Proposal

Chenru Liu 1501213463

1. the goal and meaning of this data analysis
   we have some data for the product information and user information, and the user's action information with respect to the product during 2016-02-01 to 2016-04-15. My final goal was to use all this information to recommend the products to customers that they are more likely to be interested in.

2. the process of the whole project
   1> data analysis
      we need to do some statistical analysis on the user data, product data, and action data
   2> process for the NA and other things
   3> split the data into in-sample data and out-of-sample data
   4> use the in-sample data to train model and test it on the out-of-sample data

3. Data description

Data source: data from an online shopping website

Data kinds:

User data

user_id	id for user
age	-1 to denote no info
sex	0:male，1:female，2:no info
user_lv_cd	User level . Higher number, higher level
user_reg_dt	User registration day

Product information

sku_id	Id for product
attr1	Enum type. -1 to denote no info
attr2	Enum type. -1 to denote no info
attr3	Enum type. -1 to denote no info
cate	Class ID
brand	Brand ID

Comment information

dt	Comment deadline day
sku_id	Id of product
comment_num	The number of comments for that product 0: no comment

	1: 1 comment 2: 2-10 comments 3: 11-50 comments 4: more than 50 comments
has_bad_comment	0: have no bad comment, 1:have bad comment
bad_comment_rate	The ratio of bad comments to totoal comments

Action information

user_id	The id of product's user
sku_id	The id of product
time	The time of action
model_id	The module that user click if click on the product
type	The type of the action: 1. Browse 2. Add to the shopping list 3. Delete from the shopping list 4. Purchase 5. Add the product to the interest list 6. click
cate	Class ID
brand	Brand ID

4. Some idea about the projects:

a.The process needs to be done:

1> I need to specify the people who are more likely to purchase the product. So later we will give recommendations to these people.

2>With respect to the users, we need to find products' correlation, which product will be more likely to be purchased if a user has buy A product.(for this part we may be focused on the product category instead of the specific data). For this part, we may mainly use the action information for product correlation and product information for product classification.

3> From comment information, we may find the best product for each category. So we can recommend better goods for customers. (in reality we make recommendation for specific user according to her buy habits: the price elasticity, the product bought before)

4. From action information, we can correlate the information of action with the product.

b. Final result:

1>For each user, we want to forecast whether they will buy products during 2016-4-16 and 2016-4-20.
2>For the user we believe that they will buy the product, we need to specify the product ID that they will buy.

c. The measurement of algorithm

1> The accuracy rate of our forecast about whether the user will buy the product or not at specified day.
2> The accuracy rate of our forecast about which product the customer will buy.

d. The algorithm we may include in the data process:

1> K-NN algorithm: we may use this algorithm to classify the user/products into different parts.
2> Decision tree model: we may use this model to find which attributes may finally lead to purchasing action.
3> Apriori algorithm: we may use this method to find the correlation of products.