CSP 571: Project - Proposal & Outline

1 Members

Name	Hawk ID	Email ID
Shraddhaben Patel	A20499171	spatel174@hawk.iit.edu
Nevil Jack Denis	A20474215	psubramanian@hawk.iit.edu
Naga Surya Suresh	A20492550	nlnu1@hawk.iit.edu

2 Project Proposal

2.1 Formal description and stated research goal

Don't you hate it when you go shopping and forget to pick up something you meant to?

To address the relationship between what items to purchase and, as a result, to increase and improve the company's sales and comprehend consumer behavior This study's objectives are to identify patterns in consumer purchase behavior on Instacart and suggest product combinations that might be included in various promotions.

This project also aims to predict which previously purchased products will be in a user's future order by using anonymized data on customer orders over time.

2.2 Proposed methodology/approach

On a broad scale, the goal is to develop capabilities for reorder data for specific products as well as for user preferences for products. Additionally, we construct orders based on the users' orders.

For Example everyone enjoys an Apple, so the metrics for each product should reflect the quality of a product getting reordered on its own merits. The reorder metrics for Mr. A should account for this preference as well as the reordering measures that are unique to him because Mr. A like some unusual meal that is rarely enjoyed by anyone else. Other metrics based on orders might include ordering patterns, preferred times (day of the week/hour of the day), etc.

2.3 Set of metrics which will measure analysis results

For Market basket analysis we would be using support, confidence and lift to understand the association rules of the products and items.

For Association rules, based on the predicted data we will use Confusion matrices, F1-Score, Precision, Recall, sensitivity and AUC.

3 Project Outline

3.1 Data

Data Source: Kaggle data set

Dataset Description: The data set is a relational group of files that tracks the orders that customers place over time. An anonymous sample of more than 3 million grocery orders from more than 200,000 Instacart users make up the data set. Each user receives between 4 and 100 of their order details, including the order of the things they purchased in each order, the day and week it was placed, and the amount of time between orders.

Every entity (customer, item, order, aisle, etc.) has a unique id that is connected with it.

File Description:

`orders` (3.4m rows, 206k users):

- `order_id`: order identifier
- `user_id`: customer identifier
- `eval_set`: which evaluation set this order belongs in (see `SET` described below)
- `order_number`: the order sequence number for this user (1 = first, n = nth)
- `order_dow`: the day of the week the order was placed on
- `order_hour_of_day`: the hour of the day the order was placed on
- `days_since_prior`: days since the last order, capped at 30 (with NAs for `order_number` = 1)

`products` (50k rows):

- `product_id` : product identifier
- `product_name` : name of the product
- `aisle_id` : foreign key
- `department_id` : foreign key

`aisles` (134 rows):

- `aisle_id` : aisle identifier
- `aisle` : the name of the aisle

`deptartments` (21 rows):

- `department_id` : department identifier
- `department` : the name of the department

`order_products__SET` (30m+ rows):

- `order_id` : foreign key
- `product_id` : foreign key
- `add_to_cart_order` : order in which each product was added to cart
- `reordered` : 1 if this product has been ordered by this user in the past, 0 otherwise

3.2 Data Processing

We perform various exploratory data analysis to understand the data. For better processing of data we convert various character variables to Factors. The factor conversion is done on orders, products, aisles, department tables. This helps us understand how the items are sorted in the market by department, aisle and product itself.

Next we perform analysis and processing to see what items/products are stored in terms of departments and aisles. We gather data and visualize the top products in each of the fields. We also make sure to check the bottom list of products that is being offered.

We then perform in-depth EDA of the orders table, from this table we tend to derive orders that are being placed in various time of the hours, days and week. Understanding this data helps us to get the trends in the purchasing habits.

3.3 Model Selection

We will study the data to perform association analysis using MBA. There are several uses for market basket analysis, such as cross-selling, product placement, affinity marketing, fraud detection, and consumer behavior. We will use Apriori algorithm for mining association rules and make a comparison with Frequent Pattern Growth Algorithm.

The user id and product id serve as the keys of a denormalized structure that is formed after the features are created. The issue then transforms into a classification problem that requires a classifier algorithm to be used to solve. XGBoost is the classifier that we have selected.

3.4 Tools

- **Software Packages** : RStudio, R.
- **Development**: GitHub, Notion.
- **Project Planning:** Excel, Notion Kanban boards.
- **Libraries**: data.table, dplyr, ggplot2, knitr, stringr, DT, magrittr, grid, gridExtra, , sqldf, Matrix, arules, tidyr, arulesViz, methods data.table, xgboost.

3.5 Literature review and related work

Transaction data is a set of recording data in connection with purchase and sale of commodities. Nowadays these data are widely used as a research objects in the sense of discovering new information. Some of the applications that recommends the associated products to the customer are Amazon, Movie Ticket Bookings, etc.

In our model using the association method we try to recommend the best product that the customer could get that is related to their previous purchase. There are various algorithms available to perform MBA. The one we are using is XGBoost supervised learning algorithm.

Some reference materials we found useful:

- Relevant Papers: [1],[2],[3],[4],[5],[6],[7],[8]

References

[1] Fachrul Kurniawan , Binti Umayah, Jihad Hammad, Supeno Mardi Susiki Nugroho and Mochammad Hariadi, “Market Basket Analysis to Identify Customer

Behaviors by Way of Transaction Data" Knowledge Engineering and Data Science (KEDS) -Vol 1, No 1, January 2018, pp. 20–25.

[2] Manpreet Kaur ,Shivani Kang. " Market Basket Analysis: Identify the changing trends of market data using association rule mining" , International Conference on Computational Modeling and Security (CMS 2016), Procedia Computer Science 85 (2016) 78 – 85.

[3] Verma Sheenu, Bhatnagar Sakshi. "An Effective Dynamic Unsupervised Clustering Algorithmic Approach for Market Basket Analysis". International Journal of Enterprise Computing and Business Systems 2014:4(2).

[4] Alfiqra and A U Khasanah, "Implementation of Market Basket Analysis based on Overall Variability of Association Rule (OCVR) on Product Marketing Strategy", IOP Conference Series: Material Science and Engineering, Volume 722, 3rd International Conference on Engineering Technology for Substantial Development (ICET4SD) 23-24 October 2019, Yogyakarta,Indonesia.Citation Alfiqra and A U Khasanah 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* 722 012068

[5] Andrej Trnka- Department of Applied Informatics, University Ss. Cyril and Methodius, Trnava, Slovakia. "Market Basket Analysis with Data Mining methods" 10.1109/ICNIT.2010.5508476, Publisher-IEEE, 11-12 June 2010 International Conference on Networking and Information Technology.

[6] M. Kebisek and M. Elias, "The possibility of utilization of knowledge discovery in databases in the industry", *Annals of MTeM for 2009 & Proceedings of the 9th International Conference Modem Technologies in Manufacturing*, pp. 139-142, 2009 October 8-10, 2009.

[7]A. Herman, L.E. Forcum, Joo Harry. Using Market Basket Analysis in Management Research, Journal of Management, 39 (7) (2013), pp. 1799-1824.

[8] A.A. Raorane, R.V. Kulkarni, B.D. Jitkar, Association Rule – Extracting Knowledge Using Market Basket Analysis, Research Journal of Recent Sciences, 1 (2) (2012), pp. 19-27.