# Recommending Repeat Purchases using Product Segment Statistics

Suvodip Dey
Dept. of Computer Science
Indian Institute of
Technology,Kharagpur
Kharagpur 721302, India
suvodip15@gmail.com

Pabitra Mitra
Dept. of Computer Science
Indian Institute of
Technology,Kharagpur
Kharagpur 721302, India
pabitra@cse.iitkgp.ernet.in

Kratika Gupta
SW Development Engineer
Flipkart Internet Private
Limited
Bangalore 560034, India
kratika.gupta@flipkart.com

## ABSTRACT

Repeat Purchases have become increasingly important in measuring customer's satisfaction and loyalty to e-commerce websites in regard to online shopping. In this paper, we first propose a model for estimating repeat purchase frequency in a given time period from a given product category using Poisson/Gamma model. Second, we estimate the purchase probabilities of different product types in a product category for each customer using Dirichlet model. Experimental results on data collected by a real-world e-commerce website show that it can predict a user's average repeat purchase frequency along with their product types with decent accuracy. We also argue that the output of our models can be used as prior information to enhance the performance of time-sensitive recommendation.

## Keywords

Repeat purchase, Poisson/Gamma model, Dirichlet model

## 1. INTRODUCTION

Repeat Purchase is a common phenomenon in online shopping which can depict a customer's satisfaction and loyalty to an e-commerce website. In this work we describe a model for predicting the number of repeat purchases along with their product types from a given product category, in a given time period, for a given customer. We also show that these predictions can guide existing systems for better time-sensitive recommendation.

Repeat purchases are generally predicted using time sensitive recommender system. Several studies have been done in the field of time-sensitive recommendation. Work in [3] represented the problem of recommending the right product at right time with the help of Hazards model in survival analysis. Work in [4] clusters user-product into groups having similar purchase behaviour using topic model and then filters out products from each cluster based on purchase time interval. Work in [1] addresses the problem with the help of self-exciting temporal point process(Hawkes process) where each point represents a purchase event.

There are several challenges in predicting repeat purchase events. The major hindrance is irregular purchases by customers. Model described in [3] works well when a customer purchases consistently from the e-commerce website. But real life scenarios are different. It has been observed that some customers are active for only a short time period and remains inactive for rest of the year. On the other hand frequency of purchases for majority of the customers are extremely low whereas for some customers it is uncharacteristically high. Another issue for predicting repeat purchase is data sparsity. In [1], parameters are estimated for each user-product pair. A low rank Hawkes process has been proposed as a solution and ample data is required to get a good estimate. Work in [4] tries to compute average purchase interval for each pair of products which not only increases the computational cost but also suffers from data sparsity.

One common assumption in all the existing time sensitive recommendation is that the current purchase is dependent on previous purchase/purchases. In context of repeat purchase, this assumption must hold true in ideal case. But analysis of real purchase data shows that this dependency assumption holds true for only a few customers. For the rest of the customers, it would be better to assume that purchases over a longer time periods(several weeks or months) are independent. These observations and repeat buying theory of Ehrenberg [2] serves as the motivation of our work.

For predicting number of repeat purchases along with their product types, we use a two stage top down procedure. The first stage predicts the number of repeat purchases for a given product category and time period. In the second stage, the model matches predicted repeat purchases with possible product types. If shampoo is a product category, then a group of shampoos with similar features form a product type. The model only provides a broad estimate about repeat purchases over a longer duration of time with decent accuracy. Thus it should not be used as a standalone recommendation system for predicting repeat purchase. Instead it can be used as a supporting system in parallel with existing recommendation system where the predicted information about repeat purchases and product types can be used as prior knowledge.

## 2. PROPOSED APPROACH

We propose a two stage top down method to achieve our

*RecSys '16, September 15-19, 2016, Boston , MA, USA*
© 2016 ACM. ISBN 978-1-4503-4035-9/16/09. . . $15.00
DOI: http://dx.doi.org/10.1145/2959100.2959145

goal. The first stage predicts the number of repeat purchases from a given product category in a given time period for a given user. The second stage matches the predicted repeat purchases with possible product types. We also show how this model can be used to enhance the performance of existing time-sensitive recommendation system.

## 2.1 Predicting Number of Repeat Purchases

In this stage we predict the number of repeat purchases. We build a Poisson/Gamma model to estimate the average purchase frequency of each customer for a given time period and product category. Here time period is longer - several weeks or months.

Before further detail, we state three assumptions behind the model. First, repeat purchases in one product category is independent of other product categories. Repeat purchases can be dependent across several product categories but we are restricting our scope to this independent assumption. Second, purchase frequency of each customer is independent. This assumption is pretty straightforward and can be verified from purchase data. The third and the most important assumption is that the successive purchases of a customer behave as if random over a longer duration of time. A closer observation on the purchase data clearly shows that the purchase frequency of a customer is very irregular over longer time period(several weeks or months). That is why the successive purchases in a single time period tends to be effectively independent and can be regarded as if random. Moreover, we observe a very low correlation between the average purchase frequency of different time periods. So we also assume that the average purchase frequency of different time periods are independent.

Now we describe our model for estimating average purchase frequency. The occurrence of a purchase event can be viewed as arrival of a customer to purchase a product. Then our objective can be represented as the prediction of number of times a customer will arrive/purchase in a given time period. As this is a count data problem, we assume that the purchase frequency of each customer follows a Poisson distribution for a given time period. Then according to our assumption, purchase instances of all the customers in a given time period are independent and sampled from Poisson distribution. We assume that mean of the Poisson distributions of all the customers in a given time period is sampled from a gamma prior. The basic model has been described in Figure 1. We build a separate version of the same model for different time period and product category.

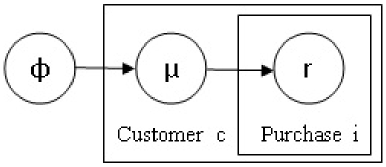

Figure 1: Distribution of each customer $c$ in a given time period $t$ and product category $m$, share information through prior $\phi$. $r_{c,i}$ is the total number of purchases in the $i^{th}$ observation by customer $c$ in time period $t$ from product category $m$. Average purchase frequency $\mu_c$ is the mean of the Poisson distribution for customer $c$ in time period $t$ for product category $m$. $\mu_c$ of all the customers are sampled from prior $\phi = Gamma(k, a)$ where $k$ and $a$ are shape and scale parameter respectively.

We use maximum likelihood estimate to infer the model parameters $\mu_c$ and $\phi$ or $\{k, a\}$ for each time period $t$ and product category $m$.

Let $D$ be the observed purchase frequency of all the customers for a given time period and product category. From our assumption, $r_{c,i} \sim Poisson(\mu_c)$ and $\mu_c \sim Gamma(k, a)$. Then we can write the likelihood function as

$$p(D|\phi) = \prod_c \frac{e^{-n_c\mu_c}\mu_c^{\sum_{i=1}^{n_c} r_{c,i}}}{\prod_{i=1}^{n_c} r_{c,i}!} \frac{{\mu_c}^{k-1}e^{-\frac{\mu_c}{a}}}{\Gamma(k)a^k}, \qquad (1)$$

where $n_c$ is the number of observations for customer $c$.

The posterior distribution($\theta_c$) of $\mu_c$ given $D$ follows a Gamma distribution.

$$p(\theta_c) = p(\mu_c|D) = Gamma(k', a'), \qquad (2)$$

where $k' = \sum_{i=1}^{n_c} r_{c,i} + k$ and $a' = \frac{a}{1+n_c a}$ .

Maximizing the likelihood in Equation 1 is equivalent to maximizing the log likelihood $L(\phi)$.

$$\begin{aligned} L(\phi) &= \ln p(D|\phi) \\ &= \sum_c -(n_c + \frac{1}{a})\mu_c + (\sum_{i=1}^{n_c} r_{c,i} + k - 1)\mu_c - \ln\Gamma(k) - k\ln(a) \end{aligned} \qquad (3)$$

We use Expectation Maximization algorithm to estimate the parameters. The process to infer parameters using EM algorithm is to iterate between the following E-step and M-step until convergence.

In the E-step, we infer $\mu_c$ as the expectation of the posterior distribution.

$$\mu_c = E[\theta_c] = k'a' \qquad (4)$$

In M-step, we infer $\phi$ or $\{k, a\}$ that maximizes the log likelihood $L(\phi)$ in equation 3.

## 2.2 Predicting Product Types

The objective of this stage is to predict the product types given the number of repeat purchases. Ideally the goal was to predict products from a given product category. As the number of products in a product category can be large and dynamic, we only predict product types. Generation of product types within a product category has been discussed later. The assumptions for this stage are - (1) Each customer's choices among the available product types follow a Multinomial distribution, (2) These choice probabilities follow a Dirichlet distribution across different customers.

Let $p_{c,j}$ denote the probability of purchasing product type $j$ by customer $c$ for a given product category. Let $x_{c,j}$ denote the number of purchases from product type $j$ by customer $c$ and $\sum_{j=1}^{k} x_{c,j} = m_c$. Then the probability mass function for the Multinomial distribution of customer $c$ can be written as

$$p(x_{c,1}, ...x_{c,k}|p_{c,1}, ...p_{c,k}, m_c) = \frac{m_c!}{\prod_{j=1}^{k} x_{c,j}!} \prod_{j=1}^{k} {p_{c,j}}^{x_{c,j}}, \qquad (5)$$

where $\sum_{j=1}^{k} p_{c,j} = 1$ .

The probability distribution of the choice probabilities of a customer $c$ given prior $\alpha = \{\alpha_1, ..., \alpha_k\}$ can be written as

$$p(p_{c,1},...,p_{c,k}|\alpha_1,\alpha_1,...,\alpha_k) = \frac{\Gamma(S)}{\prod_{j=1}^{k}\Gamma(\alpha_j)}\prod_{j=1}^{k}{p_{c,j}}^{\alpha_j-1}, \tag{6}$$

where $S = \sum_{j=1}^{k}\alpha_j$ .

Let D be the observed data of purchased product types by all customers. Then we can write the likelihood function as

$$p(D|\alpha_1,\alpha_2,...,\alpha_k) =$$
$$\prod_{c}\left\{\frac{m_c!}{\prod_{j=1}^{k}x_{c,j}!}\prod_{j=1}^{k}{p_{c,j}}^{x_{c,j}}\frac{\Gamma(S)}{\prod_{j=1}^{k}\Gamma(\alpha_j)}\prod_{j=1}^{k}{p_{c,j}}^{\alpha_j-1}\right\} \tag{7}$$

Posterior distribution($\theta_c$) of choice probabilities given D follows a Dirichlet distribution.

$$p(\theta_c|D) = Dirichlet({\alpha_1}',{\alpha_1}',...,{\alpha_k}') \tag{8}$$

where ${\alpha_j}' = \alpha_j + \sum_{{x_{c,j}}^{(i)}\in D}{x_{c,j}}^{(i)}$ .

Here we have to estimate the choice probabilities $p_{c,j}$ and prior $\alpha$ for each product category. Again we use Expectation Maximization algorithm in a similar fashion.

In the E-step we infer the choice probabilities from posterior distribution of each customer.

$$p_{c,j} = E[\theta_c] = \frac{{\alpha_j}'}{\sum_{i=1}^{k}{\alpha_i}'} \tag{9}$$

In M-step, we infer $\alpha$ that maximizes the log likelihood of equation 7.

A product type is a group of similar products within a single product category. In this work we used k-means clustering algorithm to generate product types. We used product price and brand equity as clustering features as they held good partitioning attribute for the products of our dataset. Let $x$ be total number of sales in a particular brand b. Let $y$ be the total number of purchases by the users of brand b from the entire product category of b. Then we can denote brand equity as $(y-x)/x$ where a lower value indicates higher brand equity and vice versa.

## 2.3 Application of the model

In this section we discuss how our model can be used to enhance the performance of existing time-sensitive recommendation system. Predicted product types can help reducing the search space of the recommendation. On the other hand, predicted number of repeat purchases can be used in various ways for guiding the time-sensitive recommendation. Here we discuss one such application in detail.

Let a recommender system is recommending a product at time $t$ for customer $c$ and estimates the probability of purchase as $q$. Let $t$ belongs to time period $T$. For example, if we assume time period of one month and $t$ is some date of January, then time period $T$ is the entire January month. Let the category of this product be $m$. Let $y_c$ be the number of products already purchased in $T$ by customer $c$ from category $m$. Let $\mu_c$ be the average purchase frequency of this customer in category $m$ and time period $T$. Then the probability of purchasing $(y_c+1)^{th}$ product given $\mu_c$ can be written as

$$p((y_c+1)|\mu_c) = \begin{cases} e^{-\frac{(y_c+1)-\mu_c}{\sigma}}, & \text{if } (y_c+1) > \mu_c \\ 1, & \text{otherwise} \end{cases} \tag{10}$$

We can now use $q$ and $p((y_c+1)|\mu_c)$ to get the updated purchase probability as

$$q' = q * p((y_c+1)|\mu_c) \tag{11}$$

The basic idea is to penalize the probability score when the number of purchase exceeds average purchase frequency. Here we have used an exponential kernel for this task.

Evaluation of this part requires data regarding actual recommendations along with their time. We could not run our evaluation due to lack of this ground data. Instead we give a theoretical justification about the possible impact on time-sensitive recommendation. Let a user has already purchased 3 diapers in a given time period. Let the average purchase frequency for this user in the given time period be 2. Now let a recommender system recommend another diaper to this user in the same time period. Surveys show that this type of scenario is frequent in real-world e-commerce websites which also hurts customer satisfaction with the recommender system. Thus, prior information regarding average purchase frequency can certainly help to tackle this type of situations. Penalizing the probability score is one such solution to avoid over optimistic recommendation. It can also be used in dealing with situations where number of recommendations are lesser than average purchase frequency.

# 3. EXPERIMENTAL RESULTS

## 3.1 Dataset

We have used purchase data collected on a real-world e-commerce website as our dataset. Training data contained purchase history from 2014-01-01 to 2014-12-31. Duration of test data was from 2015-01-01 to 2015-12-31. Analysis of the raw data suggested that repeat purchases in FMCG(fast-moving-consumer-goods) products are more frequent than other products. That is why we used sales data and product details of some popular FMCG products(Shampoo,Soap,Diaper and Deodorant) as our data source.

## 3.2 Evaluation Methodology and Result

We used two different methods to evaluate our two models as described next.

### 3.2.1 *Evaluation of predicted repeat purchase*

For evaluating the prediction of repeat purchase, we analyze the standard deviation of test data from the estimated average purchase frequency. Let $\mu_c$ and $x_c$ be the estimated and actual purchase frequency respectively of a customer $c$ for product category $m$ and time period $t$. Then we use the following equation to calculate the standard deviation of our model($\sigma_{Model}$) for time period $t$ and product category $m$

$$\sigma_{Model} = \sqrt{\frac{1}{N}\sum_{c=1}^{N}(x_c-\mu_c)^2} \tag{12}$$

Let $\sigma_{Mean}$ denotes the standard deviation of test data when average purchase frequency for each customer is estimated as the sample mean. We compute $\sigma_{Model}$ and $\sigma_{Mean}$

for different time period and product category. Results has been shown in Table 1 and 2.

Table 1: **Std Deviation for Time Period of 2 months**

| Period | Exp Type | Diaper | Deo | Shampoo | Soap |
|---|---|---|---|---|---|
| Jan-Feb | $\sigma_{Model}$ | 1.067 | 0.916 | 0.642 | 0.964 |
| | $\sigma_{Mean}$ | 1.293 | 1.110 | 0.804 | 1.147 |
| Mar-Apr | $\sigma_{Model}$ | 0.821 | 0.913 | 0.673 | 0.951 |
| | $\sigma_{Mean}$ | 1.066 | 1.130 | 0.871 | 1.128 |
| May-Jun | $\sigma_{Model}$ | 0.662 | 0.902 | 0.641 | 0.978 |
| | $\sigma_{Mean}$ | 0.909 | 1.113 | 0.816 | 1.201 |
| Jul-Aug | $\sigma_{Model}$ | 0.817 | 0.852 | 0.554 | 0.909 |
| | $\sigma_{Mean}$ | 0.924 | 0.901 | 0.772 | 0.941 |
| Sep-Oct | $\sigma_{Model}$ | 0.815 | 0.764 | 0.584 | 1.017 |
| | $\sigma_{Mean}$ | 0.941 | 0.803 | 0.846 | 1.155 |
| Nov-Dec | $\sigma_{Model}$ | 0.786 | 0.728 | 0.582 | 0.938 |
| | $\sigma_{Mean}$ | 0.879 | 0.859 | 0.788 | 1.014 |

Table 2: **Std Deviation for Time Period of 3 months**

| Period | Exp Type | Diaper | Deo | Shampoo | Soap |
|---|---|---|---|---|---|
| Jan-Mar | $\sigma_{Model}$ | 1.372 | 1.148 | 0.842 | 1.208 |
| | $\sigma_{Mean}$ | 1.653 | 1.402 | 1.039 | 1.421 |
| Apr-Jun | $\sigma_{Model}$ | 0.934 | 1.155 | 0.831 | 1.246 |
| | $\sigma_{Mean}$ | 1.247 | 1.415 | 1.038 | 1.489 |
| Jul-Sep | $\sigma_{Model}$ | 1.072 | 0.971 | 0.656 | 1.085 |
| | $\sigma_{Mean}$ | 1.301 | 1.318 | 0.989 | 1.228 |
| Oct-Dec | $\sigma_{Model}$ | 1.013 | 0.913 | 0.705 | 1.213 |
| | $\sigma_{Mean}$ | 1.269 | 1.245 | 0.954 | 1.343 |

Comparison shows that the value of $\sigma_{Model}$ is less than $\sigma_{Mean}$ for all the cases. Therefore our model gives better prediction of average purchase frequency in comparison to simply using sample mean for the same task. We performed the evaluation taking different time periods. Results for time period 2 and 3 months has been shown in Table 1 and 2 respectively. We observe same trend in results for different time periods which show that our model works better even if the time period is variable.

### *3.2.2 Evaluation of predicted product types*

For evaluating this section, we compute conversion rate of products using estimated purchase probability of product types. Here conversion rate is the ratio between number of products correctly recommended and number of products actually purchased. We compare the result with conversion rate using highest selling product types in a product category.

Let $T_c$ be the total number of purchases from product category $m$ by customer $c$. Let $P_c$ be the list of top $k$ product types based on estimated purchase probabilities for a given customer $c$ and product category $m$. Let $S_c$ be the total number of purchases from product types enlisted in $P_c$ by customer $c$. Let $f_m$ denote the average conversion rate of product types using $P_c$. Then $f_m$ can be written as

$$f_m = \frac{1}{N} \sum_{c=1}^{N} \frac{S_c}{T_c}, \tag{13}$$

where $N$ is total number of customers in test data.

Let $Q$ be the list of top $k$ highest selling product types from product category $m$ common to all users. Let $S_c{}'$ be the total number of purchases from product types enlisted in $Q$ by customer $c$. Let $g_m$ denote the average conversion rate of product types using $Q$. Then $g_m$ can be written as

$$g_m = \frac{1}{N} \sum_{c=1}^{N} \frac{S_c{}'}{T_c} \tag{14}$$

The results of $f_m$ and $g_m$ has been compared in Table 3. We performed this evaluation for different values of $k$ where each product category had 20 product types.

Table 3: **Evaluation of conversion rates**

| top k | Exp Type | Diaper | Deo | Shampoo | Soap |
|---|---|---|---|---|---|
| k=1 | $f_m$ | 0.618 | 0.338 | 0.434 | 0.411 |
| | $g_m$ | 0.482 | 0.217 | 0.256 | 0.255 |
| k=2 | $f_m$ | 0.816 | 0.497 | 0.598 | 0.564 |
| | $g_m$ | 0.814 | 0.352 | 0.335 | 0.412 |
| k=3 | $f_m$ | 0.892 | 0.621 | 0.708 | 0.656 |
| | $g_m$ | 0.853 | 0.471 | 0.493 | 0.631 |
| k=4 | $f_m$ | 0.931 | 0.724 | 0.791 | 0.749 |
| | $g_m$ | 0.891 | 0.652 | 0.547 | 0.704 |
| k=5 | $f_m$ | 0.951 | 0.801 | 0.832 | 0.814 |
| | $g_m$ | 0.917 | 0.726 | 0.765 | 0.731 |

Comparison shows that usage of estimated purchase probabilities results in better performance than using highest selling product types. Conversion rate of our model is dependent on the clustering of products. A more sophisticated clustering algorithm can be used for better accuracy.

## 4. CONCLUSION

In this paper, we developed two models for predicting number of repeat purchases and product types respectively. We used Poisson/Gamma model for estimating average purchase frequency whereas Dirichlet model for estimating purchase probability of a product type. We also argued that the output of our models can guide existing system for better time-sensitive recommendation. In the future, we want to explore the dependency of repeat purchases across different product categories.

## 5. REFERENCES

[1] N. Du, Y. Wang, N. He, J. Sun, and L. Song. Time-sensitive recommendation from recurrent user activities. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3474–3482. Curran Associates, Inc., 2015.

[2] A. Ehrenberg. *Repeat-buying: Facts, Theory, and Applications.* Charles Griffin Book. Griffin, 1988.

[3] J. Wang and Y. Zhang. Opportunity model for e-commerce recommendation: Right product; right time. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 303–312, New York, NY, USA, 2013. ACM.

[4] G. Zhao, M. L. Lee, and H. Wynne. Utilizing purchase intervals in latent clusters for product recommendation. In *Proceedings of the 8th Workshop on Social Network Mining and Analysis*, SNAKDD'14, pages 4:1–4:9, New York, NY, USA, 2014. ACM.