Instacart Market Basket Analysis

Introduction

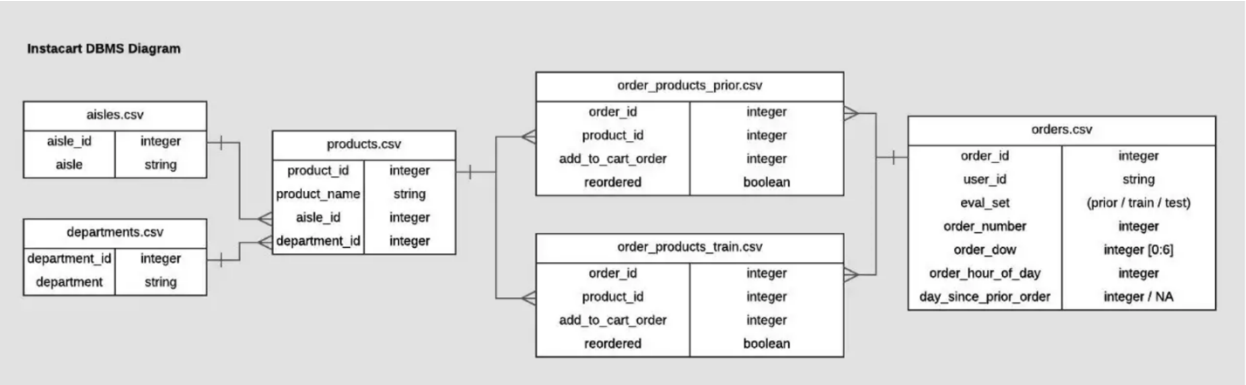
Have you ever wondered how do companies like Amazon, Netflix or Walmart suggest products that you like in the web browser? Or, have you ever gone to the grocery stores and seen a promotion: 'Buy 2 items for $10'? Intrigued by those recommendations and promotions, there are some questions need to be addressed. How do they understand our preference? Why and how do they promote buying a number of items for a deal? Also, how do they manage their inventory in a way that that a product is readily available when a customer wants to buy it?

With recommender systems, we can answer the above questions. A little bit back of history from Wikipedia: in 1979, the first recommender systems Grundy was created by Elaine Rich. It was a system that asked the user specific questions and then gave the user a recommendation based on user's stereotypes. Recommender systems are developed a lot since then.

Dataset

The dataset is taken from Kaggle, Instacart Market Basket Analysis. Instacart is an online grocery ordering through website and mobile app. It allows customer to choose their products online and deliver them to their house or have them picked up in-store. In Kaggle competition, Instacart provided anonymous data to analyzed customer's pattern and to predict which products would likely be reordered based on the customer's past purchases. There were 6 csv files:

- **aisles.csv** has 134 rows and 2 columns
- **departments.csv** has 21 rows and 2 columns
- **products.csv** has 49688 rows and 4 columns
- **orders.csv** has 3M+ rows and 7 columns
- **order_products__prior.csv** has 32M+ rows and 4 columns
- **order_products__train.csv** has 1M+ rows and 4 columns



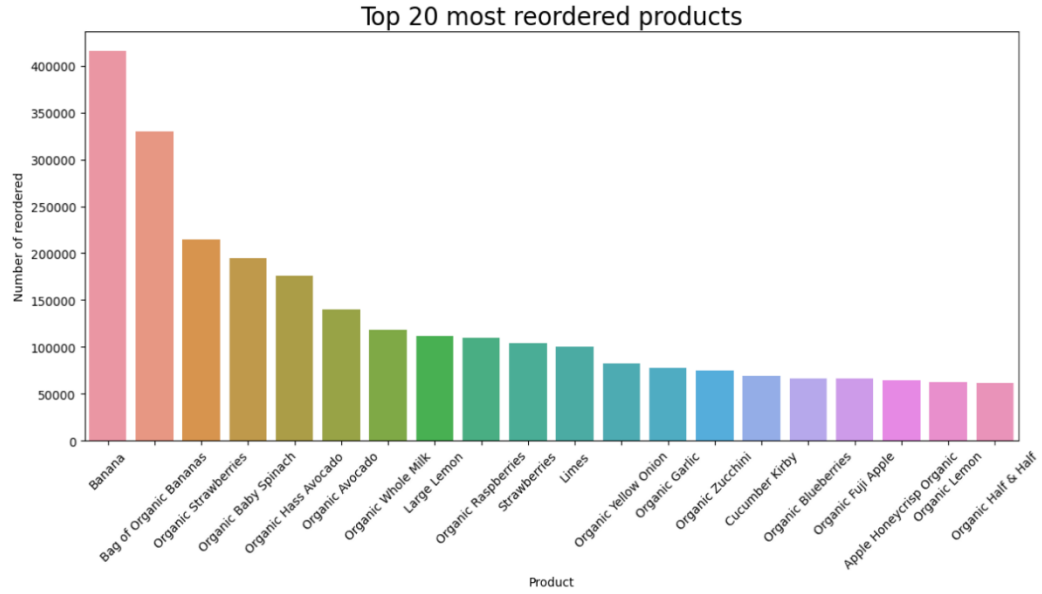
Summary of Cleaning and Preprocessing

The data is clean and has no duplicates, but has some missing value in the 'days_since_prior_order' column from orders.csv file, but it is supposed to be like that to indicate first time buyer. The data has also 3 evaluation sets which contained *prior*, *train* and *tes*t. *Prior* is a data from customer prior history of customer purchases. *Train* is the suggested data to train. *Test* is used in the competition to fill in some null value. For this Capstone Project, I would

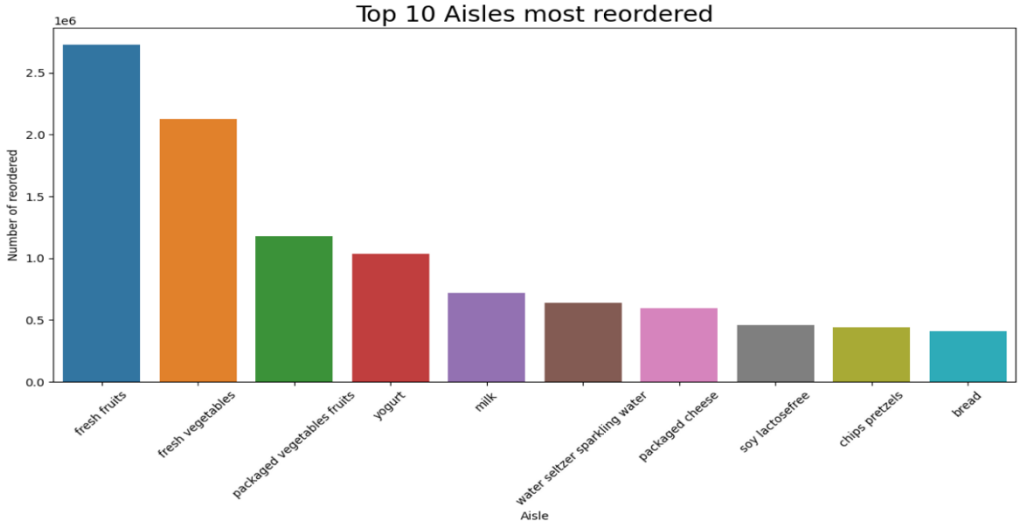
not use the *test* evaluation set. Due to computational limitation, I limit the amount of data to the last 5 transactions. In addition, I perform some feature engineering to calculate the probability of an item bought by a user as a rating since we do not have any rating on the product that is needed to do a collaborative filtering.

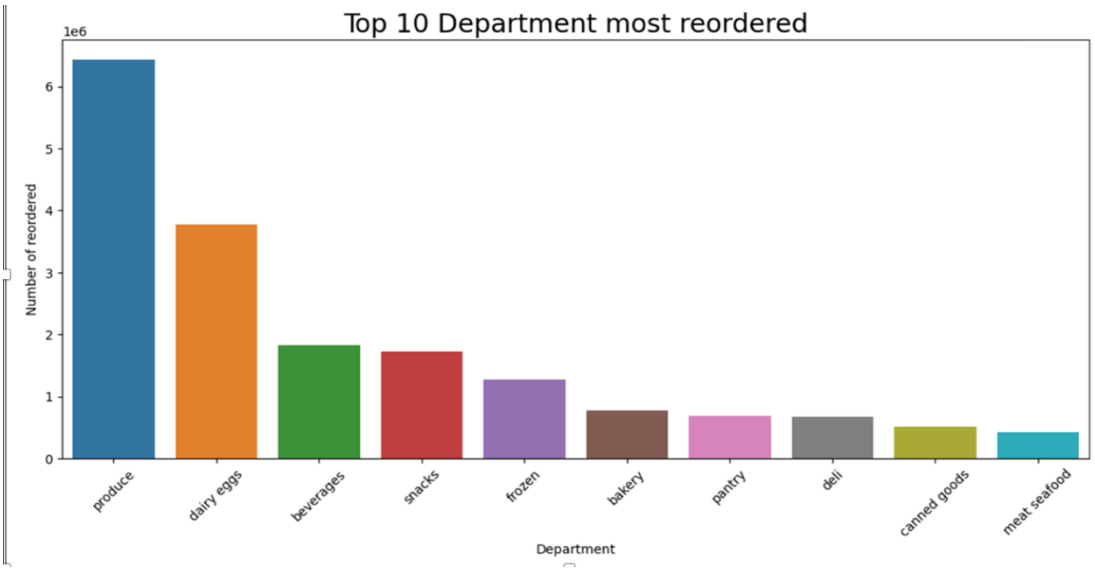
Insights, Modeling, and Results

In this project, we are interested in what products tend to be bought together, what kinds of new products can be recommended to the user and what is the prediction of customer next order. From the exploratory data analysis, we know that the top products reordered are fruits and vegetables.



It is also reflected in aisles and department where fruits and vegetables aisles, and produce department are reordered the most.





Looking at those graphs, I want to know which products tend to be bought together. Market basket analysis is a data mining technique used by retailers to uncover patterns of customer's purchases. It measures the probability of product A being bought together with product B.

First product bought	Second product bought
Oh My Yog! Pacific Coast Strawberry Trilayer Yogurt	Oh My Yog! Organic Wild Quebec Blueberry Cream Top Yogurt & Fruit
Mighty 4 Kale, Strawberry, Amaranth & Greek Yogurt Tots Snack	Mighty 4 Sweet Potato, Blueberry, Millet & Greek Yogurt Tots Snack
Unsweetened Blackberry Water	Raspberry Essence Water
Mighty 4 Sweet Potato, Blueberry, Millet & Greek Yogurt Tots Snack	Mighty 4 Kale, Strawberry, Amaranth & Greek Yogurt Tots Snack
Cream Top Strawberry on the Bottom Yogurt	Cream Top Blueberry Yogurt

Based on the above table, the customers like to buy variety flavor of the yogurt, toddler' snacks and water. The company could have a marketing strategy 'buy 3 items for $5' or 'Buy One Get One' to boost sale. I also measured the probability of aisles from which the customers tend to buy at one transaction. The customers like to buy from fresh herbs, fresh vegetables, packaged vegetables fruits and fresh fruits aisles. It shows that those aisles are frequently associated with each other.

First aisle	Second aisle
fresh herbs	fresh vegetables, packaged vegetables fruits
fresh vegetables, packaged vegetables fruits	fresh herbs
fresh herbs	fresh vegetables, fresh fruits
fresh vegetables, fresh fruits	fresh herbs
fresh herbs, packaged vegetables fruits	fresh vegetables

To recommend products that has not been bought by the customer, I perform collaborative filtering. Collaborative filtering is a technique that finds the similarity between users and items in order to recommend new products to a user. The advantage of collaborative filtering is no need of domain knowledge as it only needs information about user

id, product id, and the rating that users give to items. The disadvantage of collaborative filtering is it cannot handle cold-start problem. Cold-start problem is when you have a new user id, you don't have any prior information to train so it will not give you an accurate result. But in this case, it is not a problem since we have prior history of the customer. The models used are Funk Singular Value Decomposition (FunkSVD), Alternating Least Square (ALS), and Neural Network. FunkSVD and ALS works using matrix factorization which it decomposes high dimensional user-item matrix into low dimensional user matrix and item matrix to get latent factors representations for user and item. Matrix factorization then use latent factors to give recommendation. Neural network works in a similar way where it maps user and item into two separate embedding matrices, computes the dot product between use vector and movie vector to obtain a score and trains those matrices to give recommendation. From all of the models used, neural network collaborative filtering is the best model since it has the best accuracy.

Model	Metric	Accuracy
FunkSVD	FCP [1]	0.245
ALS	RMSE[2]	0.169
Neural Network	Binary accuracy	0.6165

Classification models (xgboost, random forest and neural network) are used to predict what products will be reordered next. They are working by classifying dataset to 1 or 0. 1 is 'they will reorder' and 0 is 'they will not reorder'. F1-score is the performance metric on which models will be evaluated. All of the model has similar f1-score and pretty balance recall and precision score, but xgboost has the highest f1-score.

Model	F1 score
Random forest	0.822
Xgboost	0.829
Neural network	0.827

Conclusions

Market basket analysis enables company identifying users' preference hence a company could do selective marketing. Recommender systems could assist customers making the right decision on buying products, saving their time searching for products and enhancing their shopping experience. Machine learning could benefit the company to predict what product the customer will reorder based on past purchases. Thus, the company could plan ahead their marketing strategy, manage their inventory, elevate customer experience and boost sales and revenue.

References

- Collaborative Filtering - Keras - https://keras.io/examples/structured_data/collaborative_filtering_movielens/
- Dataset - Kaggle - https://www.kaggle.com/competitions/instacart-market-basket-analysis/data
- Market Basket Analysis - https://pythondata.com/market-basket-analysis-with-python-and-pandas/
- Schema picture - https://suraj005.medium.com/instacart-market-basket-analysis-case-study-4ee911e0dd03
- Wikipedia - Recommender history https://en.wikipedia.org/wiki/Recommender_system

[1] FCP: Fraction of Concordant Pairs – fraction of pairs whose relative ranking order is correct.
[2] RMSE: Root Mean Square Error – a standard way to measure error of a model.