Zappos Challenge

Darshil Gohel

Masters Student
University of Texas at Dallas

PART I - Brainstorm Factors

First time visitor to the site:

For first time user, the main source of information(data) will be click stream data.

- **Items visited by user:** We can get the pages visited the user by clickstream data and hence we can get the data for the items which will be more interested to user. We can recommend this items to users.
- **Time spent on each item :** if user spends more time on one item, he is more interested in that item.
- **Number of clicks:** if the user will be more interested in one item, he will try to extract more information by clicking various details section. So this parameter is directly connected to likelihood of item by user.

- **IP address [customer information]** - We can give preference to products that are bestsellers in the customer's geographic region.

- **Calendar date / time of year** - We should show season-appropriate products (snow boots in winter, sandals in summer).

Returning customer:

- **Gender of customer**: System have the data about the gender of the customer from the first purchase and hence we should recommend the items according to the gender of the customer. Like Men footwear for men and women footwear for the women.
- **Age of customer**: different age group have different affinity towards items like young boys like more sports and casual shoes cloths rather then professional shoes and cloths. Old ladies will like differnet type of cloths then yourg girls.
- **Physical characteristics(Size) of customer:** what is size of customer?? Did he buy medium size shirt or XL type of shirt? If he has bought the small size first he must be short man and will need 39 size shoes in future.
- **Categories of products:** How much time customer have spent on each category and from which category he has bought something. We should go for that catagories as recommendation.
- **Clearance//New arrivals tendencies:** Is customer is only interested in new arrivals or Is customer is only interested in clearance goods. What is his tendency??
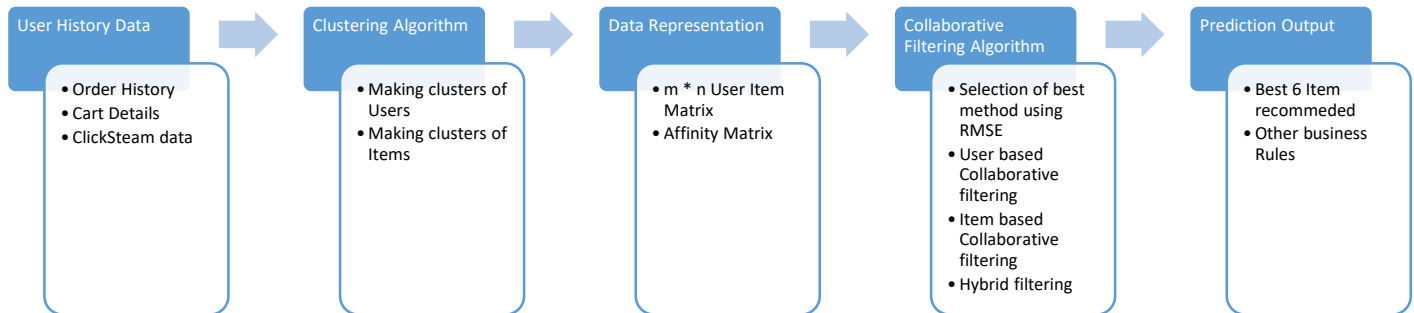
Loyal customer:

- **Buying/ Expenditure Trend:** Customer have bought more and more and spending more and more items over the time or he had initially bought more and he is spending less and

less nowadays. What is customer expenditure trend over the time?? If he is spending more , then recommend him costly items.

- **Discount/buying trend:** From the history of data, Has customer affected by the discounts given in different categories? Then we should recommend more items which discount as large as possible on particular category.
- **Hobby:** some customer would like to exercise more hence they will buy more sport cloths and shoes, some would like to swim and hence they will buy swimsuits. So, we need to find out the hobby/personality of that customer and recommend the system.
- **Personality:** Some customer will buy items like cheap cloths and cheap shoes very frequently or some customer will buy very expensive cloths and shoes two times in year.so this kind of personality finding are useful in recommendation system as we should recommendate cheap shoes and cloths to that frequent User.
- **Age/Necessity variance**: Some of the customer grown in age over time, suppose 18 year teenager will become 25 year young man in 7 years and his preference may be different over the time.25 year young man will become father of baby in coming two or three years, and hence pregnancy cloths for him wife or baby cloths for his baby should be recommend to him after 2-3 years.
- **Stuck to one category/ different category**: Has our customer bought items of only one category or he have bought other categories as well in considerable amount. Suppose, someone will come to zappos.com for only shoes and some customer will come for shoes, cloths and language.

PART II - Select Products

Recommendation System Architecture:

User History Data		Clustering Algorithm		Data Representation		Collaborative Filtering Algorithm		Prediction Output
• Order History • Cart Details • ClickSteam data	→	• Making clusters of Users • Making clusters of Items	→	• m * n User Item Matrix • Affinity Matrix	→	• Selection of best method using RMSE • User based Collaborative filtering • Item based Collaborative filtering • Hybrid filtering	→	• Best 6 Item recommeded • Other business Rules

Recommendation system in real time are far more complex then illustrated here. However, we are looking at a sample recommendation system and hence it's complexity has been reduced.

1) **User History Data**: Order History data, cart details , clickstream data of the use
2) **Clustering Algorithm**: This algorithm(e.g K means clustering algorithm) divides the user and items into certain categories. Like Men or Women, Age groups, color bias. Etc.
3) **Data Representation**: Data has to be organized in Affinity Matrix or m*n User Item Matrix. Sample Affinity Matrix:

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
U1	3	2	5	5	NA	3	4	3	0	3
U2	NA	3	3	4	1	0	4	NA	NA	2
U3	1	NA	3	0	NA	3	5	NA	1	1
U4	5	NA	5	5	2	1	NA	3	0	NA
U5	NA	3	4	NA	3	5	1	0	5	3
U6	2	0	4	0	0	4	NA	0	4	1
U7	NA	0	5	5	2	3	NA	NA	NA	NA
U8	5	0	3	3	0	3	4	5	NA	1
U9	3	2	5	1	2	0	5	2	2	4
U10	NA	1	2	0	2	4	NA	3	3	2

Where U1,U2→ Users

I1,I2→Items

And 3,2,4→ rating given by Users to some items.

4) **Collaborative Filtering Algorithm**: following algorithms has been selected as filtering algorithms.
   a. User Based Collaborative Filitering Algorithm
   b. Item Based Collaborative Filtering Algorithm
   c. Hybrid Filtering Algorithm
5) **Prediction Output:** Using among best algorithm chosen from above three and according to other business rules , best 6 items which is predicted for particular user.

Filtering algorithm selection approach For Three types of Users:

Assumptions:

- For comparison of methods, Affinity Matrix have been created according to different case of new, returning and loyal customers.
- Rating has been given on scale of 0 to 5 (6 distinct values)
- Here, we calculate the RMSE (Root Mean Square error) in runtime for all the methods and select the best method according to lowest RMSE.
- Datasets (affinity matrices) have been divided as 90% of training data and 10% as testing data for comparison of different techniques.

→First time visitor to the site

We are making the recommendation system for the sample data and affinity matrix.

Assumptions:

1) Number of Users: It is assumed here that the user is completely new and system don't know about that except the clickstream data. Hence , clustering system which the prior step of the recommendation system will failed to categorized it and hence New User has to be compared with large number of users. So, we are taking number of user as large as possible.

2) Number of Items: If I would be new user to zappos site then I will navigate the zappos site with curiosity of knowing which items have been sold rather then just jumping of particular category. So , in case of new User, they will try to explore more and more items and item categories then the loyal customer. New users are confused or still in dilemma what to buy and what to not. Hence keeping more and more item in our considering and hence our itemset for new user will be as large as possible.

3) Rating of Items: from clickstream data, for new users, if user spend more time on particular item, it is more likely that he likes that product and will try to buy that product in the future.

   Rating can be found out in following way(For clickstream data) (From 0 to 5)

Time spent on item(Seconds)	Rating

<10	0
10 to 20	1
20 to 30	2
30 to 40	3
40 to 50	4
>50	5

For existing and returning users, rating is actually rating which customer have given to particular item. We are actually comparing all the existing and new users in our cases based upon this criteria.

Creating affinity matrix procedure:

User Item based Affinity matrix has been contracted using below code :

```
m1<-                    matrix(sample(c(NA,0:5),1000000,                    replace=TRUE,
prob=c(0.50,0.30,0.15,0.02,0.01,0.01,0.01)),
     nrow=1000, ncol=1000, dimnames = list(
      user=paste('U', 1:1000, sep=''),
      item=paste('I', 1:1000, sep='')
     ))
```

Points to be look upon in this function and its value for new Users.

1) Number of Users: As large as possible, hence nrow=1000
2) Number of Items: As large as possible, hence ncol=1000
3) Probabilities:

Rating	Probabilities	Reason
NA	0.5	it is assumed that we don't have any data for user item rating for around 50% of values
0	0.3	It is assumed that New user is just exploring site and his tendency would be exploring most of items in shot time
1	0.15	It is assumed that New user is just exploring site and his tendency would be exploring most of items in shot time
2	0.02	it is very less likely that user will stuck to particular item and spend time
3	0.01	It is almost impossible to like any particular item on first visit on site
4	0.01	It is almost impossible to like any particular item on first visit on site

5	0.01	It is almost impossible to like any particular item on first visit on site

Generated m1:

m1_1.csv

Following methods have been compared:

- Item Based Collaborative Filtering
- User Based Collaborative Filtering
- Hybrid (50% Item Based Collaborative Filtering + 50% User Based Collaborative Filtering)

Following R code have been used for it .

Recommender_Syste
ms_1.r

Results:

Error calculations for the three methods:

	RMSE	MSE	MAE
UBCF	1.062457	1.128815	0.7472459
IBCF	1.102898	1.216383	0.7666330
HYBRID	1.070036	1.144977	0.7503248

Using above matrix, it is apparent that UBCF method is best as it has least Root Mean Square Error (RMSE) comparing to other two records.

Making prediction for the best six items:

Following R code has been used for the best six items recommended for the items for user 5:

Recommender_Syste
ms_1.r

Best 6 Predicted Items for the User 5 according to our algorithm.

No	Item Name
1	I736

2	I256
3	I507
4	I936
5	I486
6	I275

Twist
: Zappos' primary selling base is shoes, which accounts for about 80% of its business. Even though, Zappos expanded their inventory in 2007 to include clothing, handbags, eyewear, watches, and kids' merchandise and still in 2017 those items currently account for 20% of annual revenues. As conclusion, it is likely that customer is looking zappos as footwear site rather than other items.

Hence, it is most likely that the predicted 6 items will be footwear only(which is current case in Zappos Site !!!!). We should also include items (mostly one or two out of six) like cloths, handbags, watches which is other then footwear. Some kind of algorithm is needed at this stage to override other prediction and add items which are other then footwear.

Assuming this algorithm is created and now final list of predicted items:

No	Item Name
1	I736
2	I256
3	I507
4	I936
5	I788
6	I999

Where, item 1 to 4 are the same as predicted by our recommendation algorithm and item 5 an item 6 are items which are added by our newly created algorithm and which is other then footwear.

→**Returning customer:**

We are making the recommendation system for the sample data and affinity matrix.

Assumptions:

1) Number of Users: This user has already purchased one or two items from the site and system have some data about it like age group, gender, affinity of some categories of items. System also have some clickstream data. Hence,our clustering system will able to categorize the user and our returning user will fall in particular category and hence we will only compare returning user with the user which are present in that category. So, number of users will be somewhat less in this case.

2) Number of Items: Our returning user have bought some items but he is still unware of other items which are being sold on Zappos. So , in case of returning User, they will try to explore more and more items and item categories then the loyal customer.We need to recommend him more and

more items. Hence keeping more and more item in our considering and hence our itemset for new user will be as large as possible.

3) Rating of Items: Returning user have already rated some of our items
   Moreover, from clickstream data, for returning users, if user spend more time on particular item, it is more likely that he likes that product and will try to buy that product in the future.

   Rating can be found out in following way(For clickstream data) (From 0 to 5)

Time spent on item(Seconds)	Rating
<10	0
10 to 20	1
20 to 30	2
30 to 40	3
40 to 50	4
>50	5

For existing and returning users, rating is actually rating which customer have given to particular item. We are actually comparing all the existing and new users in our cases based upon this criteria.

Evaluating and comparing the different methods:

Creating affinity matrix procedure:

User Item based Affinity matrix has been contracted using below code :

m1<- matrix(sample(c(NA,0:5),500000, replace=TRUE, prob=c(0.40,0.20,0.10,0.10,0.10,0.05,0.03)),

nrow=500, ncol=1000, dimnames = list(
user=paste('U', 1:500, sep=''),
item=paste('I', 1:1000, sep='')
))

Points to be look upon in this function and its value for new Users.

1) Number of Users: nrow=500
2) Number of Items: As large as possible, hence ncol=1000
3) Probabilities:

Rating	Probabilities	Reason

NA	0.40	it is assumed that we don't have any data for user item rating for around 40% of values
0	0.20	Returning User know some of the item sold on zappos and he don't like them and giving them 0 rating. It is assumed that New user is just exploring site and his tendency would be exploring most of items in shot time
1	0.10	It is assumed that returning user have given 1 rating to 10% which he has visited and data collected from clickstream data
2	0.10	It is assumed that returning user have given 1 rating to 10% which he has visited and data collected from clickstream data
3	0.10	It is assumed that returning user have given 2 rating to 10% which he has visited and data collected from clickstream data
4	0.05	Still our returning user is comparatively new to our site so he will give 4 rating to very few items
5	0.03	Still our returning user is comparatively new to our site so he will give 5 rating to very few items

Generated m1:

m1_2.csv

Following methods have been compared:

- Item Based Collaborative Filtering
- User Based Collaborative Filtering
- Hybrid (50% Item Based Collaborative Filtering + 50% User Based Collaborative Filtering)

Following R code have been used for it .

Recommender_Syste
ms_2.r

Results:

Error calculations for the three methods:

	RMSE	MSE	MAE
UBCF	1.571777	2.470484	1.357055
IBCF	1.603681	2.571792	1.371418
Hybrid	1.570091	2.465185	1.356057

Using above matrix, it is apparent that hybrid method is best as it has least Root Mean Square Error (RMSE) comparing to other two records.

Making prediction for the best six items:

Following R code has been used for the best six items recommended for the items for user 5:

Recommender_Syste
ms_2.r

Best 6 Predicted Items for the User 5 according to our algorithm.

No	Item Name
1	I481
2	I609
3	I831
4	I265
5	I214
6	I433

Twist: Zappos' primary selling base is shoes, which accounts for about 80% of its business. Even though, Zappos expanded their inventory in 2007 to include clothing, handbags, eyewear, watches, and kids' merchandise and still in 2017 those items currently account for 20% of annual revenues. As conclusion, it is likely that customer is looking zappos as footear site rather than other items.

Hence, it is most likely that the predicted 6 items will be footwear only(which is current case in Zappos Site !!!!). We should also include items (mostly one or two out of six) like cloths, handbags, watches which is other then footwear. Some kind of algorithm is needed at this stage to override other prediction and add items which are other then footwear.

Assuming this algorithm is created and now final list of predicted items:

No	Item Name
1	I481
2	I609

3	I831
4	I265
5	I788
6	I999

Where, item 1 to 4 are the same as predicted by our recommendation algorithm and item 5 an item 6 are items which are added by our newly created algorithm and which is other then footwear.

→Loyal customer:

We are making the recommendation system for the sample data and affinity matrix.

Assumptions:

1) Number of Users: This user is loyal user. System have all the information about this user and it is most likely that our system will easily categorize this user and this user will be compared with the same category only and so taking fewer user number into account
2) Number of Items: This customer is loyal and it is assumed that loyal customer tend to buy the item from some few number of categories only, not all categories!!Hence,it is better to recommender him the item from those categories only in which he is most loyal to sustain its loyalty rather then just throwing some random items to him.So, taking into consideration only items of most liked categories→ Fewer no of items.
3) Rating of Items: Returning user have already rated some of our items
   Here, we are no more predict on clickstream data. This is prediction is sonely based the data system knows about the Loyal customer.

Evaluating and comparing the different methods:

Creating affinity matrix procedure:

User Item based Affinity matrix has been contracted using below code :

m1<- matrix(sample(c(NA,0:5),250000, replace=TRUE, prob=c(0.25,0.10,0.10,0.15,0.20,0.10,0.10)),

　　　nrow=500, ncol=500, dimnames = list(
　　　　user=paste('U', 1:500, sep=''),
　　　　item=paste('I', 1:500, sep='')
　　　))
Points to be look upon in this function and its value for new Users.
1) Number of Users: nrow=500
2) Number of Items: ncol=500
3) Probabilities:

Rating	Probabilities	Reason

NA	0.25	it is assumed that our loyal customer knows around 75% items and he have reated them but we don't have any data for user item rating for around 25% of values
0	0.10	Our customer is loyal and it is most likely that there will be less products which he didn't like from his favorite category.
1	0.10	Our customer is loyal and it is most likely that there will be less products which he didn't like from his favorite category.
2	0.15	There will be some items to which customer will give 2 rating
3	0.20	Most of the items our customer will like will likely have 3 rating
4	0.10	There will be some items which is loved by customer and hence customer will give them 4 rating
5	0.10	There will be some items which is loved by customer and hence customer will give them 5 rating

Generated m1:

X a,

m1_3.csv

Following methods have been compared:

- Item Based Collaborative Filtering
- User Based Collaborative Filtering
- Hybrid (50% Item Based Collaborative Filtering + 50% User Based Collaborative Filtering)

Following R code have been used for it .

R

Recommender_Syste
ms_3.R

Results:

Error calculations for the three methods:

	RMSE	MSE	MAE
UBCF	1.547359	2.394320	1.281912
IBCF	1.570394	2.466138	1.294071
HYBRID	1.543242	2.381595	1.281164

Using above matrix, it is apparent that Hybrid method is best as it has least Root Mean Square Error (RMSE) comparing to other two records.

Making prediction for the best six items:

Following R code has been used for the best six items recommended for the items for user 5:

Recommender_Syste
ms_3.R

Best 6 Predicted Items for the User 5 according to our algorithm.

No	Item Name
1	I481
2	I609
3	I831
4	I265
5	I214
6	I433

Conclusion:

For three type of customers(New,Existing or loyal), three collaborative algorithm have been compared. Following is the final algorithm applied to each customer.

User	Collaborative Filtering Algorithm
New	User based
Returing	Hybrid
Loyal	Hybrid

PART III - Improve Recommendations:

Question 1

How will you measure the effectiveness of these recommendations? As a reminder, the objective is to maximize the profitability of these recommendations. Define at least one metric that would evaluate the recommendations' success.
Ans:

For effectiveness of theses recommendation system, we can use RMSE (Root Mean square Error) method for each and every method.

Following is the equation for the RMSE.

The **root-mean-square deviation** (RMSD) or **root-mean-square error** (RMSE) is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed.

$$\text{RMSD} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}}.$$

For best recommendation system method, RMSE should be as low as possible.

For that total data needs to be divided into training and test data. From training data, the system needs to be trained and model is created. After that we need to apply that model on testing data and calculate the RMSE Values.

From the recommendation system described in section 2, the RMSE values are created from the every type of user and hence according to lowest RMSE values, best method had been chosen.

Following tables shows RMSE(Lowest) and best recommendation model everluated from every user.

User	Collaborative Filtering Algorithm	RMSE(Lowest)
New	User based	1.062457
Returing	Hybrid	1.570091
Loyal	Hybrid	1.543242

By this way, we can select the best recommendation system and increase the profitability.

Question 2 (optional)

How will your product selection process from Part II incorporate your answer to Question 1 to produce stronger recommendations over time? In essence, how will your process use these metrics in feedback loops that will affect the next iteration of recommendations? Also, take into

account that you will observe customer behaviors that may indicate they are transitioning from one customer stage to another. This can be seen, for example, in Tony's behavior below:

Ans: According to our sample recommendation system, this task will be done by our clustering algorithm.

As shown in section 2 , if new user will enter into system, our clustering algorithm fails to put it in particular category because of lack of data and hence it has to be compared to large number of users and items. We have predicted that for this type of scenario user based collaborative filtering is the best and it will predict best 6 items.

Now, we have lot more data about new users. When, new user will retune next time, he will become Returning user. This time, as we predicted in section 2 , hybrid collaborative algorithm will run and best 6 item will be selected.

After many visit of zappos site, our returning customer will become loyal customer and at that time again hybrid collaborative algorithm will run and best 6 items will be selected and recommended.

We are collected the data of users at every visit to site and hence it is more likely that user will move from one categories to another and still our recommendation system will produce the best results over the time.