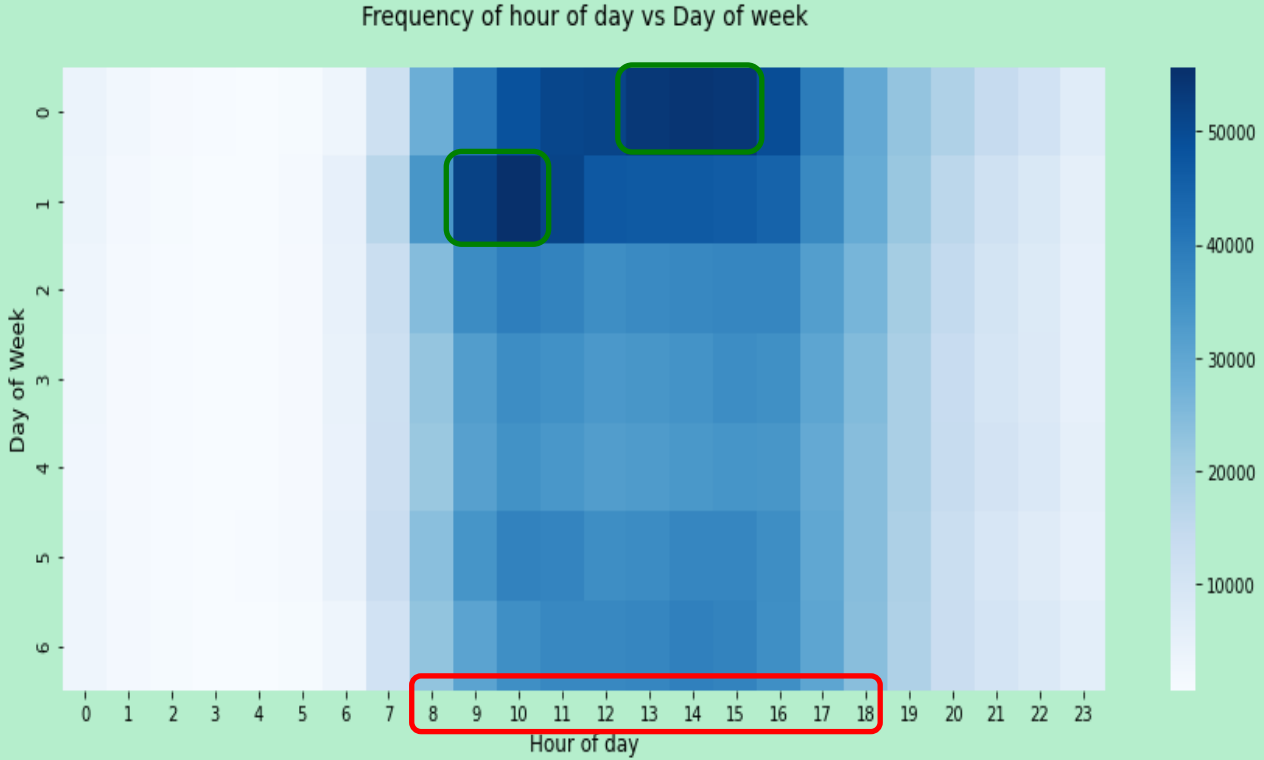
- **Fruits** and **vegetables** contributes to almost 15% of total reorders

Reordered ratio by department (Top20)



- **produce** and **dairy eggs** contribute to nearly 30% of total reorders

Daily and hourly order count heatmap



- day0 between 13-15 hour
- day1 between 9-11 hour
- Most of the orders on any day between 8-18th hour

02.

Featurization and Modeling

Featurization

Users

- user_reorder_ratio
- user_avg_prd
- mean_day_since_order
- user_aisle_count
- User_dept_count

Products

- prd_count
- prd_reordered_ratio
- prd_cart_order
- prd_avg_cart_position

User-Product

- user_prd_count
- user_prd_reorder_ratio
- user_prd_first

Last 5 orders

- order_count_last5
- reorder_count_last5
- reorder_rate_last5

Modeling : baseline f1 score

- Imbalanced target data : p(reordered=1)=0.0978

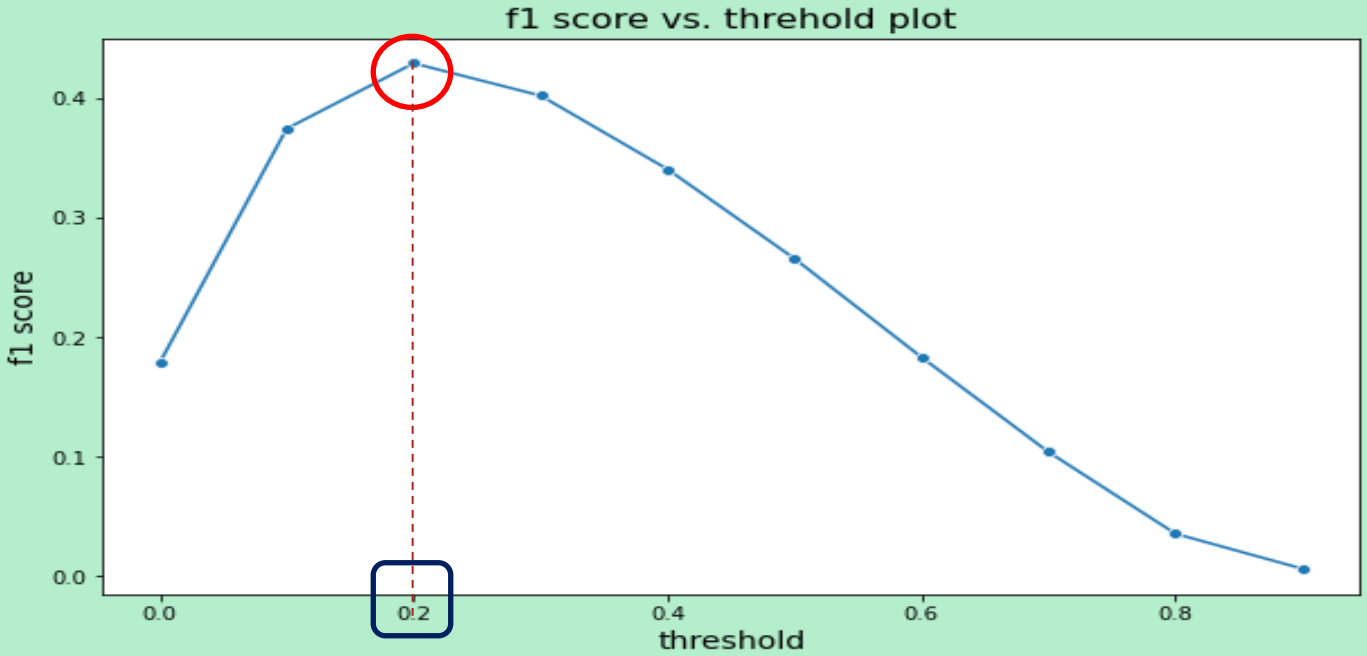
- Evaluation Metric : f1 score

- Baseline:

  Assumption: predict all as positive (reordered=1)
  precision = (true positive)/(total predict positive) = 0.0978
  recall = (true positive)/(total actual positive) = 1
  baseline f1 score=(2 * precision * recall)/(precision + recall) = **0.178**

Modeling : optimal threshold

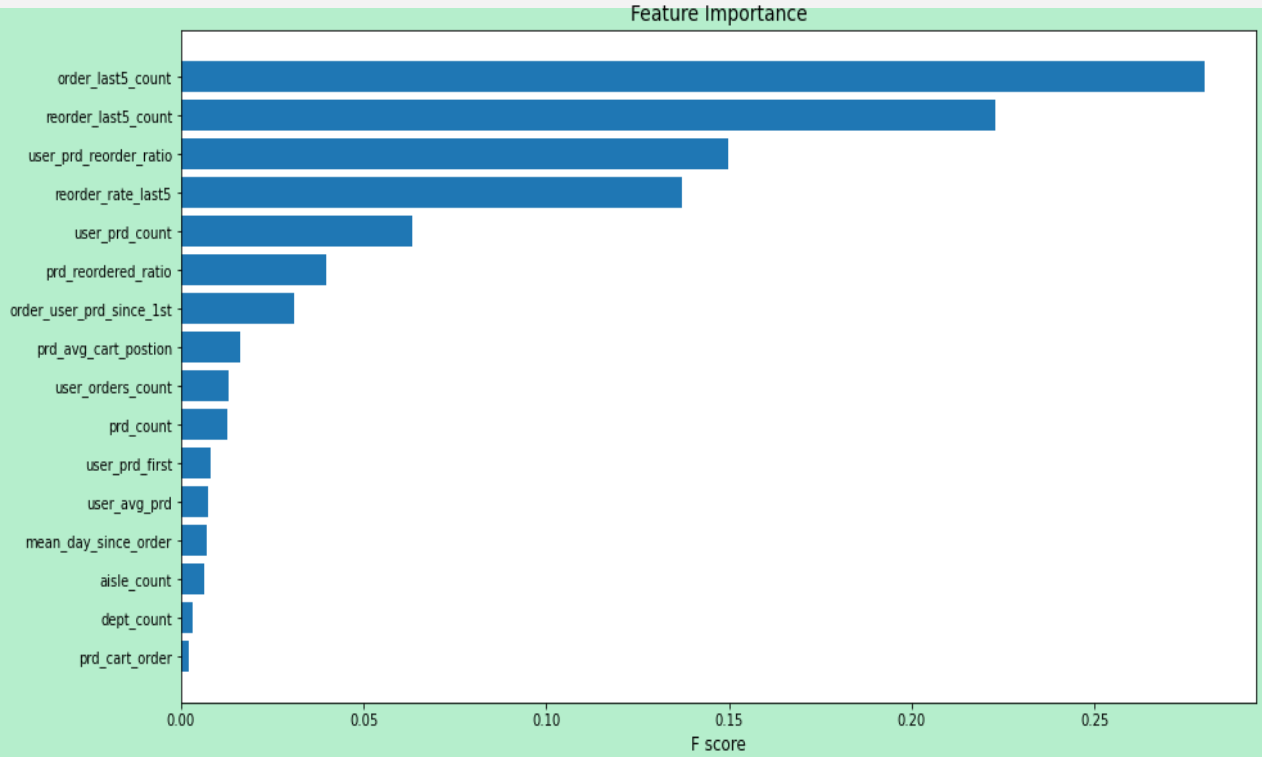
- Optimal threshold is **0.2** (with maximum f1 score)



Summary of Model score (baseline f1 score 0.178)

Classifier	Model technique	Hyper-params	Optimal Threshold	Train f1 score	Val f1 score	Test Kaggle f1 score
Logistic Regression	Basic model	{'logreg__C': 0.1}	0.20	0.427	0.429	0.363
XGBoost	Boosting-Based	{'learning_rate': 0.02, 'max_depth': 6, 'n_estimators': 1000}	0.21	0.434	0.433	0.365
Lightgbm	Boosting-Based	{'learning_rate': 0.1, 'max_depth': 3, 'num_iterations': 500}	0.22	0.431	0.432	0.365
Random Forest	**Bagging-Based**	**{'max_depth': 12, 'n_estimators': 1000}**	**0.21**	**0.437**	**0.431**	**0.368**

Top features (Random Forest)



- Last5 order and user_product strongest features
- Last5 order features – recent preference
- User_product features – long term preference

03.

Conclusion and Recommendation

Conclusion and Recommendation

Conclusion:

- Model can predict product reorder in next purchase effectively, f1 score 0.368
- Understand customer recent/long-term preference

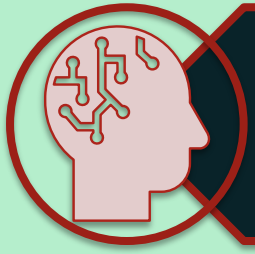
Recommendations:

- Put more effort on analysis about latest orders and user_product features
- Customer retention strategies based on predicted result:


Offer personalized discounts and credits

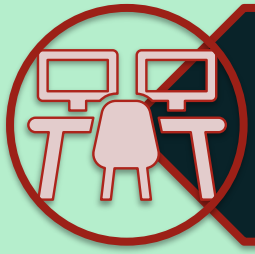

Send engaging emails to customers

Further Work



Further improvement for current model:

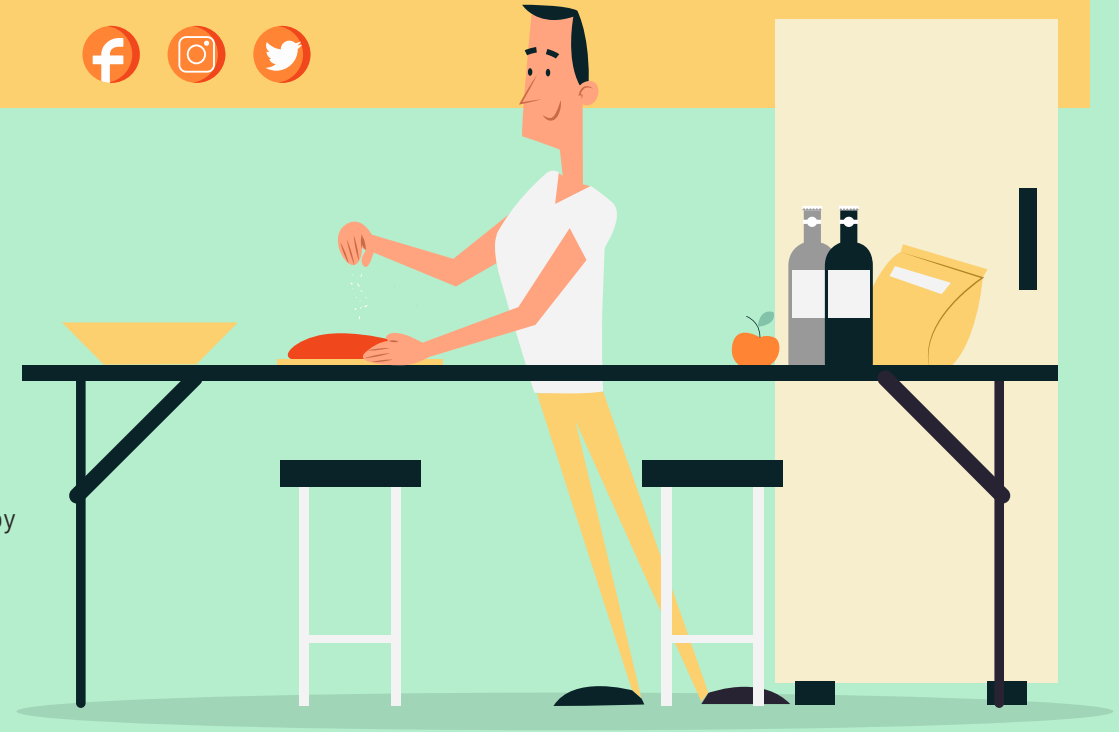
- Explore and create more features
- Varying threshold for users



Future work:

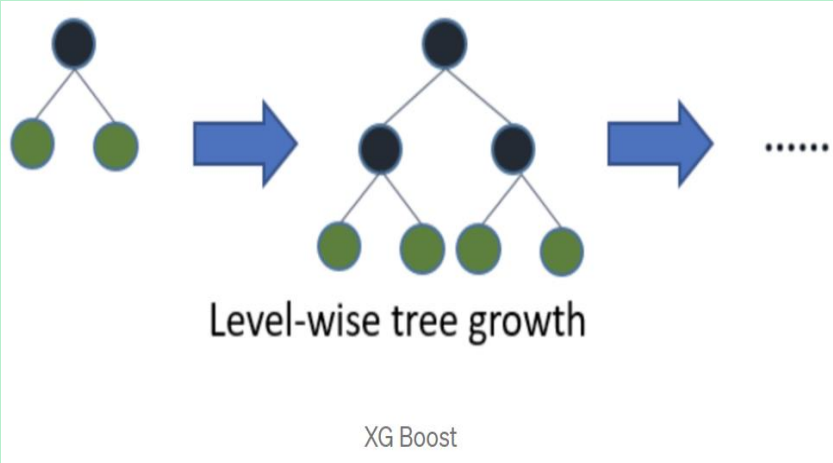
- Association rule approach
  ‘toothbrush’-->’toothpaste’

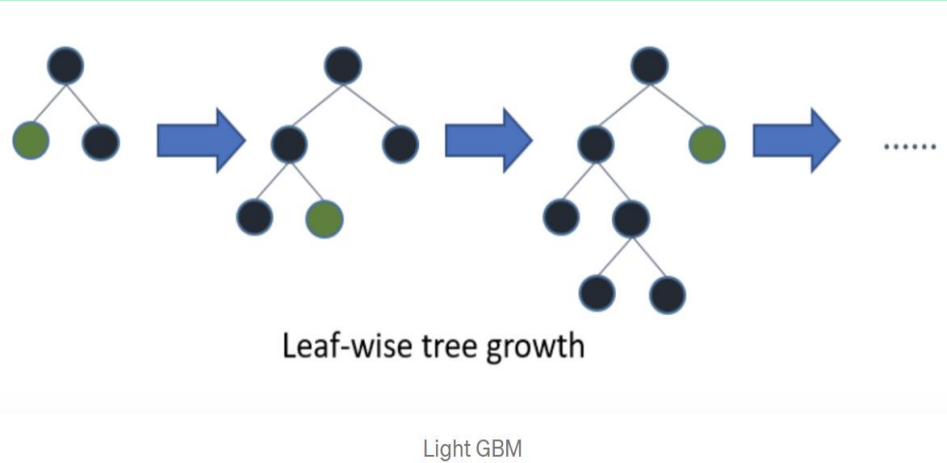
THANKS!



CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik

Backup


Level-wise tree growth
XG Boost


Leaf-wise tree growth
Light GBM