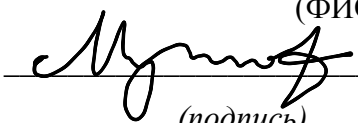


Федеральное государственное автономное образовательное учреждение  
высшего образования  
«Национальный исследовательский университет «Высшая школа экономики»  
Факультет компьютерных наук  
Образовательная программа Прикладная математика и информатика  
бакалавриат  
01.03.02 Прикладная математика и информатика

О Т Ч Е Т  
по преддипломной практике

Выполнил студент гр. 153  
Мугтасимов Данил Ренатович

(ФИО)  
  
(подпись)

Проверили:

PhD, профессор Жуков Леонид Евгеньевич

(должность, ФИО руководителя практики)

\_\_\_\_\_  
(дата)

2018/2019 уч.г.

## Содержание

Постановка задачи. Актуальность темы .....	3
Выбор методов решения поставленной задачи и его обоснование .....	4
Полученные результаты.....	5
Список изученной литературы.....	10

### **Постановка задачи. Актуальность темы.**

Внедрение рекомендательной системы является одним из подходов к увеличению продаж в бизнесе. Как правило, покупка нового продукта зависит от советов друзей, отзывов других пользователей или от сравнения с другими подобными товарами. Чтобы упростить процесс покупки и помочь клиенту сделать правильный выбор, необходимо внедрить систему рекомендаций. Механизм состоит из четырех шагов: (1) система идентифицирует клиента при каждой покупке (например, по карте лояльности); (2) предоставляет предпочтительные предложения продуктов, фильмов или ресторанов на основе предыдущего поведения; (3) контролирует реакцию на предложение; (4) улучшает список рекомендаций за счет увеличения информации о пользователе.

Целью работы является реализация системы рекомендаций, основанной на реальных данных розничной компании. Помимо задачи достижения производительности модели, стоит задача в реализации пользовательского интерфейса. Такой подход позволяет клиенту напрямую общаться с системой и получать информацию о персональных предложениях в режиме реального времени. В качестве такого канала связи будет использоваться чат-бот на платформе популярного мессенджера "Telegram".

Главные цели практики: 1. Изучить существующие алгоритмы рекомендательных систем. 2. Провести предобработку найденных данных. 3. Применить несколько разных типов моделей рекомендательных систем. 4. На основе выбранных метрик сравнить полученные модели. 5. Реализовать интерфейс работы с рекомендательной системой.

Индустрия полна хороших внедрений рекомендательных систем. Например, более 80 процентов контента на Netflix было просмотрено из личного списка предложений [1], а 60 процентов видео, просмотренного на YouTube, просматриваются из персональной рекомендации на домашней странице [2]. Тем не менее, это не относится к российскому

рынку розничных продаж. Несмотря на то, что такие системы могут принести дополнительную прибыль магазинам, количество успешно реализованных проектов в России невелико.

### **Выбор методов решения поставленной задачи и его обоснование.**

По сути, идея в том, чтобы предложить продукт клиенту. Эти продукты могут быть наиболее покупаемыми (например, трендовые) или могут иметь самые лучшие отзывы. Поэтому для составления такого списка рекомендаций достаточно данных об продажах каждого товара. В этом случае уже существуют хорошие базовые методы, созданные для того, чтобы рекомендовать наиболее популярные продукты. Этот способ хорошо подходит в случае, когда в системе появляется новый пользователь и для него отсутствует история покупок. Однако такой подход не позволяет системе развиваться на основе предыдущих предложений клиенту, а также не подразумевает предложение новинок. По этой причине мы рассмотрим более проработанные методологии.

**Метод коллаборативной фильтрации.** Данный метод основан на простой идее: если два клиента имеют схожие предпочтения, то у них будут схожими предпочтения и в будущем. Например, если от одного из двух довольно похожих пользователей пришел хороший отзыв о товаре, то с большой вероятностью этот продукт понравится второму клиенту. Преимущество методологии в том, что она не требует какой-либо специальной информации о товаре.

Чтобы понять, насколько предпочтения пользователей схожи друг с другом, сделаем следующее:

1. Построим user-item матрицу, где столбцами будут являться уникальные значения наименования товаров, а строкам будут соответствовать уникальные значения клиентов магазина.
2. Рассчитаем матрицу схожести продуктов. Будем использовать наиболее популярные подходы, такие как коэффициент Отиаи (Cosine similarity) и коэффициент корреляции Пирсона (Pearson similarity). Чтобы оценить

сходство между товарами, будем смотреть на всех клиентов, которые приобрели его. Так мы получим два соответствующих вектора, для расчета угла и расстояния между ними. Так, если косинус угла равен 1 или векторы совпадают друг с другом, то это означает полную схожесть между продуктами.

$$\text{Cosine similarity} = \cos(\varphi) = \frac{\bar{A} \cdot \bar{B}}{|\bar{A}| \cdot |\bar{B}|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

$$\text{Pearson similarity} = r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

3. Далее для каждого клиента мы прогнозируем вероятность купить продукт для тех товаров, которые он ранее еще не приобретал. Чтобы вычислить это, мы взвешиваем только что вычисленную меру сходства между целевым товаром и другими товарами, которые клиент уже купил.

Из минусов данного метода следует отметить высокий порог входа: если в системе не известны интересы пользователя, то построенные для него персональные рекомендации будут малоэффективны. Решить эту проблему предлагается несколькими способами: предлагать наиболее популярные товары среди всех клиентов используя Popularity model; классифицировать клиентов по возрастным показателям и предлагать наиболее востребованные продукты из соответствующего класса клиента.

### **Полученные результаты.**

В качестве данных использовались транзакции китайского продуктового магазина за период с ноября 2000 года по февраль 2001 года. Датасет содержит более 800 тысяч наблюдений.

В качестве обучающей выборки использовались 3 вида данных:

1. Исходный датасет с целевой переменной по количеству покупок конкретного товара
2. Датасет с добавленной dummy переменной равной единице, в случае если клиент когда-либо приобретал данный товар
3. Датасет с нормировкой по количеству покупок товара (min-max normalization) среди всех пользователей.

В качестве используемых моделей были взяты следующие:

1. Popularity model
2. Collaborative Filtering Model с использованием Cosine similarity
3. Collaborative Filtering Model с использованием Pearson similarity

Таким образом, мы получили 9 различных оценок полученных моделей. Сравним их показатели и выберем наиболее подходящую.

### Оценки моделей (1-3) на обучающей выборке 1 типа.

```
Popularity Model on Purchase Counts
+-----+-----+-----+
| cutoff | mean_precision | mean_recall |
+-----+-----+-----+
| 1 | 0.00014318442153493665 | 2.8636884306987325e-05 |
| 2 | 0.00014318442153493652 | 3.886434298805441e-05 |
| 3 | 9.545628102329142e-05 | 3.886434298805441e-05 |
| 4 | 7.159221076746826e-05 | 3.886434298805441e-05 |
| 5 | 8.591065292096246e-05 | 5.477372315860274e-05 |
| 6 | 7.159221076746835e-05 | 5.477372315860274e-05 |
| 7 | 6.136475208640137e-05 | 5.477372315860274e-05 |
| 8 | 7.15922107674684e-05 | 9.056982854233742e-05 |
| 9 | 7.954690085274287e-05 | 0.00013829796905398282 |
| 10 | 7.159221076746845e-05 | 0.00013829796905398282 |
+-----+-----+-----+
[10 rows x 3 columns]

Overall RMSE: 0.14419355268193434
```

Evaluate model Cosine Similarity on Purchase Counts

cutoff	mean_precision	mean_recall
1	0.04782359679266902	0.027942512177907275
2	0.033290378006872845	0.03626494138736405
3	0.02706185567010319	0.04230525166414334
4	0.02226517754868265	0.045050167002317204
5	0.01915807560137455	0.04809300641757925
6	0.016871897670866807	0.05059157457336369
7	0.015075274095892644	0.05238150095719277
8	0.013799398625429534	0.05414392453018335
9	0.01283886979763266	0.0563090258791334
10	0.011970217640320786	0.057928600780495346

[10 rows x 3 columns]

Overall RMSE: 1.0195117625892505

Evaluate model Pearson Similarity on Purchase Counts

cutoff	mean_precision	mean_recall
1	0.00014318442153493682	2.8636884306987352e-05
2	0.00014318442153493668	3.8864342988054366e-05
3	9.545628102329138e-05	3.8864342988054366e-05
4	0.00010738831615120242	8.65924834996998e-05
5	8.591065292096234e-05	8.65924834996998e-05
6	7.159221076746845e-05	8.65924834996998e-05
7	6.13647520864014e-05	8.65924834996998e-05
8	5.369415807560121e-05	8.65924834996998e-05
9	4.7728140511645614e-05	8.65924834996998e-05
10	5.72737686139749e-05	0.00015818469426716835

[10 rows x 3 columns]

Overall RMSE: 0.2791879905583871

## Оценки моделей (1-3) на обучающей выборке 2 типа.

Evaluate model Popularity Model on Purchase Dummy

cutoff	mean_precision	mean_recall
1	0.00014283673760891275	3.570918440222819e-05
2	7.141836880445637e-05	3.570918440222819e-05
3	4.76122458696377e-05	3.570918440222819e-05
4	0.00010712755320668477	0.00019441667063435343
5	0.00014283673760891332	0.0004800901458521823
6	0.00014283673760891348	0.0005038962687869983
7	0.00012243148937906824	0.0005038962687869983
8	0.00010712755320668447	0.0005038962687869983
9	9.522449173927526e-05	0.0005038962687869983
10	8.570204256534812e-05	0.0005038962687869983

[10 rows x 3 columns]

Overall RMSE: 0.0

Evaluate model Cosine Similarity on Purchase Dummy

cutoff	mean_precision	mean_recall
1	0.040565633480931264	0.019057996868309737
2	0.028281674046564863	0.02576639385309575
3	0.023044327000904558	0.03080223907248618
4	0.019461505499214313	0.03455799505292436
5	0.01674046564776454	0.03715932411475906
6	0.015045469694805512	0.04004258568963614
7	0.013528679576387061	0.041509723037361834
8	0.012319668618768716	0.04305887814672275
9	0.01129997301972728	0.04451025810831561
10	0.010598485930581365	0.046113770531710827

[10 rows x 3 columns]

Overall RMSE: 0.9946087139998798

Evaluate model Pearson Similarity on Purchase Dummy

cutoff	mean_precision	mean_recall
1	0.0	0.0
2	0.00014283673760891272	3.825984043095888e-05
3	9.522449173927557e-05	3.825984043095888e-05
4	7.141836880445636e-05	3.825984043095888e-05
5	8.57020425653483e-05	0.00010967820923541574
6	7.141836880445636e-05	0.00010967820923541574
7	6.121574468953404e-05	0.00010967820923541574
8	5.356377660334247e-05	0.00010967820923541574
9	4.761224586963773e-05	0.00010967820923541574
10	5.713469504356521e-05	0.00014538739363764385

[10 rows x 3 columns]

Overall RMSE: 1.0

## Оценки моделей (1-3) на обучающей выборке 3 типа.

Evaluate model Popularity Model on Scaled Purchase Counts

cutoff	mean_precision	mean_recall
1	0.0	0.0
2	0.0003512058066026694	0.00039803324748302534
3	0.0003902286740029662	0.0005541247170842112
4	0.00035120580660266926	0.0007882619214859915
5	0.00032779208616249097	0.0010223991258877705
6	0.0003512058066026688	0.0013345820650901413
7	0.0003344817205739712	0.001568719269491923
8	0.0003512058066026706	0.002036993678295484
9	0.00031218293920237195	0.002036993678295484
10	0.0002809646452821358	0.0020369936782954835

[10 rows x 3 columns]

Overall RMSE: 0.20297286225380898

Evaluate model Cosine Similarity on Scaled Purchase Counts

cutoff	mean_precision	mean_recall
1	0.002809646452821361	0.0014893727724446528
2	0.0022243034418169072	0.0022112958193501384
3	0.0022633263092171983	0.003273275282172503
4	0.0021657691407164685	0.004326892701980501
5	0.0021072348396160177	0.005329780394168103
6	0.0017950519004136418	0.005376607835048493
7	0.0019399939793290238	0.00671118990013864
8	0.0019023647857644547	0.007655543291225844
9	0.001795051900413642	0.00829942060333069
10	0.0016623741512526377	0.00847502350663203

[10 rows x 3 columns]

Overall RMSE: 0.19530710217026012

Evaluate model Pearson Similarity on Scaled Purchase Counts

cutoff	mean_precision	mean_recall
1	0.0004682744088035587	0.0002809646452821357
2	0.00046827440880355874	0.00047607898228361844
3	0.0006243658784047439	0.0011004448606883638
4	0.00046827440880355933	0.0011004448606883638
5	0.00042144696792320314	0.0013345820650901429
6	0.0003512058066026698	0.0013345820650901429
7	0.00033448172057397123	0.0014126277998907325
8	0.00029267150550222416	0.0014126277998907325
9	0.0002601524493353104	0.0014126277998907325
10	0.0002341372044017795	0.0014126277998907325

[10 rows x 3 columns]

Overall RMSE: 0.20282305840852813



Заметим, что наилучшие оценки Precision и Recall имеют модели при нормированной целевой переменной. Из оставшихся трех моделей оставим модель с наименьшим показателем RMSE, то есть Collaborative Filtering Model с использованием Cosine similarity.

Таким образом, рекомендации имеют следующий вид:

customerId	productId	score	rank
1104905	4711703122536	3.0	1
1104905	74570703074	3.0	2
1104905	4713645410122	2.0	3
1104905	4713645632036	2.0	4
1104905	4713045018096	2.0	5
1104905	2100035002364	2.0	6
1104905	20538538	2.0	7
1104905	4710498600847	2.0	8
1104905	4712172200015	2.0	9
1104905	8712045003565	2.0	10
418683	4711703122536	3.0	1
418683	74570703074	3.0	2
418683	4713645410122	2.0	3
418683	4713645632036	2.0	4
418683	4713045018096	2.0	5
418683	2100035002364	2.0	6
418683	20538538	2.0	7
418683	4710498600847	2.0	8
418683	4712172200015	2.0	9
418683	8712045003565	2.0	10

Полученный список уже используется в реализованном телеграмм боте @HseRecomSystemBot, который загружен на сервер и доступен в любое время.

Текст программы доступен по [ссылке](#), вместе с демо вариантом реализации интерфейса работы клиента с рекомендательной системой.

## **Список изученной литературы.**

- [1] Carlos A Gomez-Uribe and Neil Hunt. 2016. The netflix recommender system: Algorithms, business value, and innovation. TMIS 6, 4 (2016), 13.K. Elissa, “Title of paper if known,” unpublished.
- [2] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, and Dasarathi Sampath. 2010. The YouTube Video Recommendation System. In Recsys.
- [3] Leidy Esperanza Molina and Sandjai Bhulai. “Recommendation System for Netflix”. In: (2018).
- [4] B. Pradel, S. Sean, and N. Usunier. A case study in a recommender system based on purchase data categories and subject descriptors. In KDD, pages 377–385, 2011.
- [5] Shuai Zhang, Lina Yao, and Aixin Sun. 2017. Deep Learning based Recommender System: A Survey and New Perspectives. arXiv:1707.07435 (2017).M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.