САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

КАФЕДРА «МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ЭНЕРГЕТИЧЕСКИХ СИСТЕМ»

Григорян Ева Артуровна

Выпускная квалификационная работа бакалавра

Оценка сетевой ценности клиента

Направление 01.03.02 «Прикладная математика и информатика»

Научный руководитель,
кандидат физ.-мат. наук,
доцент
Лежнина Е. А.

Санкт-Петербург
2020

Содержание

Введение

За последние несколько лет рынок поменялся до неузнаваемости. Причиной тому стал технический прогресс. Сейчас мы не можем представить свою жизнь без онлайн-шоппинга, онлайн-кинотеатров и много другого, что подарил нам интернет. Естественно, и область маркетинга потерпела большие перемены: теперь мы не обращаем внимание на баннеры на улице или рекламу по телевизору, наш взор устремлен в телефон или ноутбук. Так как же обратить внимание владельца гаджета на нужный нам продукт, как не прогадать? За ответ на эти вопросы отвечает интеллектуальный анализ данный (Data mining). Одна из задач интеллектуального анализа данных – помочь компании определить потенциальных клиентов рынка.

Корпорации тратят баснословные суммы на маркетинг. Но как же понять оправдан он или нет? Если ожидаемая прибыль от клиента больше, чем стоимость маркетинга для нее, то маркетинг выполнил свою цель. Раньше работы в этой области в основном внутреннюю ценность клиента. Внутренняя ценность клиента – это ожидаемая прибыль от продаж ему. В своей работе я предлагаю также рассматривать сетевое значение клиента: ожидаемую прибыль от продаж других клиентов, которым он может предложить купить товар или повлиять на другого клиента.

Рассмотрим основные виды маркетинга: прямой и массовый. Прямой маркетинг – это вид продвижения, который предполагает передачу информации о продукте, услуге или компании непосредственно клиенту. К стратегиям прямого маркетинга относятся email рассылки, печатная и онлайн реклама, маркетинг на основе баз данных и прочее. В свою очередь массовый маркетинг предполагает сознательное игнорирование различий между покупателями, для продвижения товара используется единая стратегия коммуникации.

В отличие от массового маркетинга, в котором продукт предлагается всем потенциальным клиентам без разбора, в прямом маркетинге сначала

пытаются выбрать клиентов, которые с большей вероятностью принесут прибыль и работают только с ними [1]. Сбор данных играет ключевую роль в этом процессе за счет создания моделей, которые предсказывают реакцию клиента, учитывая его поведение в прошлом и любую доступную демографическую информацию [2]. При успехе, этот подход может значительно увеличить прибыль компании [3]. Одним из основных ограничений является то, что этот подход предполагает, что каждый клиент принимает решение о покупке независимо от всех других покупателей. В действительности, решение человека купить продукт часто сильно зависит от его друзей, убеждения, партнеров и т. д.

Маркетинг, основанный на системе «из уст в уста» может быть гораздо более экономичным, чем более традиционный метод, поскольку он использует клиентов, чтобы выполнить большую часть продвижения [4]. Этот тип маркетинга получил название вирусного маркетинга из-за своего сходства с распространением эпидемии. Сейчас его используют большое количество компаний, особенно в интернет-секторе.

Игнорирование сетевых эффектов при принятии о том, какие клиенты будут покупать товар, могут привести к неоптимальным решениям. Клиент, чья внутренняя стоимость ниже, чем стоимость маркетинга, на самом деле может стоить маркетинга, когда учитывается ее сетевая ценность. Наоборот, маркетинг для прибыльного клиента может быть избыточным, если сетевые эффекты уже делают очень вероятной покупку товара.

Хоть определение сетевой ценности клиента на первый взгляд является чрезвычайно сложной задачей, эта величина зависит не только от самого клиента, но и от потенциала его в сети и состояния самой сети. В результате, маркетинг с сильными сетевыми эффектами является либо хитом, либо провалом. Многие компании инвестируют значительные средства, чтобы приобрести клиентов, которые будут основой сети, что иногда приводит к банкротству, когда желаемы сетевые эффекты не реализуются. С другой

стороны, некоторые компании являются более успешными, чем ожидалось. Таким образом, основа действий на сетевых рынках может значительно снизить риск компаний на этом рынке.

Увеличение использования интернета привело к наличию большого количества данных, из которых может быть добыта необходимая сетевая информация. В своей работе я предлагаю общую схему для этого и для использования результатов оптимизации выбора клиентов на рынок. Решение основано на моделировании социальной сети в виде марковых случайных полей, где вероятность покупки каждого клиента зависит как от внутреннего желания приобрести продукт, так и от влияния других клиентов. Затем я фокусируюсь на базах данных, как источнике данных для интеллектуальных сетей влияния. Я применяю математическую модель к общедоступной базе данных MovieLens [5] с рейтингом 9 тысяч фильмов и демонстрирую ее преимущества по сравнению с традиционным прямым маркетингом.

Обзор литературы

Тема маркетинга в социальных сетях достаточно актуальна и популярна, поэтому многие авторы уделяют ей достаточное внимание.

В книге «The Complete Database Marketer: Second Generation Strategies and Techniques for Tapping the Power of Your Customer Database» автора Arthur Hughes [1] рассматривается вопрос о необходимости маркетинговой базы данных. Ее используют для установления и поддержания выгодных пожизненных отношений с клиентом. Независимо от того, что производит компания, для эффективного маркетинга нужна база клиентов, чтобы быть связанными с постоянными клиентами, которые хотят знать о новых предложениях, а также чтобы понять поведение клиентов и общаться с ними.

В статье Charles X. Ling и Chenghui Li «Data Mining for Direct Marketing: Problems and Solutions» [2] рассматривается процесс извлечения данных из баз данных для прямого маркетинга и возникающие при этом проблемы.

Как меру ценности клиента, которую я буду использовать для постоения модели и вычислений, я выбрала стоимость жизни клиента, но такой подход не нов. В свою очередь, D. R. Mani, James Drew, Andrew Betz и Piew Datta [6] построили модель для точного прогнозирование стоимости жизни клиента.

В книге Губанова Д. А. «Социальные сети: модели информационного влияния, управления и противоборства» [7] подробно описаны результаты исследования математических моделей социальных сетей, а также методы влияния членов социальных сетей друг на друга.

Мной было принято решение, работать с такой рекомендательной системой как коллоробтивная система фильтрации. В статье John S. Breese, David Hackerman и Carl Kadie «Empirical Analysis of Predictive Algorithms for Collaborative Filterig» [8] она подробно описана. В этой работе рассмотрены несколько алгоритмов для рекомендательной системы, включая методы

основанные на коэффициентах корреляции, которые я и использовала для своей модели.

Содержательная постановка задачи

В своей работе я рассматриваю рынок, как социальную сеть, включающую в себя пользователей онлайн-кинотеатра. В моем распоряжении информация о фильмах (жанр) и рейтингах, которые им поставили пользователи

Цель: разработать модель для определения сетевого значения клиента

Задача: вычислить сетевое значение клиента

Гипотеза: прямой маркетинг является лучшей стратегией, чем массовый.

Математическая постановка задачи

Рассмотрим множество n потенциальных клиентов и пусть X_i - логическая переменная, которая принимает значение 1, если клиент i покупает продукт, который продается, и 0 в противном случае. Далее для простоты изложения за X_i будем обозначать и самого i-ого клиента. Пусть соседями являются X_i клиенты, которые непосредственно влияют на X_i: $\mathbf{N}_i = \{X_{i,1}, \dots, X_{i,n_i}\} \subseteq \mathbf{X} - \{X_i\}$, где $\mathbf{X} = \{X_1, \dots, X_n\}$. Другими словами, X_i не зависит от $\mathbf{X} - \mathbf{N}_i - \{X_i\}$. Пусть $\mathbf{X}^k(\mathbf{X}^u)$ – клиенты, чье значение (то есть, купили ли они продукт) известны (неизвестны), и пусть $\mathbf{N}_i^u = \mathbf{N}_i \cap \mathbf{X}^u$. Предположим, что продукт описывается набором характеристик $\mathbf{Y} = \{Y_1, \dots, Y_m\}$. Пусть M_i - переменная, представляющая маркетинговое действие для клиента i. Например, M_i может быть логической переменной, где $M_i = 1$, если клиенту предлагается определенная скидка, $M_i = 0$ в противном случае. В качестве альтернативы M_i может собой представлять собой непрерывную переменную, указывающую размер предлагаемой скидки, или номинальную переменную, указывающее какое из нескольких возможных действий принято. Пусть $\mathbf{M} = \{M_1, \dots, M_n\}$.

Необходимо смоделировать и оценить сетевое значение клиента.

Математическая модель

Модель построена на основе подхода, описанного в статье P. Demingas, M. Richardson [9].

Для всех $X_i \notin \mathbf{X}^k$

$$
\begin{aligned}
P(X_i|\mathbf{X}^k, \mathbf{Y}, \mathbf{M}) & \\
&= \sum_{C(\mathbf{N}_i^u)} P(X_i, \mathbf{N}_i^u \big| \mathbf{X}^k, \mathbf{Y}, \mathbf{M}) \\
&= \sum_{C(\mathbf{N}_i^u)} P(X_i|\mathbf{N}_i^u, \mathbf{X}^k, \mathbf{Y}, \mathbf{M})\ P(\mathbf{N}_i^u \big| \mathbf{X}^k, \mathbf{Y}, \mathbf{M}) \\
&= \sum_{C(\mathbf{N}_i^u)} P(X_i|\mathbf{N}_i,\ \mathbf{Y}, \mathbf{M}) P(\mathbf{N}_i^u \big| \mathbf{X}^k, \mathbf{Y}, \mathbf{M})
\end{aligned}
$$

где $C(\mathbf{N}_i^u)$ - множество всех возможных комбинаций неизвестных соседей X_i (то есть множество $2^{|\mathbf{N}_i^u|}$ всевозможных присваиваний 0 и 1 им). Далее мы аппроксимируем $P(\mathbf{N}_i^u \big| \mathbf{X}^k, \mathbf{Y}, \mathbf{M})$ по его максимальной оценке энтропии с учетом предельных вероятностей $P\big(X_j \big| \mathbf{X}^k, \mathbf{Y}, \mathbf{M}\big)$, $X_j \in \mathbf{N}_i^u$ [10].

Это дает

$$
P(X_i|\mathbf{X}^k, \mathbf{Y}, \mathbf{M}) = \sum_{C(\mathbf{N}_i^u)} P(X_i\ |\mathbf{N}_i, \mathbf{Y}, \mathbf{M}) \prod_{X_j \in \mathbf{N}_i^u} P\big(X_j \big| \mathbf{X}^k, \mathbf{Y}, \mathbf{M}\big) \qquad (1)
$$

Множество переменных $\mathbf{X}^u$ с совместной вероятностью, обусловленной $\mathbf{X}^k$, $\mathbf{Y}$ и $\mathbf{M}$, описанное уравнением (1), является примером случайного Маркого поля [11, 12, 13]. Так как уравнение (1) выражает вероятности $P(X_i|\mathbf{X}^k, \mathbf{Y}, \mathbf{M})$ как функцию от них самих, то его можно применять итеративно, чтобы найти их значения, начиная с подходящего начального присваивания. Эта операция гарантированно сходится к локальным значениям, если начальное присваивание достаточно близко к ним [10]. Естественным выбором для инициализации является использование вероятностей без сети $P(X_i|\mathbf{Y}, \mathbf{M})$.

Заметим, что число членов в уравнении (1) является экспоненциальным по числу неизвестных соседей X_i-ого. Если это число мало (например, 5), то это не должно быть проблемой; в противном случае требуется приблизительное решение. Стандартным методом для этой цели является выборка Гиббса [14]. Альтернатива, основанная на эффективном алгоритме k-кратчайшего пути представлена в [15].

Учтем, что $\mathbf{N}_i$, $\mathbf{Y}$ и X_i должны быть независимыми от маркетинговых действий для других клиентов. Используя наивную байесовскую модель как функцию от $\mathbf{N}_i$,$Y_1, \ldots, Y_m$ и M_i [16],

$$P(X_i|\mathbf{N}_i, \mathbf{Y}, \mathbf{M}) = P(X_i|\mathbf{N}_i, \mathbf{Y}, M_i) = \frac{P(X_i)P(\mathbf{N}_i, \mathbf{Y}, \mathbf{M}|X_i)}{P(\mathbf{N}_i, \mathbf{Y}, M_i)}$$

$$= \frac{P(X_i)P(\mathbf{N}_i|X_i)P(M_i|X_i)}{P(\mathbf{N}_i, \mathbf{Y}, \mathbf{M})} \prod_{k=1}^{m} P(Y_k|X_i)$$

$$= \frac{P(\mathbf{N}_i|X_i)P(M_i|X_i)}{P(Y, M_i|\mathbf{N}_i)} \prod_{k=1}^{m} P(Y_k|X_i) \qquad (2)$$

где $P(Y, M_i|\mathbf{N}_i) = P(Y, M_i|X_i = 1)P(X_i = 1|\mathbf{N}_i) +$
$P(Y, M_i|X_i = 0)P(X_i = 0|\mathbf{N}_i)$. Соответствующая вероятность без сети $P(X_i|\mathbf{Y}, \mathbf{M}) = P(X_i)P(M_i|X_i)\prod_{k=1}^{m} P(Y_k|X_i)P(\mathbf{Y}, M_i)$. В соответствии с уравнением (2), для вычисления уравнения 1 нам необходимо знать следующие вероятности: $P(X_i|\mathbf{N}_i), P(X_i), P(M_i|X_i), P(Y_k|X_i)$ для всех k. За исключением $P(X_i|\mathbf{N}_i)$ все они легко вычисляются за один проход через данные. Форма $P(X_i|\mathbf{N}_i)$ зависит от механизма, с помощью которого влияют друг на друга. Далее я опишу вычисление данной вероятности, когда $\mathbf{X}$ – это набор пользователь совместной системы фильтрации.

Для простоты предположим, что $\mathbf{M}$ является булевым вектором (то есть рассматривается только один тип маркетинговых действий, например, предложение клиенту определенной скидки). Пусть c – стоимость маркетинга

одному клиенту (предполагается, что константа), r_0 – доход от продажи клиенту, если маркетинговые действия не выполняются, r_1 – доход от продажи клиенту, если маркетинговые действия выполняются. r_0 и r_1 будут одинаковыми, если маркетинговые действия не будут предполагать скидку. Пусть $f_i^1(\mathbf{M})$ – результат присвоение присвоения M_i 1, и оставляя остальную часть вектора $\mathbf{M}$ неизменной, аналогично с $f_i^0(\mathbf{M})$. Ожидаемый подъем прибыли (the expected lift in profit) от маркетинга клиенту в изоляции (то есть игнорируя влияние на него других клиентов) вычисляется [17]

$$ELP_i(\mathbf{X}^k, \mathbf{Y}, \mathbf{M}) = r_1 P\left(X_i = 1 \middle| \mathbf{X}^k, \mathbf{Y}, f_i^1(\mathbf{M})\right) - r_0 P\left(X_i = 1 \middle| \mathbf{X}^k, \mathbf{Y}, f_i^0(\mathbf{M})\right) - c \quad (3)$$

Пусть $\mathbf{M}_0$ – нулевой вектор. Глобальный подъем ожидаемой прибыли, который получается в результате определенного выбора, вычисляется следующим образом

$$ELP(\mathbf{X}^k, \mathbf{Y}, \mathbf{M}) = \sum_{i=1}^{n} r_i P(X_i = 1 \middle| \mathbf{X}^k, \mathbf{Y}, \mathbf{M}) - r_0 \sum_{i=1}^{n} P(X_i = 1 \middle| \mathbf{X}^k, \mathbf{Y}, \mathbf{M}_0) - |\mathbf{M}| c$$

где $r_i = r_1$, если $M_i = 1$ и $r_i = r_0$, если $M_i = 0$, а $|\mathbf{M}|$ – количество единиц в векторе $\mathbf{M}$.

Внутренняя ценность клиента задается уравнением (3). Общая стоимость клиента (внутренняя плюс сетевая) – это и есть EPL, полученный путем маркетинга для него: $ELP\left(\mathbf{X}^k, \mathbf{Y}, f_i^1(\mathbf{M})\right) - ELP\left(\mathbf{X}^k, \mathbf{Y}, f_i^0(\mathbf{M})\right)$. Сетевой показатель клиента является разницей между ее суммарными и внутренними ценностями. В целом, это значение зависит от того, какие клиенты собираются купить товар, а какие уже купили.

Извлечение социальных сетей из баз данных совместной фильтрации

Возможно, десять лет назад было бы трудно практиковать использование такой модели, как уравнение (1), из-за отсутствия данных для оценки влияния вероятностей $P(X_i|\mathbf{N}_i)$. В наше время люди влияют друг на друга онлайн. Любая форма онлайн-сообщества является потенциально богатым источником данных для разработки социальный сетей. Я сосредоточилась на потенциально очень эффективном источнике данных: коллаборативная система фильтрации, широко используемые коммерческими сайтами (например, amazom.com), чтобы рекомендовать продукцию пользователям.

В коллаборативной системе фильтрации пользователи оценивают набор элементов (например, фильмы, книги, публикации в новостных группах), и эти оценки затем используются, чтобы рекомендовать другие элементы, которые могут быть интересны пользователю [18]. Рейтинги могут быть неявными (например, купил ли пользователь продукт или нет) или явными (например, пользователь дает рейтинг от нуля до пяти звезд продукту). Основная идея заключается в том, чтобы предсказать рейтинг пользователя, который он поставит товару, как средневзвешенное значение рейтингов, данных аналогичными пользователями, а затем рекомендовать элементы с высокими прогнозируемыми рейтингами. Сходство пары пользователей *(i, j)* измеряется с использованием коэффициента корреляции Пирсона:

$$W_{ij} = \frac{\sum_k (R_{ik} - \bar{R}_i)(R_{jk} - \bar{R}_j)}{\sqrt{\sum_k (R_{ik} - \bar{R}_i)^2 \sum_k (R_{jk} - \bar{R}_j)^2}} \quad (4)$$

где R_{ik} − рейтинг пользователя i на продукт k, $\bar{R}_i$ − среднее значение рейтингов пользователя i (аналогично для пользователя j). Суммы и средние

вычисляются по элементу k, которые оценены для обоих пользователей i и j. Для элемента k, который пользователь i не оценил, спрогнозированный рейтинг выглядит следующим образом

$$\hat{R}_{ik} = \bar{R}_i + \rho \sum_{X_j \in \mathbf{N}_i} W_{ij}\left(R_{jk} - \bar{R}_j\right) \tag{5}$$

где $\rho = {1}\Big/{\sum_{X_j \in \mathbf{N}_i} |W_{ij}|}$ является нормализующим фактором, $\mathbf{N}_i -$ множество n_i пользователей наиболее похожих на пользователя i в соответствии с уравнением (4). Множество $\mathbf{N}_i$ может быть всей базой данной пользователей, но по соображениям надёжности шума и вычислительной эффективности n_i берется намного меньше. Для соседей, которые не оценили товар R_{jk} усыновлен в $\bar{R}_j$.

Ключевым преимущество коллаборативной системы фильтрации как источника для разработки социальной сети для вирусного маркетинга является то, что механизм, с помощью которого пользователи влияют друг н друга, известен и хорошо изучен – это и есть алгоритм коллаборативной фильтрации. Пользователь i влияет на пользователя j, когда j видит рекомендацию, часть рейтинга которой составляет оценка от i. Предполагая, что i и j не знают друг друга в реальной жизни, нет другого способа, которым они могут существенно влиять друг на друга.

Рассмотрим пространство $(\mathbf{X}, \mathbf{Y})$, где $\mathbf{Y} -$ набор свойств объекта (известно), а X_i означает оценил ли пользователь i товар или нет. Для простоты мы предполагаем, что если пользователь оценил товар, то он купил его и наоборот. $P(X_i)$ можно определить как долю элементов, оцененных пользователем i. Условные вероятности $P(Y_k|X_i)$ могут быть получены путем подсчета числа появления каждого значения Y_k с каждым значением X_i. Для оценки $P(M_i|X_i)$ требуется фаза сбора данных, в которой пользователи на рынок выбираются случайным образом, и их ответы регистрируется. Так как

для меня такие данные недоступны, я устанавливала $P(M_i|X_i)$ с использованием различных исследований об эффективности данного типа маркетинга.

Множество соседей $\mathbf{N}_i$ для каждого i является множеством соответствующего пользователя в системе коллаборативной фильтрации. Если рейтинги соседей известны, $\hat{R}_i$ является детерминированной функцией $\mathbf{N}_i$ (по уравнению (5)). Если оценки некоторых или всех соседей неизвестны, мы можем оценить их как их ожидаемые значения с учетом атрибутов товара. Другими словами, вклад соседа с неизвестным рейтингом будет $E\left[R_j|Y_k\right] - \bar{R}_j$. $P(R_j|\mathbf{Y})$ может быть оценена с использованием наивной байесовской модели. Пусть $\hat{R}_i(\mathbf{N}_i)$ будет значением $\hat{R}_i$ получается таким образом. Затем, рассматривая это как детерминированную величину,

$$P(X_i|\mathbf{N}_i) = \int_{R_{min}}^{R_{max}} P(X_i|\hat{R}_i, \mathbf{N}_i) dP\left(\hat{R}_i|\mathbf{N}_i\right) = P\left(X_i\middle|\hat{R}_i(\mathbf{N}_i), \mathbf{N}_i\right)$$
$$= P\left(X_i\middle|\hat{R}_i(\mathbf{N}_i)\right)$$

Все, что остается, оценить $P\left(X_i\middle|\hat{R}_i\right)$. Это можно рассматривать как одномерную регрессионную проблему с $\hat{R}_i$ в качестве входа и $P\left(X_i\middle|\hat{R}_i\right)$ в качестве выхода. Наиболее подходящая функциональная форма для этой регрессии будет зависеть от полученных данных. Я использовала кусочно-линейную модель $P\left(X_i\middle|\hat{R}_i\right)$, полученную делением диапазона $\hat{R}_i$ на бункеры, вычислением среднего $\hat{R}_i$ и $P\left(X_i\middle|\hat{R}_i\right)$ для каждого бункера, затем оценкой $P\left(X_i\middle|\hat{R}_i\right)$ для каждого произвольного $\hat{R}_i$ с помощью линейной интерполяции между двумя ближайшими средними.

Числовой эксперимент

Применив методологию, изложенную ранее, к проблеме маркетинга кинофильмов, я использовала базу данных онлайн-кинотеатра MovieLens [5]. Она содержит 100836 тысяч рейтингов 9742 фильмов, полученных 610 пользователями, собранных в период с 29 марта 1996 г. по 24 сентября 2008 г. Рынок кинопроката - интересное приложение для методов, которые мы предлагаем, потому что, как известно, успех фильма сильно зависит от эффекта «сарафанного радио» [19].

В базе данных содержаться данные трех типов: информация о пользователях, информация о фильмах (id, название, жанр и год выпуска) и рейтинги поставленные пользователями (от 0 до 5).

Значение переменных в области MovieLens следующее: X_i - видел ли человек i рассматриваемый фильм. **Y** содержит атрибуты(характеристики) фильма. R_i - рейтинг (от нуля до пяти звезд), присвоенный фильму человеком i. Для простоты, в этом разделе мы предполагаем, что $\hat{R}_i$ сконцентрирован в нуле (то есть $\bar{R}_i$ вычтено из $\hat{R}_i$; см. Уравнение (5))

1.1 Модель

В представленной моделе $\mathbf{Y} = \{Y_1, \ldots, Y_{20}\}$- двадцать логический переменных, описывающих жанр фильма. Таким образом, $P(\mathbf{Y}|X_i)$ является моделью жанровых предпочтений пользователя. Так как каждый пользователь из базы данных оценил более 20 фильмов, они все будут входить в сеть (в противном случае они не содержали бы никакой полезной информации). Веса W_{ij} были определены с использованием модифицированного коэффициента Пирсона. Далее мне предстояло рассчитать оптимальное число соседей. $n_i = 5$ число, которое, по моему мнению, обеспечивает разумный компромисс между точностью и скоростью модели.

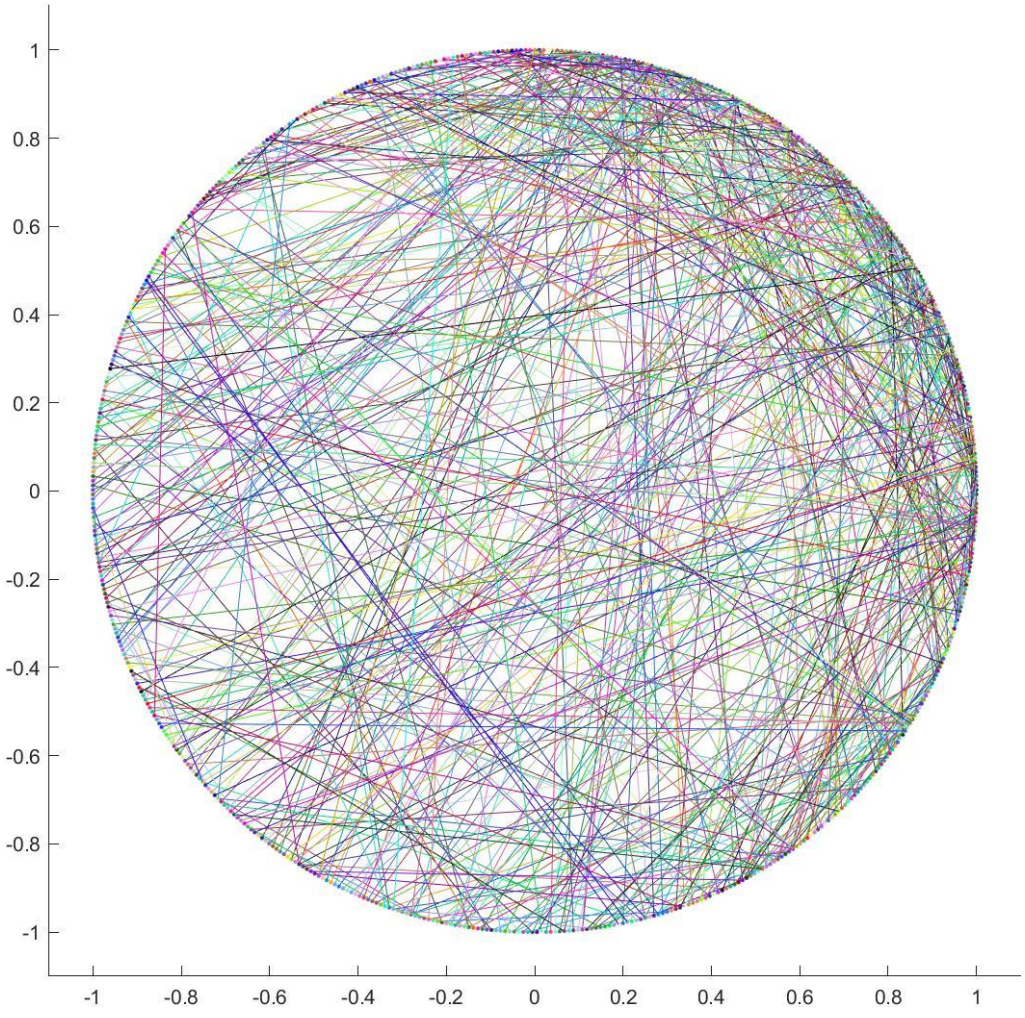

Рис.1 Сеть клиентов онлайн-кинотеатра

Среднее значение W_{ij} для соседей составило 0,0934. Повторение экспериментов со значениями $n_i = 10$ и $n_i = 20$ не привело к значительным изменениям модели. Стоить отменить, что сеть, полученная в каждом случае, была полностью сведена, то есть не содержала изолированных подграфов (см. Рис. 1).

Как я предлагала ранее, вычисления $P(X_i|\mathbf{X}^k, \mathbf{Y}, \mathbf{M})$ требует оценки $P\left(X_i\middle|\hat{R}_i\right), P(X_i)$ и $P(Y^k|X_i)$. $P(X_i)$ – это доля фильмов, оценённых i-ым пользователем. $P\left(X_i\middle|\hat{R}_i\right)$ моделировалась с использованием кусочно-линейной функции. Мы измеряем $P\left(X_i\middle|\hat{R}_i\right)$ для кажного из девяти отрезков, границы которых были -5,0; -2,0; -1,0; -0,5; -0,1; 0,1; 0,5; 1,0; 2,0 и 5,0. Стоит обратить внимание, что хоть R_i находится в диапазоне от 0 до 5, $\hat{R}_i$ представляет собой взвешенную сумму разности соседей от их среднего значения и, таким образом, может варьироваться от -5 до 5.

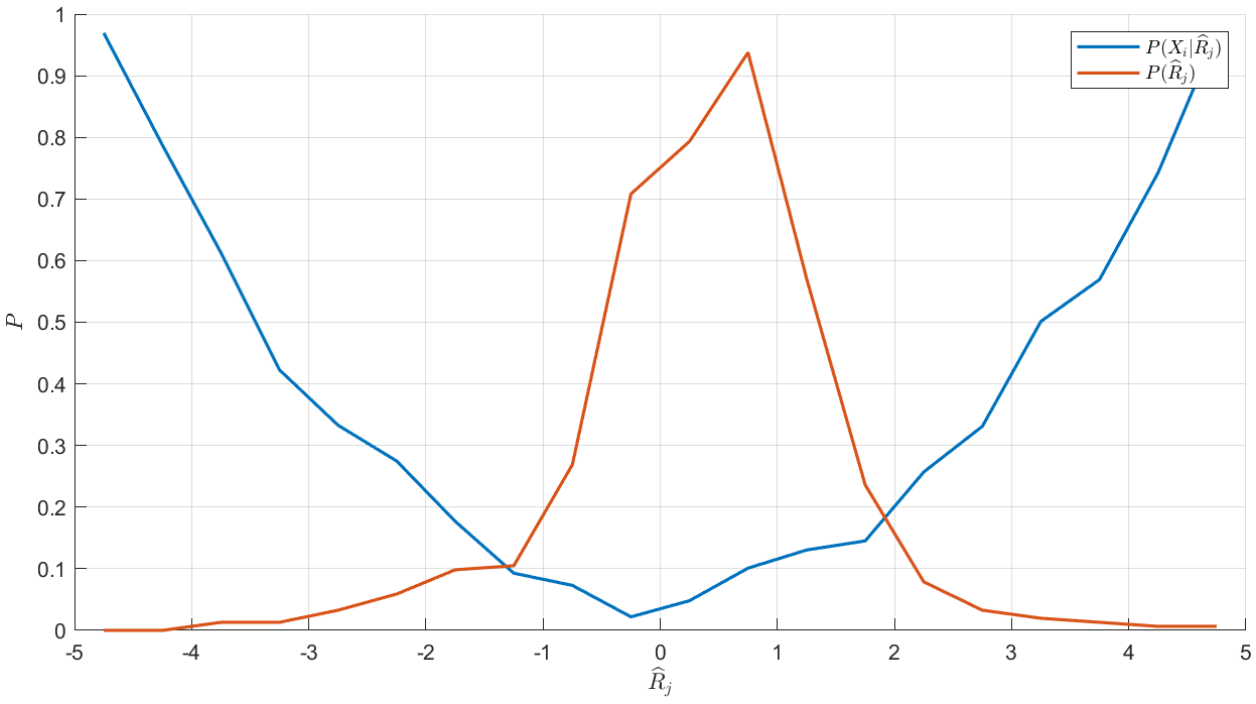

Рис. 2 Эмпирическое распределение $\hat{R}_i$ и X_i с учетом $\hat{R}_i$.

Изначально, я ожидала, что $P\left(X_i\middle|\hat{R}_i\right)$ будет монотонно увеличиваться с $\hat{R}_i$. Фактическая форма, показанная на рисунке 2, показывает увеличение

$P\left(X_i\middle|\hat{R}_i\right)$, когда $\hat{R}_i$ значительно отклоняется от 0 в любом направлении. Это явление обусловлено корреляцией между $\left|\hat{R}_i\right|$ и популярностью фильма: для популярного фильма $\hat{R}_i$ с большей вероятностью будет отклоняться дальше от 0, а X_i с большей вероятностью будет равно 1.

1.2 Сетевое значение

Возьмем постоянные переменные следующим образом: $r_0 = 1, r_1 = 0{,}5$ и $c = 0{,}1$. Проведя вычисления мы получили значения, изображенные на рисунке 3.

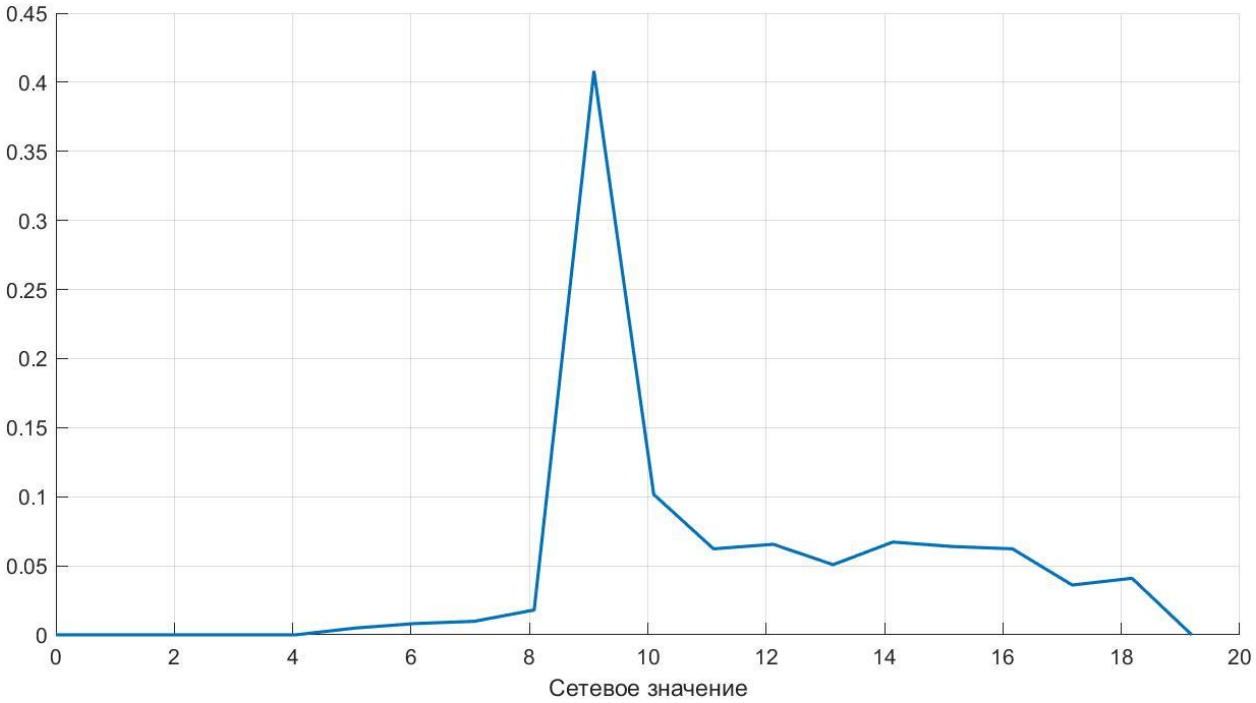

Рис. 3 Сетевые значения клиентов

Единица стоимости на этом графике – это средний доход, который был бы получен путем маркетинга для покупателя в отдельности. Таким образом сетевая ценность равная 10 подразумевает, что, продавая данному пользователю товар, мы, по сути, получаем бесплатный маркетинг для 10 клиентов. График показывает, что некоторые пользователи имеют очень высокую стоимость сети. Это идеальная ситуация для прямого вирусного маркетинга, так как онлайн-кинотеатр может предоставлять скидку лишь клиентам с высоким значением сети.

1.3 Маркетинг

Было проведено исследование трех маркетинговых стратегий: массовый маркетинг, отсутствие маркетинга и прямой маркетинг. Для массового маркетинга мы берем $M = M_0$, для массового $M_i = 1$ для любого i. Для прямого маркетинга $M_i = 1$, если $ELP_i(X^k, Y, M) > 8$.

В результате экспериментов массовый маркетинг дает отрицательную прибыль, в случае с массовой рекламой прибыль варьировалась от -132 до -234. Это значит, что игнорирование сетевых эффектов неизбежно приводит к потере прибыли. Опыт показывает, что выбор клиента с наибольшим сетевым значением помогает найти выгодные рыночные возможности, которые могли быть упущены. Без учета сетевых значений прямой маркетинг может привести к потере прибыли. Клиент, который выглядел прибыльным сам по себе, мог на самом деле иметь отрицательную общую стоимость. Это наглядно демонстрирует, что игнорирование сетевых эффектов может не только привести к упущенным маркетинговым возможностям, но и сделать невыгодную маркетинговую акцию выгодной.

Подведя итоги маркетинговых экспериментов, можем сказать, что алгоритм совместно фильтрации можно модифицировать, чтобы он назначал соседями лишь активных пользователей. Это может привести к значительному росту прибыли.

Выводы

Используя проделанные исследования, можно разработать некую стратегию для выбора клиентов на рынок. Перспективного и «хорошего» клиента отличают следующие качества:

1. Более вероятно, что он даст высокую оценку товару
2. Он имеет сильный вес при определении прогнозируемого рейтинга соседей
3. Имеет много соседей, на которых он легко влияет
4. Имеет высокую вероятность покупки товара
5. Имеет много соседей, обладающими свойствами 1-4.

Говоря простыми словами, все сводится к поиску человека, который

1. Получит удовольствие от этого фильма
2. Имеет много близких друзей
3. На чье мнение легко повлиять
4. Скорее всего посмотрит фильм, который поступит в продажу
5. Имеет друзей обладающих такими же свойствами

Заключение

В представленной работе была построена модель, вычисляющая сетевое значение клиента онлайн-кинотеатра. Это помогло понять, что массовый маркетинг во времена социальных сетей и интернета избыточен, более того может привести к убыткам.

При построении стратегии маркетинга специалистам не стоит игнорировать сетевые эффекты. С их учетом прямой маркетинг дает положительное значение прибыли для компании.

Список литературы

1. A. M. Hughes The Complete Database Marketer: Second Generation Strategies and Techniques for Tapping the Power of Your Customer Database. Irwin, Chicago, IL: 1996.
2. C. X. Ling, C. Li Data Mining for Direct Marketing: Problems and Solutions // Proceeding of the Forth International Conference on Knowledge Discovery and Data Mining. New York, NY: AAAI Press, 1998. C. 73-79.
3. G. Piatetsky-Shapiro, B. Masand Estimating campaign benefits and modeling lift // Proceeding of Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Can Diego, CA: ACM Press, 1999. C. 185-193.
4. S. Jurvetson What exactly is viral marketing? // Red Herring. 2000. №78. C. 110-112.
5. MovieLens Database // GroupLens URL: https://grouplens.org/datasets/movielens/ (дата обращения: 10.03.20).
6. D. R. Mani, J Drew, A Betz, P Datta Statistic and data mining techniques for lifetime value modeling // Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining . New York, NY: ACM Press, 1999. C. 94-103.
7. Д. А. Губанов, Д. А. Новиков, А. Г. Чхартишвили Социальные сети: модели информационного влияния, управления и противоборства. Москва: Физматлит, 2010.
8. J. S. Breese, D. Heckerman, C. Kadie Empirical analysis of predictive algorithms for collaborative filtering // Proceedings of Fourteenth

Conference on Uncertainty in Artificial Intelligence. Madison, WI: Morgan Kaufmann, 1998.

9. P. Domingos, M. Richardson Mining the Network Value of Customers // Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining. San-Francisco, CA: ACM Press, 2001. С. 57-66.
10. L. Pelkowitz A continuous relaxation labeling algorithm for Markov random fields // IEEE Transactions on Systems, Man and Cybernetics. 1990. №20. С. 709-715.
11. J. Besag Spatial interaction and the statistical analysis of lattice systems // Journal of the Royal Statistical Society. 1947. №36. С. 192-236.
12. R. Kindermann, J.L. Snell Markov Random Fields and Their Applications. RI: Providence, 1980.
13. R. Chellappa, A. K. Jain Markov Random Fields: Theory and Applications. Boston, MA: Academic Press, 1993.
14. S. Geman, D. Geman Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1984. №6. С. 721-741.
15. S. Chakrabarti, B. Dom, P. Indyk Enhanced hypertext categorization using hyperlinks // Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data. Seattle, WA: ACM Press, 1998. С. 307-318.
16. P. Domingos, M. Pazzani On the optimality of the simple Bayesian classifier under zero-one loss // Machine Learning. 1997. №29. С. 103-130.

17. D. Heckerman, D. M. Chickering A dicision theorwtic approach to targeted advertisisng // Proceedings of Sixteenth Annual Conference on Uncertainty in Artificial Intelligence. Stanford, CA: Morgan Kaufmann, 2000.

18. P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl GroupLens: An open architecture for collaborative filtering of netnews // Proceedings of the ACM 1994 Conference on Computer Supported Cooperative Work. New York, NY: ACM Press, 1994. C. 175-186.

Приложение

Ниже представлен код в MATLAB для вычисление матрицы корреляционных значений,

```
ratings = ratings_table{:, 3};
W = zeros(totalUsers, totalUsers);
for i = 1:totalUsers - 1
    for j = i + 1:totalUsers
        list1 = userInit(i):userInit(i) + count(i) - 1;
        list2 = userInit(j):userInit(j) + count(j) - 1;
        [list, len, indx1, indx2] = compare_movies(movieId_r(list1), movieId_r(list2));
        if len ~= 0
            ER1 = mean(ratings(list1(indx1(1:len))));
            ER2 = mean(ratings(list2(indx2(1:len))));
            sum1 = 0; sum2 = 0; sum3 = 0;
            for k = 1:len
                sum1 = sum1 + (ratings(list1(indx1(k))) - ER1) * (ratings(list2(indx2(k))) - ER2);
                sum2 = sum2 + (ratings(list1(indx1(k))) - ER1)^2;
                sum3 = sum3 + (ratings(list2(indx2(k))) - ER2)^2;
            end
            if sum2 ~= 0 && sum3 ~= 0
                W(i, j) = sum1 / sqrt(sum2 * sum3);
            end
        end
    end
end
W = W + W';
for i = 1:totalUsers
    W(i, i) = 1;
end
```