ISM6136 Data Mining- Final project Group 11:

Nutrition(flavor) and ratings-based Food Recommendation system

Kritika Jajoo

Sai Sujitha Reddy Mullangi

Vatsal Borda

NUTRITION(FLAVOUR) & RATINGS BASED FOOD RECOMMENDATION SYSTEM

Background

Offering online personalized recommendation services helps to improve customers' satisfaction and needs. Conventionally, a recommendation system is considered as a success if customers purchase the recommended products. However, the act of purchasing itself does not guarantee satisfaction and a hugely successful recommendation system should be one that aids the customer's choice decision in today's market with numerous alternatives. In this project, we build the recommendation system based on matchbox filtering. Two models are tested: K-means clustering and Matchbox filtering.

Personalization of product information has become one of the most key factors that impact a customer's product selection and satisfaction in today's competitive and challenging market. Personalized recommendation requires firms to understand customers and offer food and services that meet their requirements. Successful firms are those provide the right products to the right customers at the right time and for the right price. The ratings and reviews are very influential with consumers who have the same tastes, so consumers get the same recommendations with other consumers with the same taste. Recommendation systems are decision aids that analyze customer's prior online behavior and present information on products to match customer's preferences. Through analyzing the customer's food preferences or communicating with them, recommendation systems employ quantitative and qualitative methods to discover the food products that best suit the customer.

Motivation for solving a problem

Working in dining services at hub we have noticed confused faces of students with burden of N number of choices in front of them. Most of them are bored with what usual they have tasted on other hand hesitant to try out untasted as they might not like. If students with meal coupons (They can eat whatever they want unlimited!) are pondering around what to eat, it is not hard to infer the crucial bottleneck of food businesses especially those who have lengthy menu. Where customer must bear the cost too of just trying out new. Additionally, it is being observed that people eat their usual food item with regularity with certain amount of gap of time rather than try new until they get bored. So, recommending food items using Machine Learning restaurants/food delivery business gives variability of dishes which customers likely to find tasty will enrich consumer experience and will boost restaurant’s existing revenue generated from each customer.

Solution Methodology

We have used McDonald’s menu dataset https://www.kaggle.com/mcdonalds/nutrition-facts to build a recommender system for customers. Plus, one dummy dataset of user ratings exclusively for matchbox algorithm along with nutritional dataset of menu. Overall, nutrition dataset contains 15 attributes which are used most of the time for building a recommendation system as those determines the taste.

Like most other recommendation systems, ours uses two components to form the input, the first one being a curated database of food items containing their nutritional values. The second component is the set of user profiles for which the recommendations would be generated. This was constructed by obtaining user reviews for the food items. User reviews help us understand a user’s preferences, which in turn helps build user profiles for recommending food items.

Two algorithms were used in the implementation of this classification solution – K means Clustering and Matchbox Algorithm. K- Means Clustering suggests an approach wherein the dishes are clustered into distinct groups and recommendations are made for the group (i.e., sweet, spicy, sour, umami) effectively eliminating the personalized recommendation component of the recommendation system. In clustering nearest neighbor of dish user likes might be the recommendation.

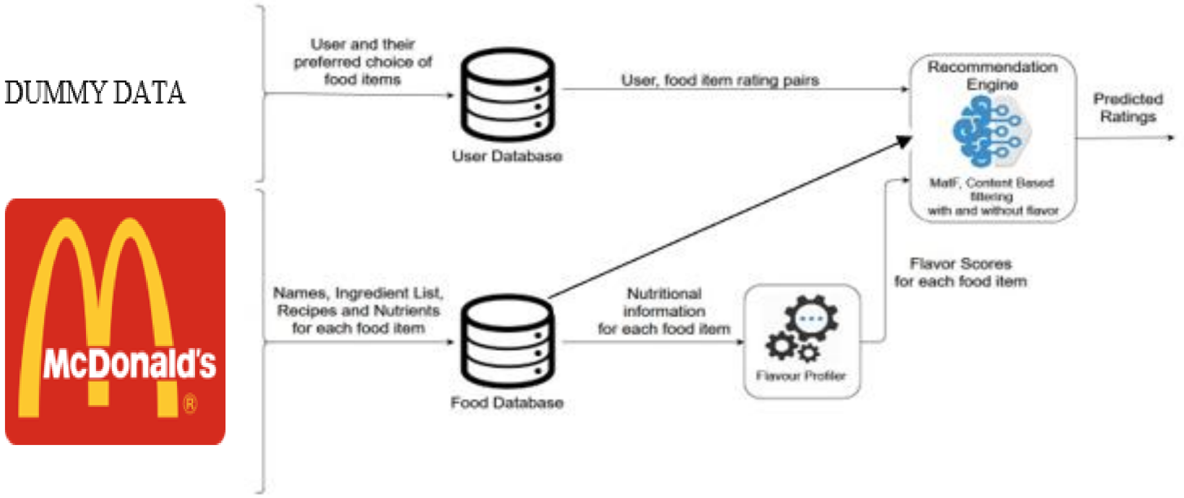
The Matchbox recommender combines collaborative filtering with a content-based approach. It is therefore considered a **hybrid recommender**.

Collaborative Filtering: The Collaborative Filtering approach for recommendation looks to make predictions regarding a user’s preference by collecting preferences from multiple similar users. The assumption in Collaborative Filtering is that people who view and evaluate items in a similar fashion are likely to assess other food dishes in a comparable manner. Matrix Factorization is a Collaborative Filtering algorithm that takes as input a User-Item Rating Matrix. This matrix is sparse since it is not likely that a user has rated all dishes in the food database. The approach aims to break down the User-Item matrix into two matrices of latent user and item representation. The intent of this approach is to reform the original User-Item matrix while filling in the missing ratings. When making recommendations for a particular user, the Collaborative Filtering Algorithm only considers other similar users. It does not consider the content or features of an item; hence, food flavor cannot be incorporated when using this method to make recommendations.

Content Based Filtering: Content-Based Recommender systems stem from the idea of using the content, properties or description of an item for recommendation purposes. Items are described with a set of descriptor terms or tags which form the basis for item-based comparison.

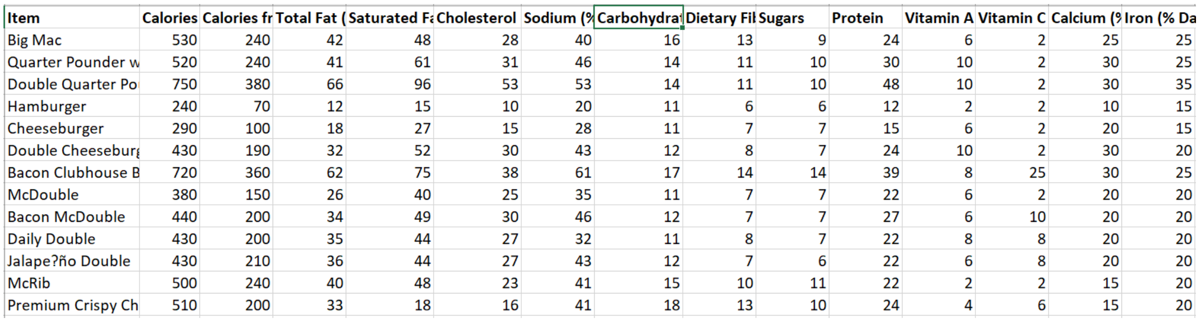
The **Train Matchbox Recommender** module reads a dataset of user-item-rating triples and, optionally, some user and item features. It returns a trained Matchbox recommender. You can then use the trained model to generate recommendations, find related users, or find related food items, by using the Score Matchbox Recommender module.

Most impressive USP of Matchbox recommender is it combines both filtering in such a way that it eliminates famous ‘cold start’ problem by putting threshold of having enough data points of user ratings before executing collaborative filtering as it performs even better than content-based filtering if it has enough data of user, else it utilizes content-based filtering in absence of enough user ratings.


DUMMY DATA
McDonald's
User and their preferred choice of food items
User Database
User, food item rating pairs
Recommendation Engine
MatF, Content Based filtering with and without flavor
Predicted Ratings
Names, Ingredient List, Recipes and Nutrients for each food item
Food Database
Nutritional information for each food item
Flavour Profiler
Flavor Scores for each food item

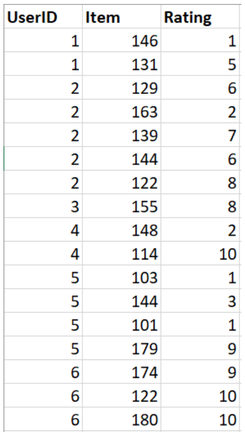
Dataset Description:

Food dataset



Item	Calories	Calories fr	Total Fat (	Saturated Fa	Cholesterol	Sodium (%	Carbohydrat	Dietary Fil	Sugars	Protein	Vitamin A	Vitamin C	Calcium (%	Iron (% Da
Big Mac	530	240	42	48	28	40	16	13	9	24	6	2	25	25
Quarter Pounder w	520	240	41	61	31	46	14	11	10	30	10	2	30	25
Double Quarter Po	750	380	66	96	53	53	14	11	10	48	10	2	30	35
Hamburger	240	70	12	15	10	20	11	6	6	12	2	2	10	15
Cheeseburger	290	100	18	27	15	28	11	7	7	15	6	2	20	15
Double Cheeseburg	430	190	32	52	30	43	12	8	7	24	10	2	30	20
Bacon Clubhouse B	720	360	62	75	38	61	17	14	14	39	8	25	30	25
McDouble	380	150	26	40	25	35	11	7	7	22	6	2	20	20
Bacon McDouble	440	200	34	49	30	46	12	7	7	27	6	10	20	20
Daily Double	430	200	35	44	27	32	11	8	7	22	8	8	20	20
Jalape?ño Double	430	210	36	44	27	43	12	7	6	22	6	8	20	20
McRib	500	240	40	48	23	41	15	10	11	22	2	2	15	20
Premium Crispy Ch	510	200	33	18	16	41	18	13	10	24	4	6	15	20

User dataset



UserID	Item	Rating
1	146	1
1	131	5
2	129	6
2	163	2
2	139	7
2	144	6
2	122	8
3	155	8
4	148	2
4	114	10
5	103	1
5	144	3
5	101	1
5	179	9
6	174	9
6	122	10
6	180	10

The McDonald’s dataset consists of 91 menu items and 14 variables. The independent variables in the dataset are calories, calories by fat, total fat, saturated fat, cholesterol, sodium, carbohydrates, dietary fiber, sugars, protein, vitamin A, vitamin C, calcium, and iron.

Following is the description of each variable:

1. **Calories:**
   A calorie is a variable which is a unit of energy. When you hear something contains 100 calories (about 8 minutes of running), it is a way of describing how much energy your body could get from eating or drinking it**.**

2. **Calories by fat:**
   Calories by fat is a variable which can be defined as calories that can come from fat as a source. Fat can be a substantial source of calories in food,

as it has more than double the number of calories per gram as protein and carbohydrates.

3. **Total fat (% daily value):**

   Total fat is a variable which indicates how much fat is in a single serving of a food per daily value percentage of fat, based on a 2,000-calorie diet. The current daily value (DV) for fat is 78 g.

4. **Saturated fat (% daily value):**

   A fat that contains only saturated fatty acids, is solid at room temperature, and comes chiefly from animal food products.

5. **Cholesterol (% daily value):**

   Dietary cholesterol refers to cholesterol that enters the body through foods such as red meats, eggs, and fatty dairy products.

6. **Sodium (% daily value):**

   Sodium is a variable which is a mineral that occurs naturally in many of the foods we eat. Sodium is essential for maintaining fluid balance in the body, for transmitting nerve signals, and for helping your muscles contract and relax. The recommended maximum intake of sodium is 2,300 milligrams a day.

7. **Carbohydrates (% daily value):**

   Carbohydrates is a variable which is a measure of a nutrient which provides the body with glucose, which is converted to energy used to support bodily functions and physical activity. The current daily value (DV) for carbohydrates is 275 g.

8. **Dietary Fiber (% daily value):**

   Fiber is a variable which is a type of carbohydrate that the body can't digest. Though most carbohydrates are broken down into sugar molecules, fiber cannot be broken down into sugar molecules, and instead it passes through the body undigested. Fiber helps regulate the body's use of sugars, helping to keep hunger and blood sugar in check. The recommended amount of fiber per day is 20 to 35 grams per day for adults.

9. **Sugars:**

   Sugars are a variable which are carbohydrates. Like all carbohydrates, they provide a source of energy in our diet. Sugar is a term that includes all sweet carbohydrates, although the term most often is used to describe sucrose or table sugar, a ‘double sugar’.

10. **Protein:**

    Protein is a variable which provides calories, or “energy” for the body. Each gram of protein provides 4 calories.

11. **Vitamin A (% daily value):**

    Vitamin A is a variable which is the generic term for a group of fat-soluble compounds important for human health. They are essential for many processes in your body.

12. **Vitamin C (% daily value):**

    Vitamin C is a variable which represents a nutrient that is a water-soluble vitamin that is found in many foods, particularly fruits and vegetables. The current daily value (DV) for vitamin C is 90 mg.

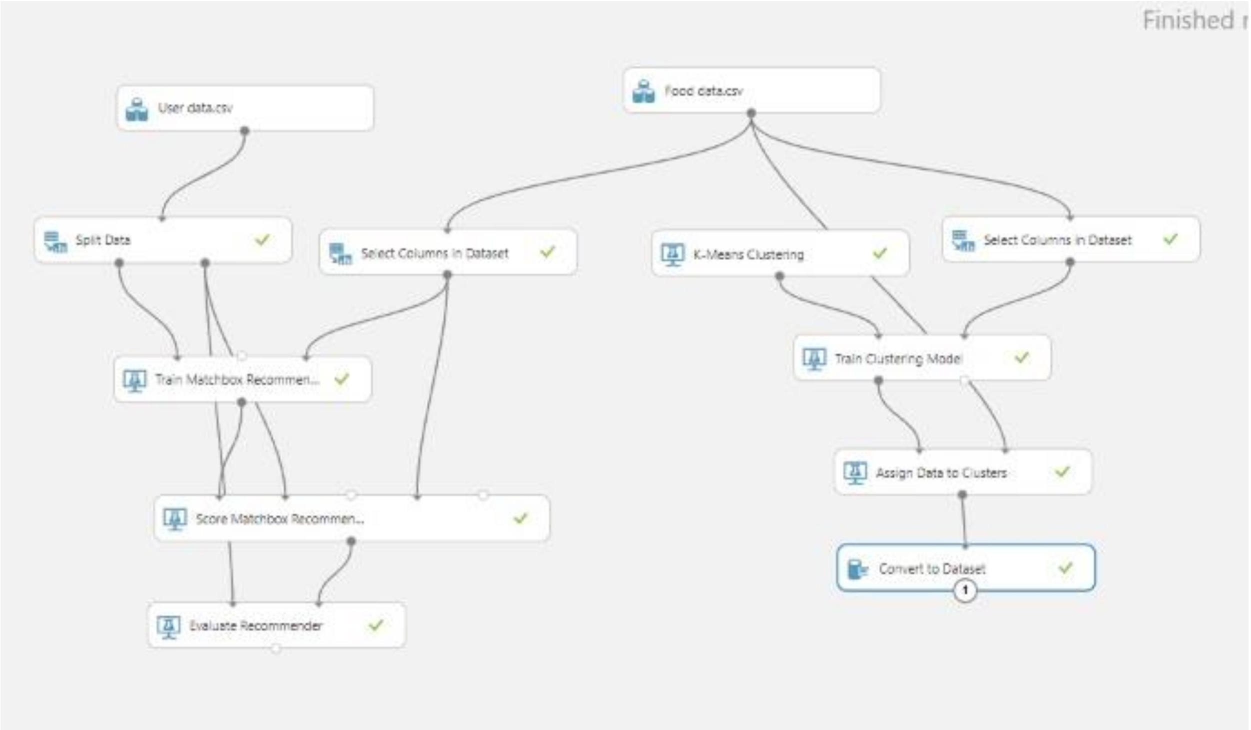
13. **Calcium:**

    Calcium is a variable which can be defined as a nutrient which can be found in a variety of foods, including dairy products, leafy vegetables, fish etc. The current daily value (DV) for calcium is 1300 mg.

14. **Iron:**

    Iron is a variable which can be defined as a nutrient which can be found in animal foods such as meat, seafood and poultry that provide both types and are better absorbed by the body. The current daily value (DV) for iron is 18 mg.

Algorithm Comparison:

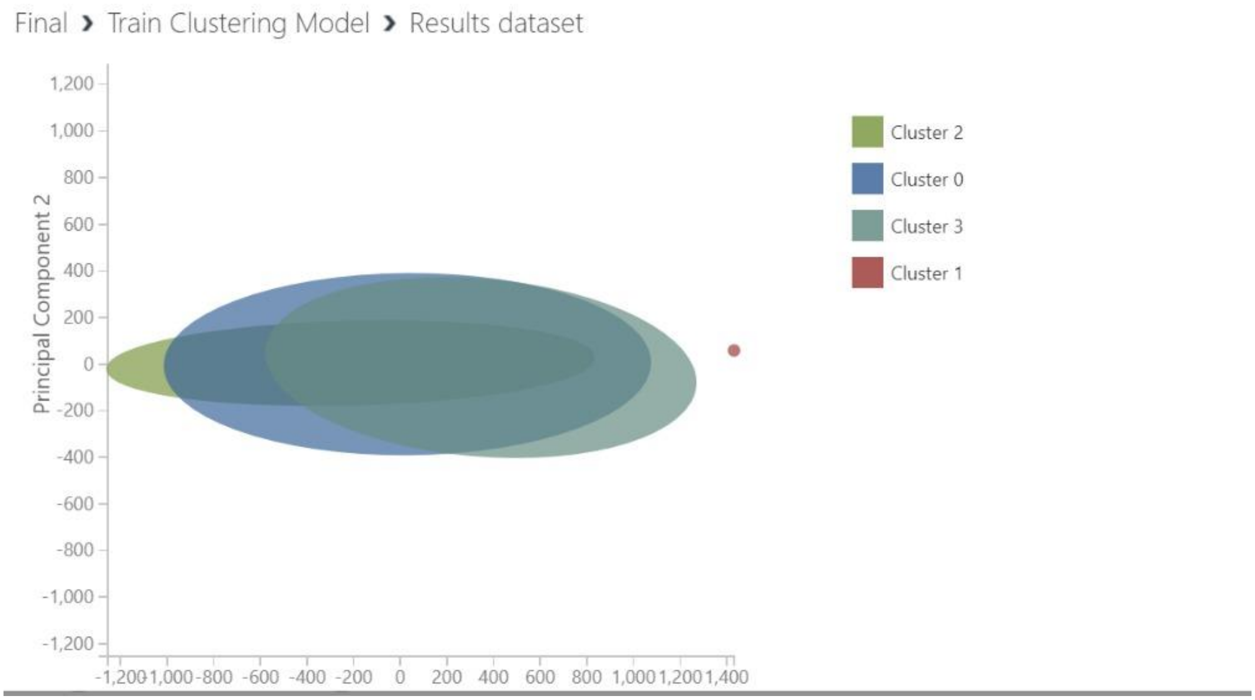


K means clustering Algorithm:

- First, we used K-means Clustering algorithm to classify food items based in their relevant nutritional information.
- Parameter mode:

  (I) Small number of centroids (4)

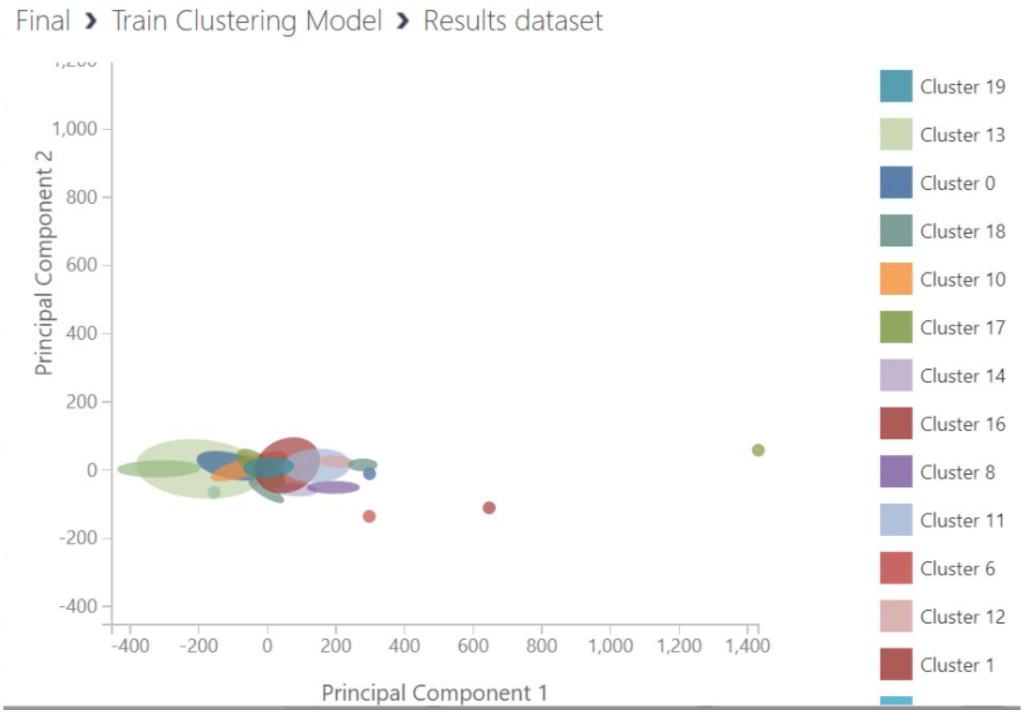
  a. Applying 4 number of centroids gave us output with highly overlapping clusters at least according to visualization but as prof taught visualizations do not give enough idea regarding clusters of food items



(II) High number of centroids with too high iterations (20)

a. Certainly, huge iteration eats computational power but giving out accuracy same applies for a greater number of centroids but it might over fit the model important indication being very miniscule data points in each cluster

Output we got had segregated clusters but very smaller clusters.

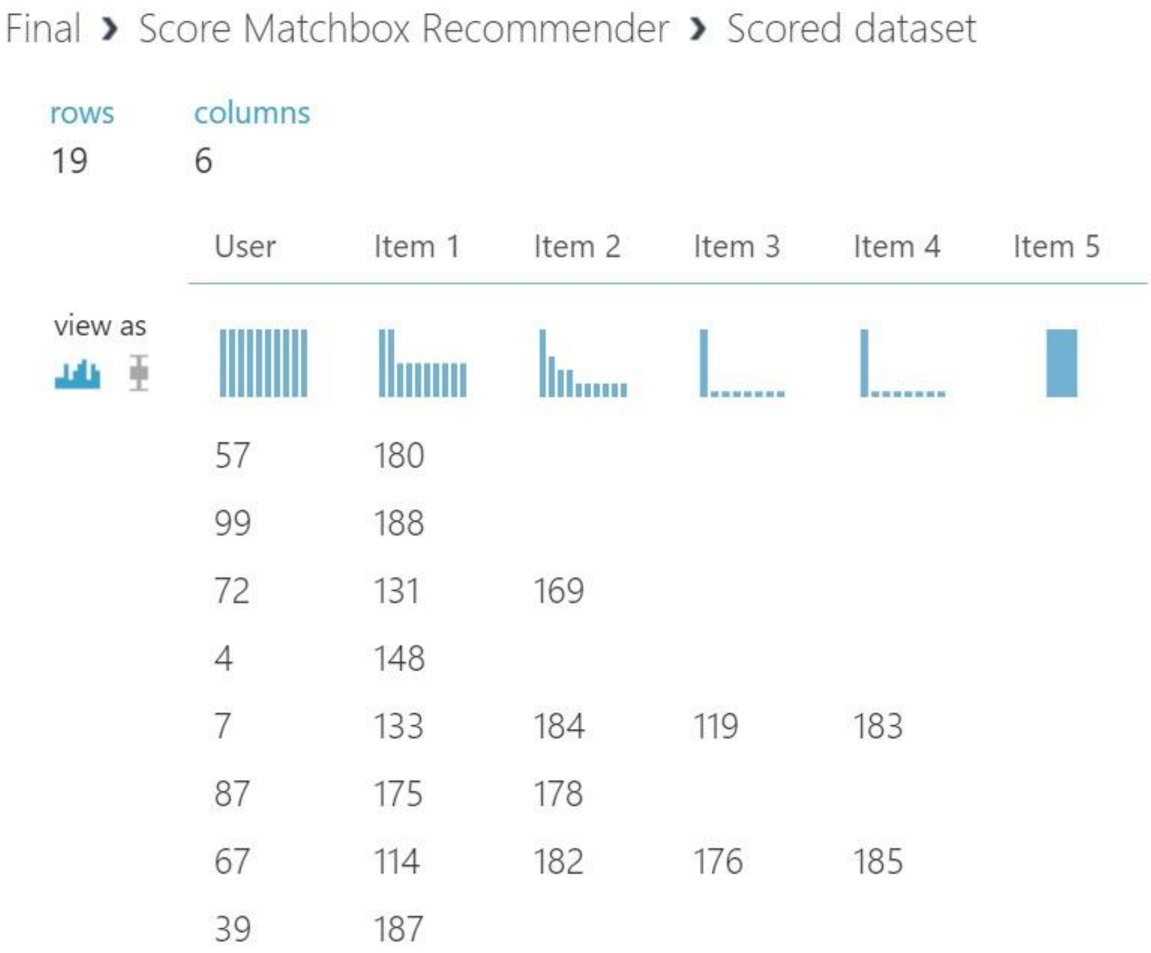


Matchbox Recommender:

When a user is new to the system, predictions are improved by making use of the feature information about the user, thus addressing the well-known "cold-start" problem. However, once you have collected enough ratings from a particular user, it is possible to make fully personalized predictions for them based on their specific ratings rather than on their features alone. Hence, there is a smooth transition from content-based recommendations to recommendations based on collaborative filtering. Even if user or item features are not available, Matchbox will still work in its collaborative filtering mode.

One of important parameter is NUMBER OF TRAITS (more latent traits or type about user) in the algorithm increment of which provides better precision as it stems more nuances in identifying user and usually alters ranking of recommended item with precision

(I) Match box recommender with number of traits: 20



Final › Score Matchbox Recommender › Scored dataset

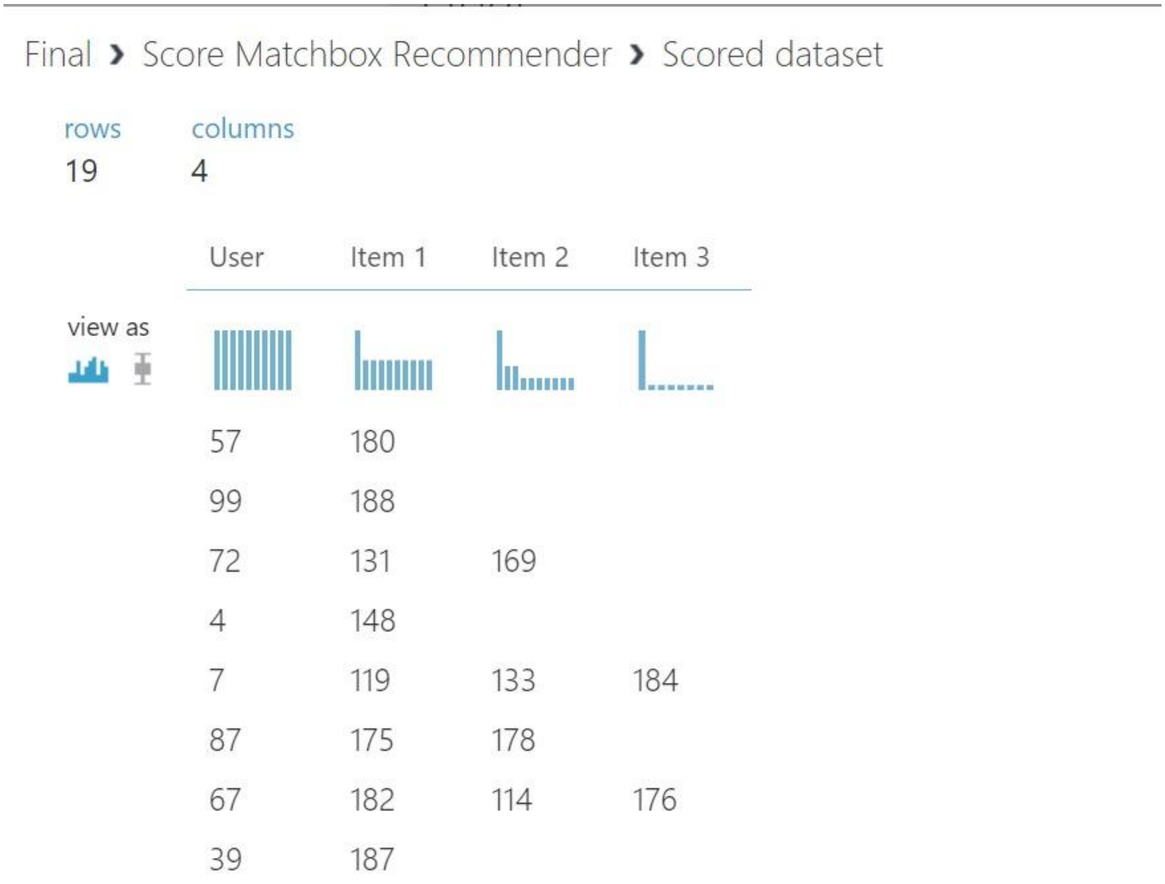
rows 19

columns 6

view as

User	Item 1	Item 2	Item 3	Item 4	Item 5
57	180				
99	188				
72	131	169			
4	148				
7	133	184	119	183	
87	175	178			
67	114	182	176	185	
39	187				

II) Matchbox recommender with No. of traits: 2



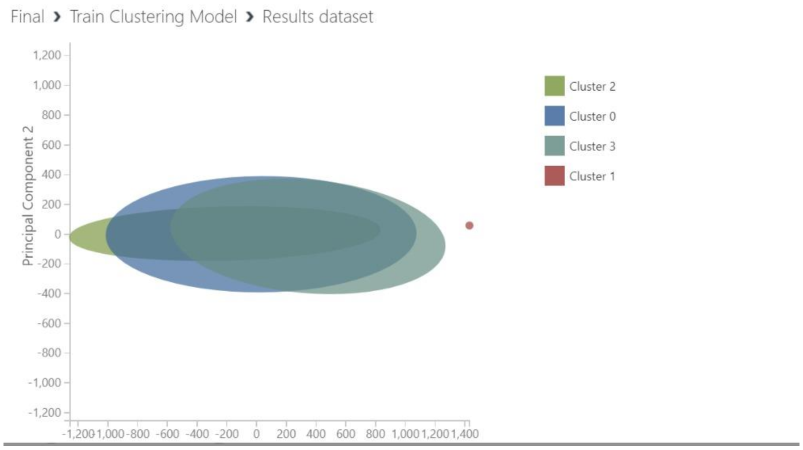
Final › Score Matchbox Recommender › Scored dataset

rows 19

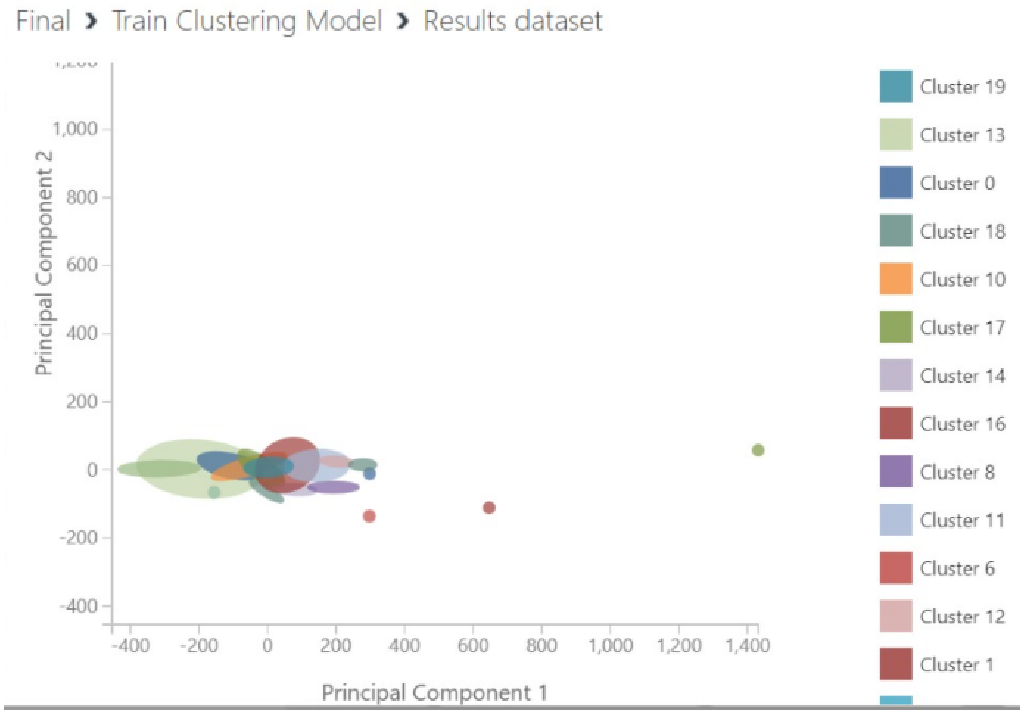
columns 4

User	Item 1	Item 2	Item 3
57	180		
99	188		
72	131	169	
4	148		
7	119	133	184
87	175	178	
67	182	114	176
39	187		

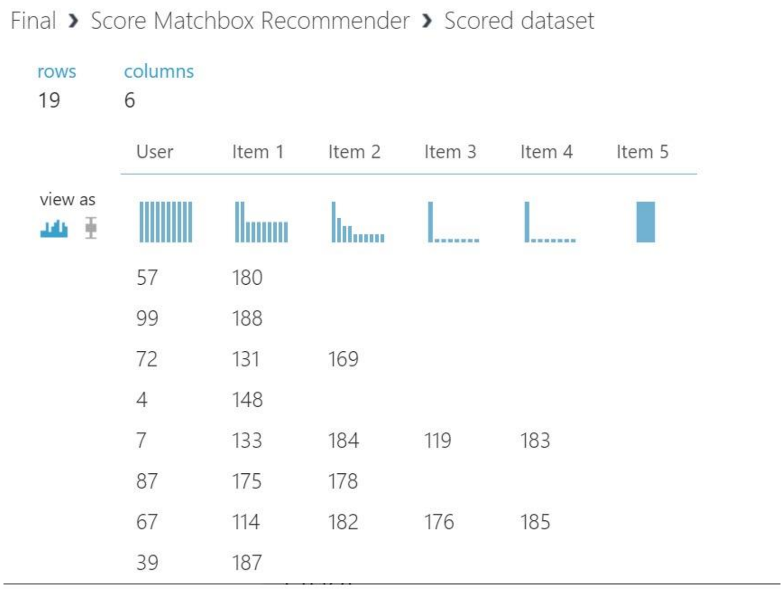
We can see that even with smaller dataset of users model altered preferences for some of the users which indicates the difference of power of prediction due to change in number of traits



Final › Train Clustering Model › Results dataset

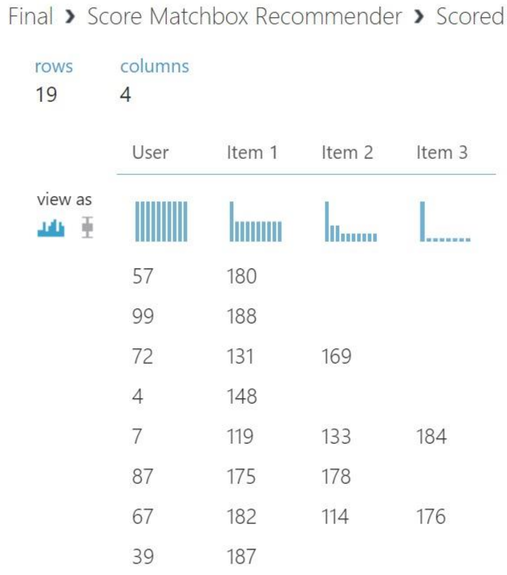


Final › Train Clustering Model › Results dataset


Final › Score Matchbox Recommender › Scored dataset

rows	columns
19	6

User	Item 1	Item 2	Item 3	Item 4	Item 5
57	180				
99	188				
72	131	169			
4	148				
7	133	184	119	183	
87	175	178			
67	114	182	176	185	
39	187				


Final › Score Matchbox Recommender › Scored dataset

rows	columns
19	4

User	Item 1	Item 2	Item 3
57	180		
99	188		
72	131	169	
4	148		
7	119	133	184
87	175	178	
67	182	114	176
39	187		

CONCLUSION:

Overall, we have observed that Matchbox recommender performed better than the K Means Clustering algorithm on the Azure Machine Learning Studio platform. Considering the severity of the scenario we are dealing with we have chosen Accuracy, Misclassification rate and the implementation complexity as the metrics for choosing the best algorithm.

This algorithm we found is highly sophisticated and with more advanced precision as it smartly combines two powerful algorithm and treats each limitation with right strategy plus this unsupervised algorithm also considers dishes features too unlike former which provides edge in recommending precisely.

What you need to do as a Decision maker:

- Gathering enough data to perform our analysis.
- Run the model which will recommend food items based on customer preferences.
- Among those recommendations, Customer will choose most profitable dish.
- Business can offer complementary food as BOGO / discounted while try it out as campaign.
    - Factors to Consider while giving discount:
        - Cost you bear on that recommended food product
        - margin associated with it
        - profit generated already from that customer (loyalty and trust)
    - If you are "not taking risk at all" type you can run marketing campaign 'TRY IT OUT' and just freely recommend customers without any rewards (you will bear minimal cost but some proportion of customers may not ignore recommendations so it is all about risk appetite)

    - Incentivize consumer to have variety of food which will drive revenue for sure
    - Monitor dashboard which shows your revenue per customer gets boosted or not due to these recommendations.