Mid-Report(1)

Session-based Purchase Prediction and Explanation

Team GAZUA

2011147064 JuHyuk Lee

2012147562 Inho Choi

2014154009 YooJin Kim

**Contents**

1. Subject
   1. Session-based Purchase Prediction and Explanation
   2. Necessity
2. Previous Research and limitation
   1. Previous Research
   2. Limitation & Our development
3. Current Progress
   1. Data Analysis
4. Future Plan
5. Reference
6. **Subject**
   1. **Session-based Purchase Prediction and Explanation**

Purchase prediction is about predicting user’s product purchase probability, to know whether a user will purchase the product or not. In this project, we’ll use session-based user data to predict the user’s product purchase by developing a neural model.

Also, using neural network model could be hard to understand since it is a black-box model. People wouldn’t trust the output from simple black-box model without reasonable explanation about why the model is reliable. According to this problem, we’ll use an explanation model LIME(Local Interpretable Model-agonistic Explanation) to explain the predictions of our classifier.

* 1. **Necessity**

Recommending a worthy item to a user is important for sellers to gain high profit. Since online market becomes popular to users in these days, lots of sellers use online market for their marketing. Not for only off-line marketing, now also for on-line marketing product recommendation is necessary. To recommend a worthy item for a user in eCommerce market, seller should know which product user is interested in, and wants to purchase finally. For this reason, purchase prediction now becomes important in eCommerce market.

Since session data is a reasonable data to get in online, session-based purchase prediction will be a good solution. And with some reasonable explanation, it could be a reliable solution for eCommerce marketers to manage customers.

1. **Previous Research and limitation**
   1. **Previous Research**
      1. Purchase prediction with user data

In paper about predicting online purchase conversion for retargeting, researchers used logined user data to predict the user purchase model. Given an e-commerce site, each customer has a unique cookie ID. Researchers use this unique cookie ID and its page visit (click or hit) and product purchase.

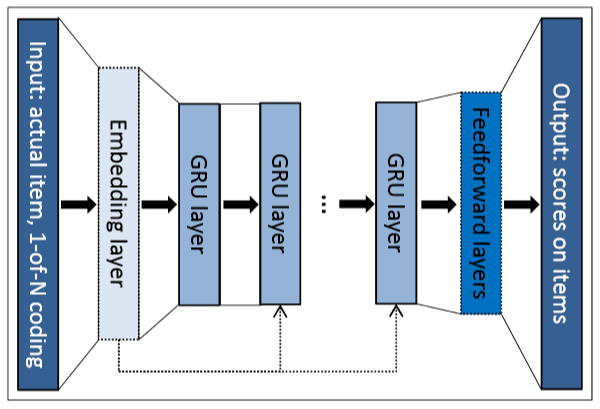
Using this data, researchers study about the conversion rate, which means the rate of conversion from window-shoppers to a real buyer. For the result of their experiments, researchers proposed a joint modeling of customer-level and product-level conversion patterns based on the buying decision process, to get the reliable conversion rate.

* + 1. Session-based Recommendation with RNN

In paper about session-based recommendation with RNN, researchers used short session-based data to build a recommender system with RNN. Since RNN is used for sequential data modeling, their basic idea was trying to modeling the whole session as it is a sequential data.

Researchers used GRU-based RNN, which means elaborating model of an RNN unit dealing with vanishing gradient problem which is showed in Figure 1. Also they used ranking loss function to modify classic RNN, and use ranking method to get top items that the user might be interested in.

For the result of these experiments they got a recommendation system for eCommerce marketers to predict user’s purchase and recommend some reasonable product for the user.



<Figure 1>

* 1. **Limitation & Our development**
     1. Limitation

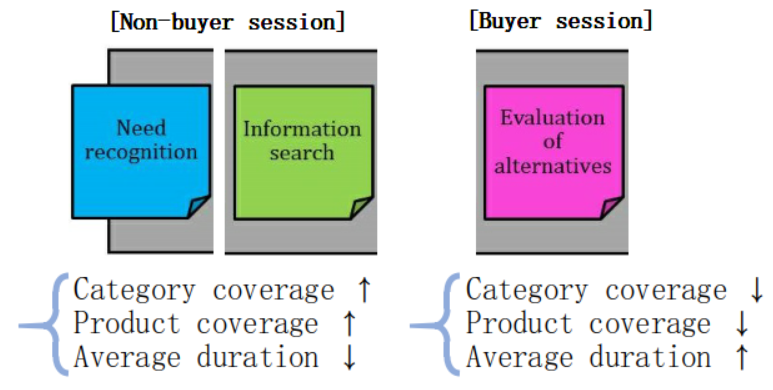
The previous studies have two major big limitations. First, the study about purchase prediction with user data used only logined data of users to purchase their prediction. In their research, used data was with user cookie ID data. Since people usually visit online shop without login, we thought that there must be more reasonable data that related with all users who does not logined.

The study about session-based recommendation with RNN also has a serious problem. In this research, researchers used RNN model to make prediction. However, since RNN is a black-box model, it is hard to understand or trust the output from the model. It needs more reliable explanation model for the output. So we try to find out some more explainable model by refer some explanation method, LIME (Local Interpretable Model-agnostic Explanations).

* + 1. Our development
       1. Our research will find a explainable model for predicting user’s purchase.
       2. To develop a prediction model, we’ll use session-based user data.
       3. Since the dataset is not sufficient, we’ll use ACGAN to oversample the imperfect dataset to use.
       4. We expect that our new prediction model could be applied to eCommerce customer management.

1. **Current Progress**
   1. **Data Analysis**

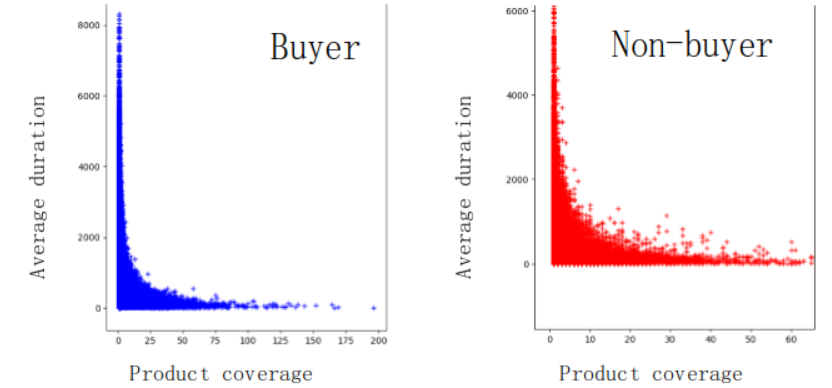
We had some assumptions about category coverage, product coverage, average duration. First, Category coverage of Non-buyer session would be bigger than Buyer session. Second, Product coverage of Non-buyer session would be bigger than Buyer session. Third, Average duration of Non-buyer session would be shorter than Buyer session.



In order to check if our assumptions are correct or not, we did data analysis by next procedure.

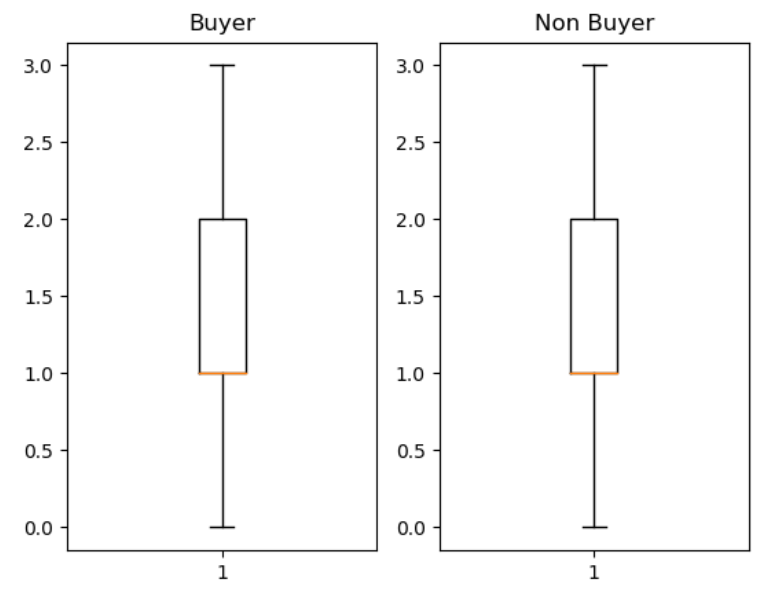
Our data is composed of buyer session(509,696) and non-buyer session(8,740,033). We sampled same number(509,696) of non-buyer session with buyer session. After that, we visualized following four graphs and analyzed them.

**Product coverage - Average duration**



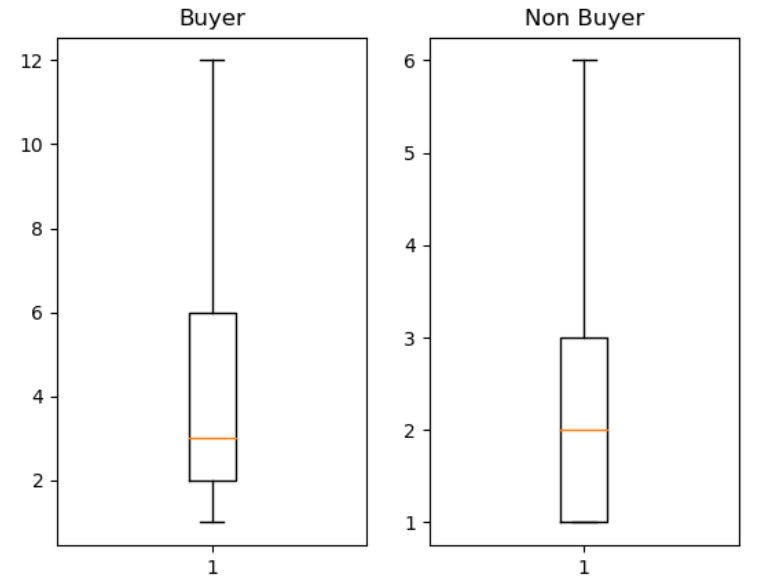
⇒ Product coverage and Average duration are inversely proportional.

**Category coverage**



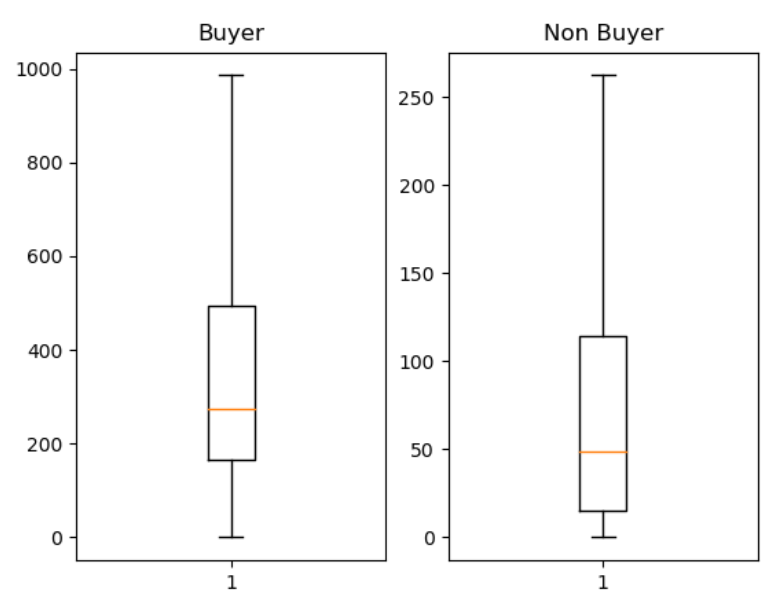
⇒ Category coverage is not an important factor because each graph shows similar coverage.

**Product coverage**



⇒ Product coverage of buyer session is bigger than it of non-buyer session.

**Average duration**

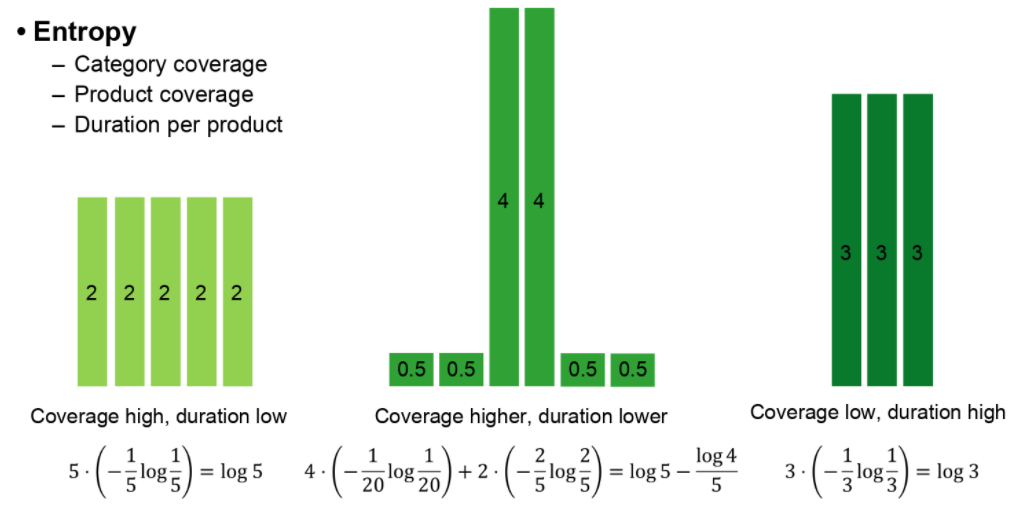


⇒ Average duration of buyer session is much bigger than it of non-buyer session.

Result: We can figure out that our second and third assumptions are correct but the category coverage is not an important factor. Also, we get the relation between product coverage and average duration.

* 1. **Entropy representation for buyer decision process**

We also analyzed Entropy of session while change the value of Product coverage and Average duration. Entropy is a good way to measure the uncertainty of specific state. The bigger the entropy, uncertainty of that state is bigger.



* 1. **Session representation for GAN**

Our GAN will learn the distribution.



There are two parts of distribution for that representation.

**Distribution for characteristic of buyers**

* + Category preference : (0, 0, 3, 0, 2, 0, 0, 1, 0, 0) // each number stands for the rank of the category.
  + Brand preference : between 0 and 1
  + Sale preference : between 0 and 1
  + Season : Summer or Winter (0 or 1)
  + High/Low purchaser : between 0 and 1
  + Phase : One of three phases(Need Recognition, Information Search, Alternative Evaluation) // expressed by entropy
  + Ex : (0, 0, 3, 0, 2, 0, 0, 1, 0, 0, 0.8, 0.2, 1, 0.5, 0.31)

Entropy

**Distribution for sequential data given buyer**

* + S=(click1, click2, click3, …, clickN)
  + click = (product, category, duration, purchase, price, quantity)
  + N = max(n) ⇒ S∈ matrix with 0’s

Based on this representation, we get the similarity between two session (b1 and b2) by Euclidean distance of them.

Example.

Then, the Euclidean distance between and (d()) is :

d() = 

However, we found that the value of category preference is bigger than other factors. In that case, the other factors except for category preference cannot be considered even though they are important information. Therefore, the values of above representation should be rescaled.

1. **Future Plan**

We decided the real session representation for GAN input. Therefore, choosing appropriate GAN model is left.

* 1. **Choosing GAN model**

There are four candidate models.

1. ACGAN with imbalanced training: Basic model is ACGAN and this model is trained with imbalanced batch which proportion is same with training data. i.e. buy : non-buy = 33:1.

2. ACGAN with balanced training: Basic model is ACGAN and this model is trained with balanced batch. i.e. buy : non-buy = 1:1.

3. WACGAN with imbalanced training: Basic model is WGAN + ACGAN and this model is trained with imbalanced batch.

4. WACGAN with balanced training: Basic model is WGAN + ACGAN and this model is trained with balanced batch.

We are going to test these models and find appropriate model for our research.

* 1. **Prediction and Explanation**

After generating fake buyer session, we are going to combine generated data and original data to train neural model to predict whether sessions in the test set is buyer or non-buyer. Then, we are going to explain the prediction with buyer decision process.

1. **Reference**
2. "Why Should I Trust You?" Explaining the Predictions of Any Classifier (<http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf>)
3. Session-based Recommendations with Recurrent Neural Networks (<https://arxiv.org/pdf/1511.06939.pdf>)
4. 1시간만에 GAN(Generative Adversarial Network) 완전 정복하기 (<https://www.youtube.com/watch?v=odpjk7_tGY0>)
5. Conditional Image Synthesis With Auxiliary Classifier GANs (<https://arxiv.org/pdf/1610.09585.pdf>)
6. eCommerceGAN : A Generative Adversarial Network for E-commerce (<https://arxiv.org/abs/1801.03244>)
7. Predicting Online Purchase Conversion for Retargeting (<https://dl.acm.org/citation.cfm?id=3018715>)
8. Predicting User Purchase in E-commerce by Comprehensive Feature Engineering and Decision Boundary Focused Under-Sampling (<https://dl.acm.org/citation.cfm?id=2813517>)
9. RecSys Challenge 2015 (<http://2015.recsyschallenge.com/challenge.html>)