Team No - 29
Project No - 8

User Profiling Engine

Predicting potential buyers and their buying behavior on E-commerce websites

Project Members
Utkarsh Agarwal (201301184)
Viplav Sanghvi (201505573)
Lalit Kundu (201201062)

Mentor
Anurag Tyagi

1. Abstract

Number of people who uses internet and websites for various purposes is increasing at an astonishing rate. More and more people rely on online sites for purchasing rented movies, songs, apparels, books etc. The competition between numbers of sites forced the web site owners to provide personalized services to their customers. So the recommender systems came into existence. RS providers are recording users' activities, computing statistical models, and delivering recommendations on demand. Since the past 10 years, there has been an increased interest in providing recommendation of products, discounts, etc to the consumers for increasing the sales. Such information is of high value to an e-business as it can indicate not only what items to suggest to the user but also how it can encourage the user to become a buyer. In this paper we present a method which will help the e-commerce businesses in providing the user some dedicated promotions, discounts, recommendation of products, etc. Here we use the Yoochoose Dataset provided in the Recsys Challenge 2015. First we study and analyse the dataset and extract the relevant features and then associate it with the Random Forest Classifier to find out its efficiency. We have even show a comparative study of features.

2. Introduction

Many small and mid-sized e-commerce businesses use services from recommender system (RS) providers to outsource the implementation and operation of their RS. RS providers are recording users' activities, computing statistical models, and delivering recommendations on demand.

Due to the nature of the collected data, which is usually implicit, the task of the RS provider is to deliver top-N recommendations and not really to predict what will be the rating that a user will give an item. YOOCHOOSE GmbH is such a RS provider that is specialized in calculating best matching top-N recommendations on a user base for different use cases like generating cross- or up-sell, exploit long tail and last but not least to keep the user entertained.

In this challenge, YOOCHOOSE is providing a collection of sequences of click events; click sessions. For some of the sessions, there are also buying events. The goal is hence

to predict whether the user (a session) is going to buy something or not, and if he is buying, what would be the items he is going to buy. Such an information is of high value to an e-business as it can indicate not only what items to suggest to the user but also how it can encourage the user to become a buyer. For instance to provide the user some dedicated promotions, discounts etc'. The data represents six months of activities of a big e-commerce businesses in Europe selling all kinds of stuff such as garden tools, toys, clothes, electronics and much more.

3. Project Description

The goal of our project is to create a window of opportunity for the E-commerce sites to increase their sell value by understanding the customers and their needs. The more we know about the customers, the easier it is to identify opportunities to sell them new products and target them with appropriate offers.

Our system will take a sequence of click events performed by some user during a typical session in an e-commerce website as input. The aim is to predict whether the user is going to buy something or not, and if he is buying, what would be the items he is going to buy. The task could therefore be divided into two sub goals:

1. **Is the user going to buy items in this session?**
2. **If yes, what are the items that are going to be bought?**

4. Dataset

The german company Yoochoose, which provides recommendations for e-commerce platforms, news and media, sponsors the competition. They give us a dataset of

completely anonymized (unless otherwise proven) implicit feedback data coming from an e-commerce business located in Europe. The business sells all kind of stuff such as garden tools, toys, clothes, electronics and much more.

The dataset has:

33 MILLION CLICKS

1.1 MILLION BUYS

9 MILLION SESSIONS (USERS)

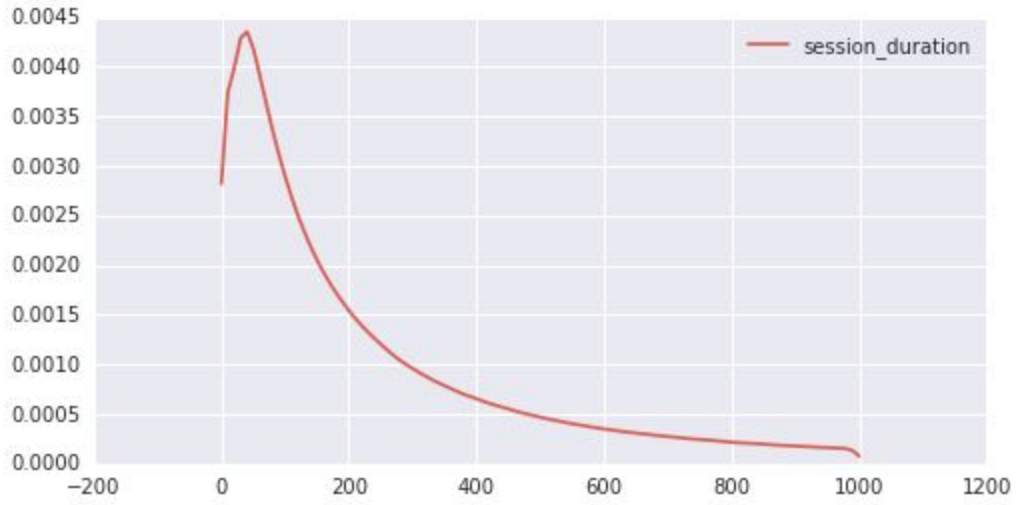
53K ITEMS

The data represents six months worth of clicks grouped in sessions, from April 1st to September 30th. For some of the sessions (around 5%), there are also buying events, which is the interesting part, as the goal of this challenge is to predict whether the user (a session) is going to buy something or not, and if she is actually buying, what are the items bought. At this point it is important to note that user and sessions are equivalent in this problem. Since the data is anonymized, the terms session and users are interchangeable.

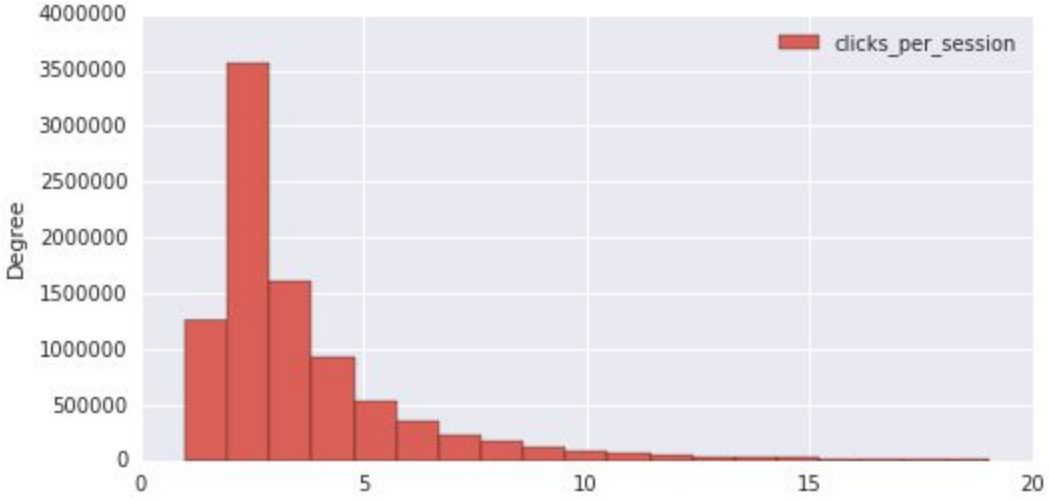
The YOOCHOOSE dataset contain a collection of sessions from a retailer, where each session is encapsulating the click events that the user performed in the session. For some of the sessions, there are also buy events; means that the session ended with the user bought something from the web shop.

Dataset Analysis:

Session Duration

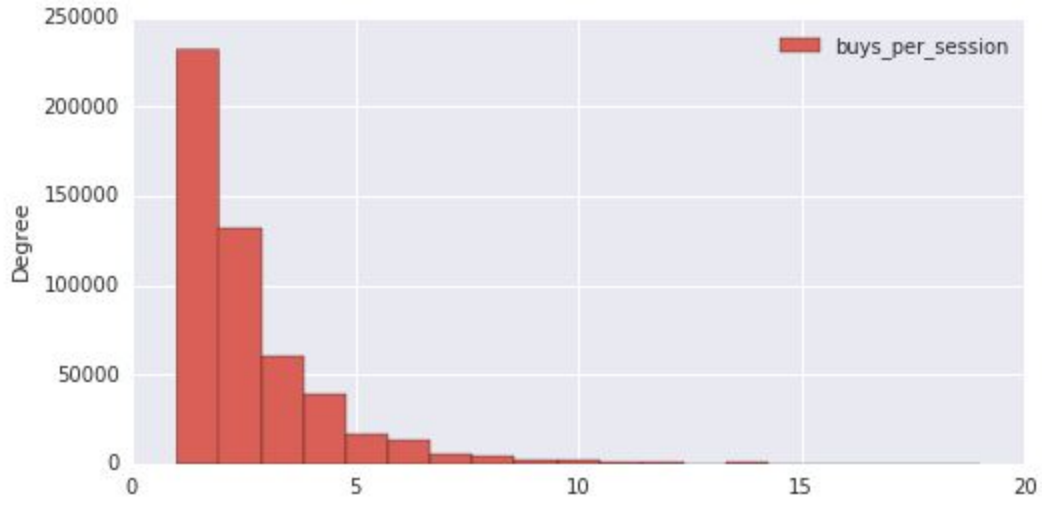


Clicks Per Session



	clicks_per_session
count	9249729.000000
mean	3.568098
std	3.787520
min	1.000000
25%	2.000000
50%	2.000000
75%	4.000000
max	200.000000

Buys Per Session



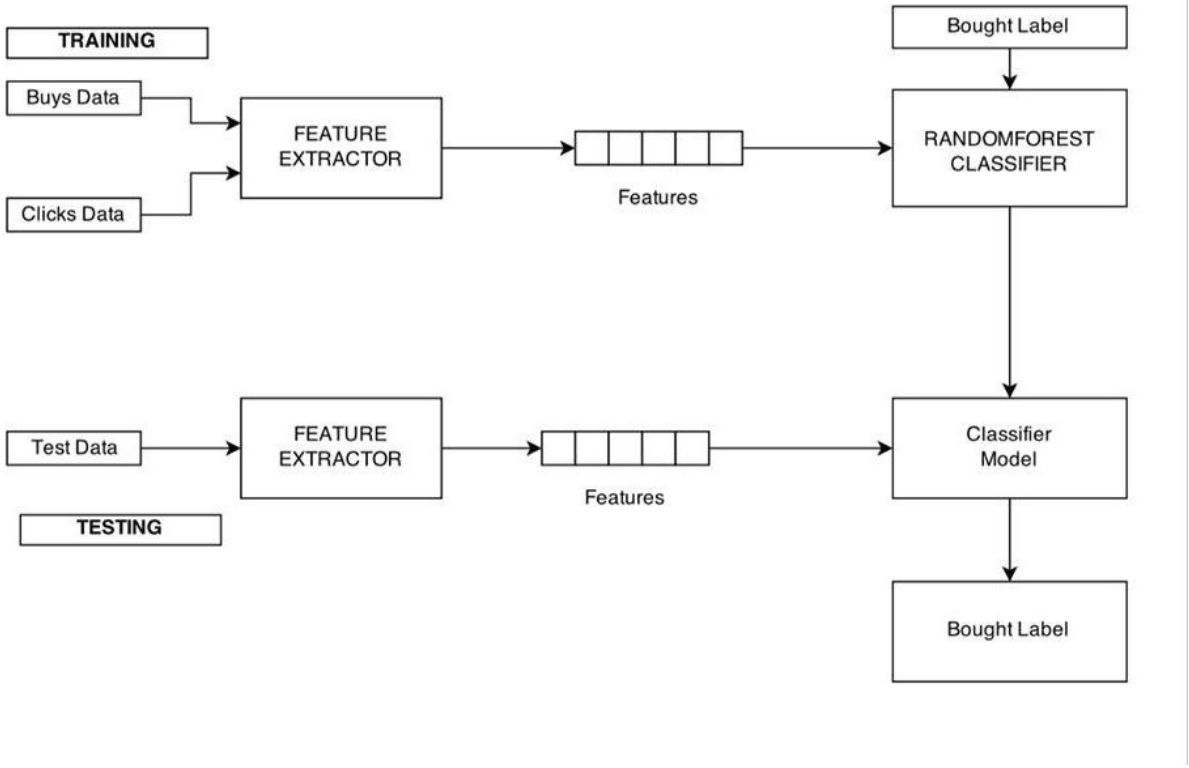
	buys_per_session
count	509696.000000
mean	2.257724
std	1.933342
min	1.000000
25%	1.000000
50%	2.000000
75%	3.000000
max	144.000000

Challenges in Dataset:

1. Yoochose-clicks.dat file is very large to handle(1.5 GB).
2. Number of items bought are very less than number of clicks.
3. Lack of common features in yoochooseclicks.dat and yoochose-buys.dat
4. The category variable in yoochooseclicks.dat is too sparse and requires to be binned.

5. Approach

The task is divided into two modules: Feature Extraction and Classification. In feature extraction we extract the relevant features like buys/clicks, popularity, hour of day, etc. from the dataset and use these features as an input to our classifier. In the classification module we train and generate a classifier model through which we classify later whether those feature vectors from test dataset will label the item as bought or not bought.



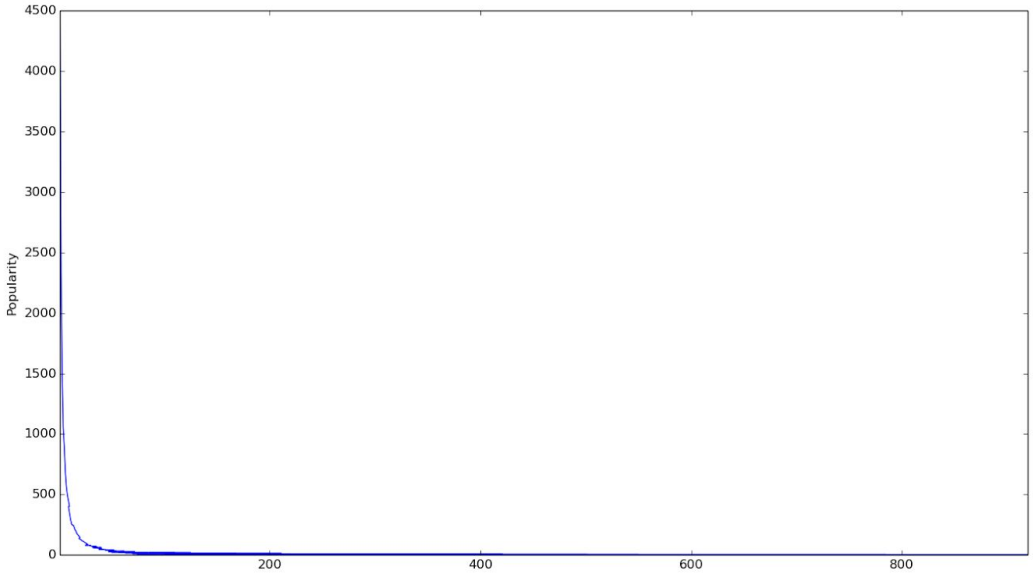
A. Feature Extraction

We have extracted item based features in every session and consequently predicted whether an item would be bought or not.

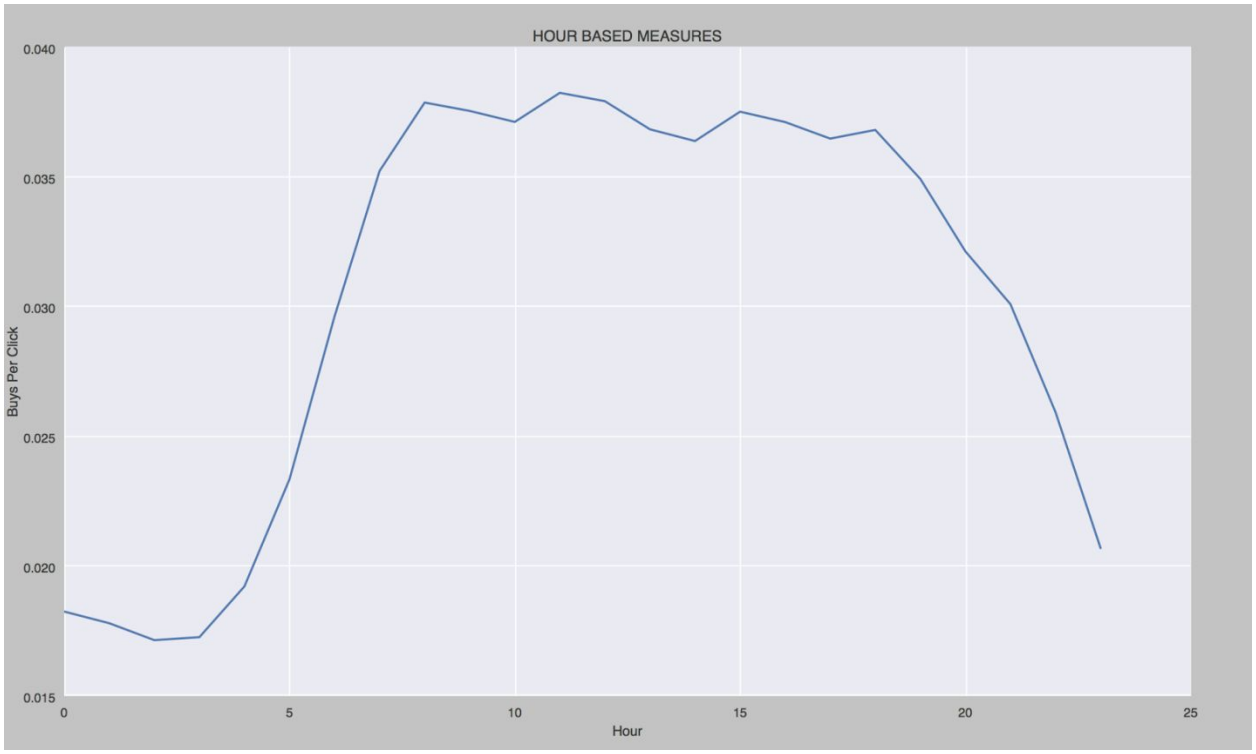
The following are the important features:

1. Buys per Click: Items which have a high proportion of buys with respect to clicks in the training set have a higher chance of being bought in the test clicks as well.

2. Item Popularity: Globally popular items in the training data have a higher chance of being bought in the test data as well. Items which are clicked once and bought once have a high value with respect to previous feature - buys per clicks.

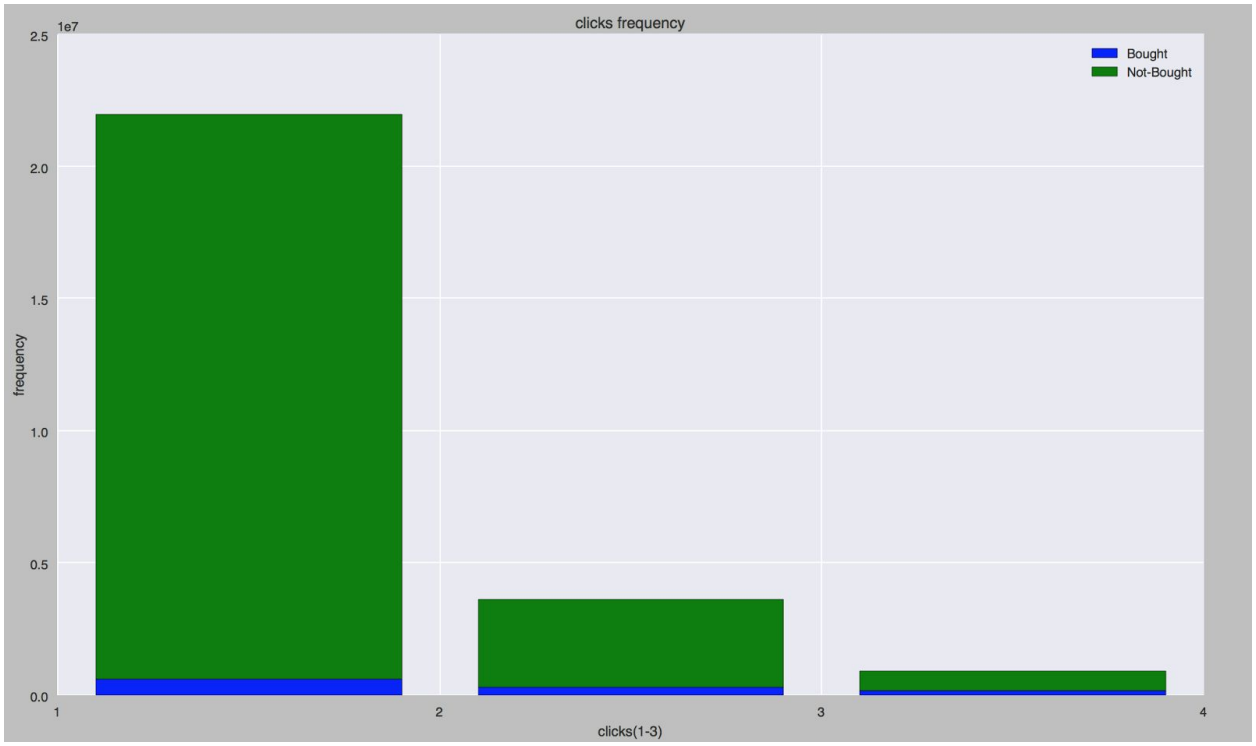


3. Hour of Item Click: In the training data, hours 8-17 of a day have been observed to have high buys/click rate compared to others. We have categorized our hour by serially numbering them from 0 to 23.

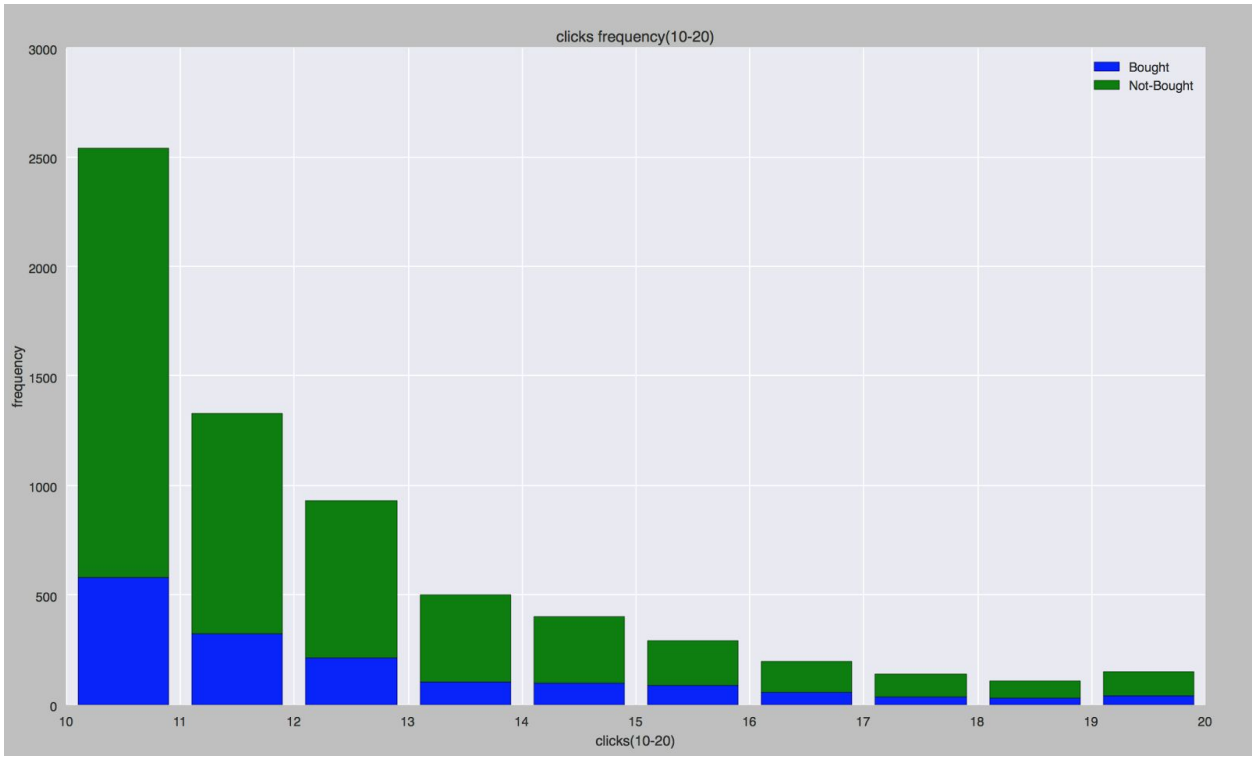


Hour based measures

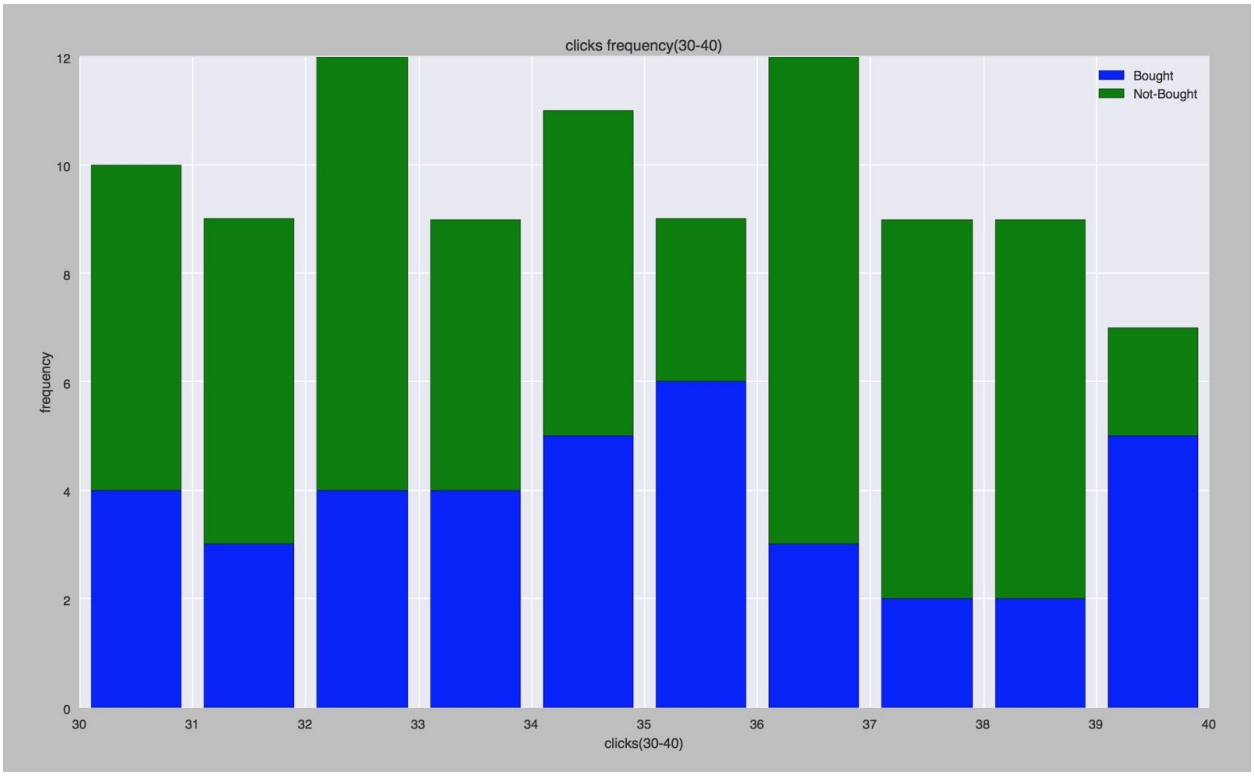
4. Item Clicks: An item which receives more number of clicks in a session is more likely to be bought by the user.



Clicks frequency



Clicks frequency (clicks ranging from 10-20)



Clicks frequency (clicks ranging from 30-40)

5. Day of the Week: In the training data, weekends have been observed to give higher number of buys as compared to other days for some items. We categorize the clicks into seven categories:

0 for session clicks on Monday
1 for session clicks on Tuesday
2 for session clicks on Wednesday
3 for session clicks on Thursday
4 for session clicks on Friday
5 for session clicks on Saturday
6 for session clicks on Sunday

6. Session Duration: The most likely reason that a user spends a lot of time on a particular item can be that he is interested in the item and is taking his time in deciding whether to buy the item or not and hence a high value of this feature increases the chances for the item in being bought.

7. Day Of Month: The information in this feature is already extracted by including features such as Day of Week and Week of Year.

8. Month Of Year: We have given a numeric value to each month of an year from April to September.

4 for April
5 for May
6 for June
7 for July
8 for August
9 for September

B. Classification

We are using Random Forest Classifier Model to classify the test set into respective classes.

Random Forest

In random forests , each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. In addition, when splitting a node during the construction of the tree, the split that is chosen is no longer the best split among all features. Instead, the split that is picked is the best split among a random subset of the features. As a result of this randomness, the bias of the forest usually slightly increases (with respect to the bias of a single non-random tree) but, due to averaging, its variance also decreases, usually more than compensating for the increase in bias, hence yielding an overall better model.

We have used the random forest Implementation in sklearn library of python - **sklearn.ensemble.RandomForestClassifier**

Reasons for using Random Forest Classifier:

- Can Build Complex decision regions.
- Combining Classifier outputs helps in preventing Over fitting.
- Capable Of Dealing with Imbalanced Datasets.

- Capable of handling millions of input samples unlike SVM which would work only if the number of input samples are in thousands.

6. Evaluation Measure

The evaluation considers taking into consideration the ability to predict both aspects – whether the sessions end with buying event, and what were the items that have been bought. Let's define the following:

- **Sl** – sessions in submitted solution file
- **S** - All sessions in the test set
- **s** – session in the test set
- **Sb** – sessions in test set which end with buy
- **As** – predicted bought items in session s
- **Bs** – actual bought items in session s

then the score of a solution will be :

$$Score(\boldsymbol{Sl}) = \sum_{\forall s \in \boldsymbol{Sl}} \begin{cases} if\ s \in \boldsymbol{Sb} \rightarrow \frac{|S_b|}{|S|} + \frac{A_s \cap B_s}{A_s \cup B_s} \\ else \rightarrow -\frac{|S_b|}{|S|} \end{cases}$$

7. Results

Following are the final scores obtained with various features .

Item clicks + Buys/click = 40163

Item clicks + Buys/click + popular items + Hour of Day = 40428

Item clicks + Buys/click + popular items + Hour of Day + Month Of Year = 45821

Best Score Obtained = **45821**

8. Conclusion

Proper Features influence the score of prediction. Not all features are useful in determining buyer behavior and some of them may even prove to be detrimental.

Features that take too many values may actually overfit, it is better to bin the values into fewer value sets.