УДК 330, 339, 65

Соловьева А.И.,
учащаяся школы № 67,
Россия, г. Москва
Рихтер А.А.,
канд. техн. наук, м.н.с., научно-исследовательский институт аэрокосмического мониторинга «АЭРОКОСМОС»
Россия, г. Москва

АНАЛИЗ МАГАЗИННЫХ ЧЕКОВ ДЛЯ ПРОГНОЗИРОВАНИЯ КЛИЕНТСКИХ ПРЕДПОЧТЕНИЙ

Аннотация: *Работа посвящена созданию программы, прогнозирующей выбор покупателей в магазинах. Написанный программный код позволяет проанализировать потребительские корзины и на основе обработанных данных осуществить подсчет вероятности выбора того или иного товара и сформулировать прогнозы по единичным продуктам для корректировки плана закупок. Разработанная схема показала свою эффективность, а именно точный и корректный для данного этапа результат. Она может быть полезна как покупателю, получающему персональное предложение, так и продавцу для увеличения объема продаж и повышения доверия клиентов.*

Ключевые слова: *программы лояльности, магазинные чеки, клиентские предпочтения, виртуальные данные, транзакции, программирование, маркетинговые технологии.*

Annotation: *The work is devoted to the creation of a program that predicts the choice of customers in stores. The written program code allows you to analyze consumer baskets and, based on the processed data, calculate the probability of choosing a product and formulate forecasts for individual products to adjust the procurement plan. The developed scheme has shown its effectiveness, namely, the exact*

and correct result for this stage. It can be useful as a buyer who receives a personal offer, and the seller to increase sales and increase customer confidence.

Keywords: *loyalty programs, store checks, customer preferences, virtual data, transactions, programming, marketing technologies.*

Введение

Одной из сфер, в которой базы данных постоянно увеличиваются, является продажа товаров и услуг. Для потребителя и продавца в условиях постоянного растущих спроса и предложения немаловажными становятся программы лояльности. У многих магазинов есть накопительные карты, бонусные программы, скидки на определенные продукты. Но в большинстве случаев эти программы построены по одному стандарту, и из-за того, что они нацелены сразу на всю аудиторию, возникает целая группа убыточных клиентов, скидка перестает стимулировать потребление и люди чувствуют неудовлетворённость, т.к. на их личные предпочтения не было обращено внимание.

Данные проведенного опроса клиентов различных магазинов и продавцов товаров и услуг позволили выделить следующее несоответствие: 80% компаний считают, что они информированы о своих клиентах и знают, что им предложить, при этом лишь 15% клиентов согласны с этим утверждением [1].

Необходимость разрешения указанного противоречия подтверждает актуальность работы. Целью работы является разработка вспомогательного алгоритма для проведения прогноза потребности в покупке товара в соответствие с анализом чеков и текущего состояния покупательной способности.

В соответствии с поставленной целью были определены задачи работы: 1) осуществить отбор данных о покупках в различных магазинах для дальнейшего анализа и преобразования; 2) проанализировать существующие методы работы с данными о покупках в магазинах; 3) разработать концепцию анализа данных магазинных чеков для дальнейшего прогнозирования покупок; 4) выбрать оптимальный для поставленной цели язык программирования; 5) разработать

алгоритм и получить первичные экспериментальные данные, проверить их достоверность; 6) определить направления дальнейшей работы над проектом.

Ход работы

В табл. 1 представлена дорожная карта (траектория) проектно-исследовательской работы.

Таблица 1.

Дорожная карта проектно-исследовательской работы (по месяцам)

	09.2018	10.2018	11.2018	12.2018	01.2019	02.2019
Постановка задачи. Анализ рынка						
Поиск данных и разработка начальной схемы						
Первичная обработка данных						
Поиск и интеграция алгоритма подсчета вероятностей						
Преобразование вероятностей в таблицы						
Поиск заказчиков, сравнение с конкурентами						

Анализ и работа с виртуальными данными

В настоящее время в мире насчитывается около 33 Збайт виртуальных данных, а по прогнозам к 2025 году их количество возрастет уже в пять раз до 175 Збайт.

В соответствии с такими же прогнозами 11.8% составит среднегодовой темп роста мирового рынка услуг по анализу данных для бизнеса в период 2016-

2020 гг. По оценкам IBM, к 2020 году количество рабочих мест для специалистов по анализу данных вырастет на 28%. [2] Уже сейчас крупные компании прибегают к анализу данных, и наиболее распространенным направлением для аналитики является поведение клиентов.

В исследовании были использованы данные о транзакциях в магазинах, выложенные для участников хакатона Boosters [3]. После регистрации на нужном хакатоне, скачиваются тестовые данные (выборка реальных данных), которые можно использовать не только для участия в хакатоне, но и в личных целях. На рис. 1. мы видим загрузку основных библиотек (см. ниже). Данные были получены в виде таблицы, показанной на рис. 2, состоящей из времени покупки, номера чека, названия продукта, его стоимости. На данном этапе была проведена обработка данных ID(fsId) покупателей и названий продуктов (name), которые были ими выбраны. Если схема оказывается успешной при таком малом количестве признаков, т.е. результат будет около 30%, то данные можно считать корректными. Также можно будет повышать точность, рассматривая корреляцию с датой, суммой чека и другими внешними факторами. Т.к. данных много, для прототипа и тренировки взяты первые 106 продуктов для получения эффективной схемы с наименьшей затратой времени на обработку.

```
import pandas as pd
import gc
import math

%pylab inline
```

```
Populating the interactive namespace from numpy and matplotlib
```

Рисунок 1. Загрузка основных библиотек

```
In [4]: data=pd.read_csv('FNS_category.csv', sep=';')
```

```
In [5]: data
```

	dateTimeField	dateTimeFieldHour	documentId	fsId	name	unit	price	quantity	sum
0	2017-02-02	01	101	8710000100198805	Губка FRESH антибактериальная Арт. F3065	NaN	4999	1.0	4999
1	2017-02-02	01	114	8710000100198805	Жев рез DIROL X-Fresh арбузный лед 16г	NaN	3499	1.0	3499
2	2017-02-02	01	262	8710000100199156	Фисташки JAZZ соленые 40	NaN	12600	1.0	12600
3	2017-02-02	01	262	8710000100199156	Пакет-майка ПЕРЕКРЕСТОК 30х55с	NaN	400	1.0	400
4	2017-02-02	01	400	8710000100098775	Корм д/кошек Вискас пауч д/котят паште	NaN	1790	1.0	1790

Рисунок 2. Загрузка данных и их представление

```
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori
```

Рисунок 3. Загрузка библиотеки Mlxtend

Выбор языка программирования и библиотек для работы с данными

Эксперты TIOBE назвали Python языком программирования 2018 года – по их словам, Python стал неотъемлемой частью многих IT-сфер [4]. Он лидирует по использованию в статистике, научных вычислениях. По данным на февраль 2019 года, Python занимает 3-е место в общем мировом рейтинге языков программирования.

Для того, чтобы работать непосредственно с табличными данными, были использованы библиотеки numpy, pandas [5-6] и mlxtend [7]. Numpy необходим для работы с числовыми значениями, pandas облегчает работу с таблицами, а mlxtend считает вероятности выбора тех или иных связок. На рис. 3. – загрузка библиотеки Mlxtend и нужных блоков.

```
data['name']=data['name'].str.lstrip('1234567890')
data['name']=data['name'].str.lstrip(' ')
data['name']=data['name'].str.lstrip('1234567890')
data['name']=data['name'].str.rstrip('65х40см')
data['name']=data['name'].str.rstrip('кг1КГВСПЕСОК')
data['name']=data['name'].str.rstrip('-')
data['name']=data['name'].str.rstrip('ПНД 12')
data['name']=data['name'].str.rstrip('ВЕС')
data['name']=data['name'].str.rstrip('-1234567890')
```

Рисунок 4. Код для очистки записи продуктов

АИ	пакет
Пакет ПЯТЕРОЧКА	а
ПАКЕТ-МАЙКА ДИКСИ 38Х	бананы
809 Бананы	мандарины
95ULT	ult
: 3305976 Пакет ПЕРЕКРЕСТОК майка	картофель
БАНАНЫ	яйц
* 3182856 Яйца кур.С1 стол.фас.10шт	лимоны

Рисунок 5. Пример очистки записи продуктов

```
te = TransactionEncoder()
te_ary = te.fit(dataset).transform(dataset)
df1 = pd.DataFrame(te_ary, columns=te.columns_)
df1
```

апельсины	вода	йогурт	лук	мандарины	молоко	салат	хлеб	яблоки	яйц
False	False	True	True	True	True	False	True	False	True
False	False	True	True	True	False	True	True	False	True
False	False	False	False	False	True	True	False	True	True
False	True	True	False	True	True	True	False	False	False
True	False	False	True	True	False	True	False	False	True

Рисунок 6. Преобразование таблицы с чеками

В связи с тем, что требуется разработка программы с возможностью её дальнейшего внедрения в маркентинговые технологии и производство, логичнее выбрать более эффективные, быстрые и проверенные инструменты. Заметим, что можно обойтись и без этих библиотек, прописав их функции самостоятельно.

Из-за отсутствия единого стандарта записи товаров, в системах разных магазинов они записываются по-разному. Это значит, что для эффективной работы с этими данными, нужно выделить чистые наименования продуктов (без дополнительной информации). Названия товаров приведены к стандартному виду: нижний регистр, все на русском, снята лишняя информация (цифры, знаки препинания и т.д.).

Так, на рис. 4 показан код для очистки записи продуктов, а на рис. 5 – результат преобразований. Каждой строчке таблицы соответствует один

продукт. Если в столбце fsID номер повторятся в других строках, значит эти продукты были в одной корзине.

Был создан словарь, состоящий из сгруппированных продуктов в корзинах по их номерам с помощью функции groupby. Затем из полученного словаря были сделаны списки, в которых продукты также были собраны по корзинам. Для примера рассмотрим строку 0 на рис. 6. Мы видим, что на пересечении таких товаров, как йогурт, лук, мандарины, молоко, хлеб и яйца, написано True. Это говорит о том, что именно эти товары покупатель и выбрал для своей корзины.

Для работы с библиотекой mlxtend полученные списки были преобразованы в таблицу, где по горизонтали отложены все виды продуктов, по вертикали – номер чека, а на пересечении – True или False, в зависимости от наличия товара в корзине.

Просматривается вероятность выбора продукта, считая отношение выбранных связок ко всем товарам, которые входят в состав этой связки.

0.294779	(лукрепчатый, бананы)
0.297992	(мандарины, бананы)
0.283936	(картофель, лукрепчатый)
0.284739	(мандарины, картофель)

Рисунок 7. Пример вывода вероятностей выбранных связок

В итоге выводится таблица, состоящая из двух столбцов (рис. 7.). В одном находится связка продуктов, в другом вероятность выбора этой связки продуктов (от 0 до 1, 1 – это 100%). Это ассоциации, то есть не просто самые часто встречающиеся продукты, а именно их связь друг с другом. Например, если в связке находится хлеб и молоко, то с вероятностью 28.4% мандарины и картофель возьмут вместе, а не по отдельности. Поэтому если человек взял хлеб и вероятность выбора его с молоком окажется выше заданного порогового значения, то ему можно предложить хлеб.

Заказчики

Можно сделать вывод, что при внедрении этой программы в магазины, выгоду будут получать как продавцы, так и клиенты. Планируется предлагать разработанную схему крупным магазинам, где люди покупают достаточное

количество товаров. Это может быть полезно для того, чтобы программа могла рассчитывать более точные и необходимые потребности покупателей. Для примера, можно взять любой сетевой супермаркет, средний чек которого варьируется между оптовой закупкой на месяц вперед и стоимостью в минимаркетах, в которых покупают по несколько необходимых продуктов. Магазинам в свою очередь остаётся изменить ассортимент на кассах, оставляя продукцию, которая будут предлагаться чаще всего (для каждого магазина устанавливается свой товарный ассортимент).

Аналоги и конкуренты

Определена область торговых компаний, для которой будут проведены исследования ассортимента и покупательских предпочтений. Для сравнения производился выбор из тех магазинов, которые соответствуют заданным критериям (средний чек, сетевой магазин, продолжительность и стабильность на потребительском рынке), наиболее приемлемыми оказались Перекресток [8] и Избенка [9]. Программы лояльности этих магазинов являются представителями основных существующих и используемых программ, и были сопоставлены с разработанной программой лояльности (табл. 2).

Таблица 2.

Критерии сравнения программ лояльности

	Перекресток	**Избенка**	**Разработанная программа**
Способ хранения данных	Карты и купоны на чеках или фишки	Карта и приложение	Все в базах данных
Персональные предпочтения	Нет	Человек сам выбирает продукт для скидки, нужно делать регулярно	Да

Внесение корректировок в программу	Редко	Редко	При каждом новом чеке обновляются данные
Способ поощрения клиента	Накопление баллов	Накопление баллов	Скидка сразу при покупке

Можно заметить, что несмотря на достаточную развитость используемых в настоящее время программ лояльности, данная программа имеет ряд преимуществ, в частности, персональный подход к покупателю, отсутствие носителя информации о покупателе и необходимости заполнения форм для получения поощрений.

Выводы

Таким образом, была разработана рабочая программа, которая соответствует поставленной цели. Процент выбора определенных связок в среднем составляет 25-30%, что и требовалось на этом этапе. Есть прогнозы для единичных продуктов, что также может быть полезным для магазинов с точки зрения корректировки закупки товаров. В разработке находится алгоритм для поиска связок с упоминанием конкретного продукта, а также более общего алгоритма с учётом тенденций рынка, потребностей потребителей и продавцов.

Разработанная схема оказалась эффективной с точки зрения её практического применения, планируется её усовершенствование. Чтобы увеличить точность прогноза, в настоящий момент рассматривается более объемная выборка, а также корреляция с другими факторами, предоставленными в таблице (дата покупок, время суток, итоговая сумма чека, стоимость товаров, категории товаров) и внешними факторами (продажи сезонных товаров, спрос на которые зависит от праздников или времени года, деление магазинов на разные уровни качества и их контингент). Планируется создание страницы, которая будет являться прототипом будущего сервиса для того, чтобы

продемонстрировать заказчикам, как могла бы выглядеть эта система в их магазинах.

Анализ данных – важная проблема современного мира, особенно в области торговых отношений, клиентского сервиса и информационных и коммуникационных технологий. Продвигая эти технологии, с применением данного подхода можно принести не только выгоду частным компаниям, но также и задать стимул и ориентир для всей сферы клиентского сервиса. Были достигнуты все поставленные задачи и цели, а именно была разработана готовая к использованию программа, которая считает вероятности выбора связок продуктов, на основе чего можно предлагать клиенту товар, в котором он нуждается.

ИСПОЛЬЗУЕМЫЕ ИСТОЧНИКИ

1. Маркетинг-Микс Моделирование [Электронный ресурс]. URL: https://www.nielsen.com/ru/ru/solutions/capabilities/marketing-mix.html (дата обращения: 10.11.2018).
2. BigData&AnalyticsHub [Электронный ресурс]. URL: https://www.ibmbigdatahub.com/blog/quant-crunch-demand-data-science-kills (дата обращения: 10.11.2018).
3. Boosters.pro [Электронный ресурс]. URL: https://boosters.pro/championship/evotor1 (дата обращения: 10.11.2018).
4. TIOBE [Электронный ресурс]. 2019. URL: https://www.tiobe.com/tiobe-index (дата обращения: 10.11.2018).
5. Хейдл М. Изучаем pandas. Высокопроизводительная обработка и анализ в Python./ Майкл Хейдл. – Издательство: ДМК-Пресс, 2018 г. – 438 с.
6. Coursera [Электронный ресурс]. URL: https://www.coursera.org/specializations/jhu-data-science (дата обращения: 10.11.2018).
7. MLxtend [Электронный ресурс]. URL: http://rasbt.github.io/mlxtend/#contact

8. Перекрёсток [Электронный ресурс]. URL: https://www.perekrestok.ru/info/pravila_club_perekrestok (дата обращения: 10.11.2018).
9. ВкусВилл [Электронный ресурс]. URL: https://vkusvill.ru/loyalty-program (дата обращения: 10.11.2018).