

Абраменко Максим Владимирович

бакалавр направления подготовки
«Прикладная математика и информатика»
Национальный исследовательский ядерный
университет «МИФИ»
Россия, Москва

Паршин Михаил Николаевич

бакалавр направления подготовки
«Информатика и вычислительная техника»
Национальный исследовательский ядерный
университет «МИФИ»
Россия, Москва
zojloverr@gmail.com

Maxim V. Abramenko

Bachelor of studies "Applied Mathematics
and Informatics"
National Research Nuclear University MEPhI
Russia, Moscow

Mikhail N. Parshin

Bachelor of the field of study "Informatics
and Computer Engineering"
National Research Nuclear University MEPhI
Russia, Moscow

**ПРЕДСКАЗАНИЕ БУДУЩИХ ПОКУПОК
КЛИЕНТА МАГАЗИНА НА ОСНОВЕ
ИСТОРИИ АКТИВНОСТИ**

**PREDICTION OF FUTURE PURCHASES OF
A STORE CUSTOMER BASED ON
THE HISTORY OF ACTIVITY**

Аннотация

Таргетированная реклама является недорогим и эффективным способом увеличения прибыли магазинов. Основным условием успеха рекламной компании является правильное определение интересов клиента. Целью данной работы было создание инструмента для определения товаров, которые клиент купит с наибольшей вероятностью.

Ключевые слова:

регрессия, RFM-анализ, финансовые данные

Abstract

Targeted advertising is an inexpensive and effective way to increase store profits. The main condition for the success of an advertising campaign is the correct definition of the client's interests. The purpose of this work was to create a tool for determining the goods that a client will buy with the greatest probability.

Keywords:

regression, RFM-analysis, financial data

Введение

Клиент-ориентированный подход хорошо зарекомендовал себя в сферах торговли и услуг, поскольку позволяет увеличить прибыль с каждого клиента. Работа с каждым клиентом «вручную» является крайне трудоемким и долгим процессом. Целесообразнее использовать для решения некоторых классов задач, таких как, например, разбиение клиентов на категории, методы математической статистики и алгоритмы машинного обучения. Исходя из данных о транзакциях банки и онлайн-магазины могут, например, выдавать клиенту таргетированную рекламу, либо же персональные скидки на основе его предыдущей покупательской деятельности.

Методология

В данной работе будет решаться задача разработки алгоритма, который на основе синтетических данных из открытых источников будет предсказывать

некоторые числовые и категориальные признаки клиентов на основе его предыдущего поведения в системе.

Алгоритм структурно разделен на четыре блока:

1. Блок предобработки данных.
2. Регрессионная модель, которая на основании признаков, имеющихся данных предсказывает сумму покупок за рассматриваемый период.
3. Блок группировки, разделяющий пользователей на категории с помощью новых выделенных признаков.
4. Модель классификации, предсказывающая категорию товара, приобретаемого клиентом на основании некоторого числа признаков.

Описание данных

В работе использовались данные о транзакциях с 01/12/12 по 09/12/2011 зарегистрированного британского онлайн магазина, который специализируется на уникальных подарках [1]. Многие клиенты магазина являются оптовиками.

В датасете содержится информация о 541909 транзакциях. О каждой транзакции известно:

1. Уникальный номер клиента.
2. Дата и время поступления.
3. Название товара.
4. Уникальный код товара.
5. Количество товара за транзакцию.
6. Страна, в которой проживает покупатель.

RFM-анализ как метод группировки клиентов

RFM-анализ – анализ клиентской базы на основе истории покупок, который проводится по трём показателям [2]:

- 1) Recency – давность покупки – промежуток времени с момента последней покупки.
- 2) Frequency – частота покупок – количество покупок за данный период.
- 3) Monetary – сумма покупок – сумма всех покупок за данный период.

В результате RFM-анализа обнаруживаются группы самых лояльных, прибыльных и неактивных клиентов.

Результаты

Платежные истории клиентов были разбиты на периоды длиной в месяц и решать задачу предсказания суммы общей суммы трат клиента за следующий период. По признакам транзакций текущего и предыдущего периода была построена регрессионная модель [3]. По рисунку 1 можно сделать вывод, что оптимальным решением будет длина периода равная 50 дням. На графике так же можно увидеть некоторую сезонность – раз в 15-20 дней траты возрастают. Эти наблюдения связываются с периодом выплаты зарплат, и как следствие, большими тратами со стороны клиентов. Точность угадывания на сгруппированных при помощи RFM-анализа данных составила 68.3%.

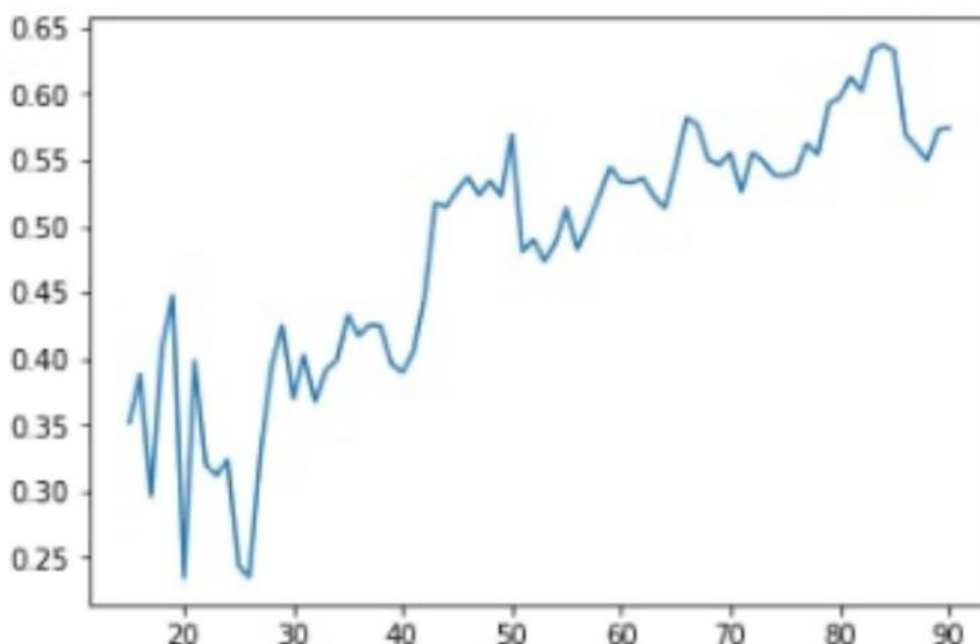


Рисунок 1 – График зависимости длины периода в днях от точности предсказания регрессионной модели

Реализация алгоритма разбиения товаров на заранее не заданные категории. Всего в данных о покупках клиентов содержится 400104 различных купленных товаров, из которых 3896 являются уникальными. Все множество наименований товаров на английском языке было разбито на отдельные слова. Был создан словарь `stop_words`, содержащий наиболее работе алгоритма слова, такие как, например, «and», «six» и «everyone». Выбор оптимального числа кластеров строился по правилу локтя [4]: при нем строится график зависимости инерции от количества кластеров и

берется значение на изгибе полученной кривой. Инерция – метрика, показывающая насколько кластеры связаны между собой, вычисляется по формуле (1).

$$\sum_{i=1}^n \min(|x_i - x_j|)^2 \quad (1)$$

где n – число кластеров, m – среднее значение в каждом из кластеров, или же центроид.

На рисунке 2 показан график зависимости величины инерции от числа кластеров при разбиении.

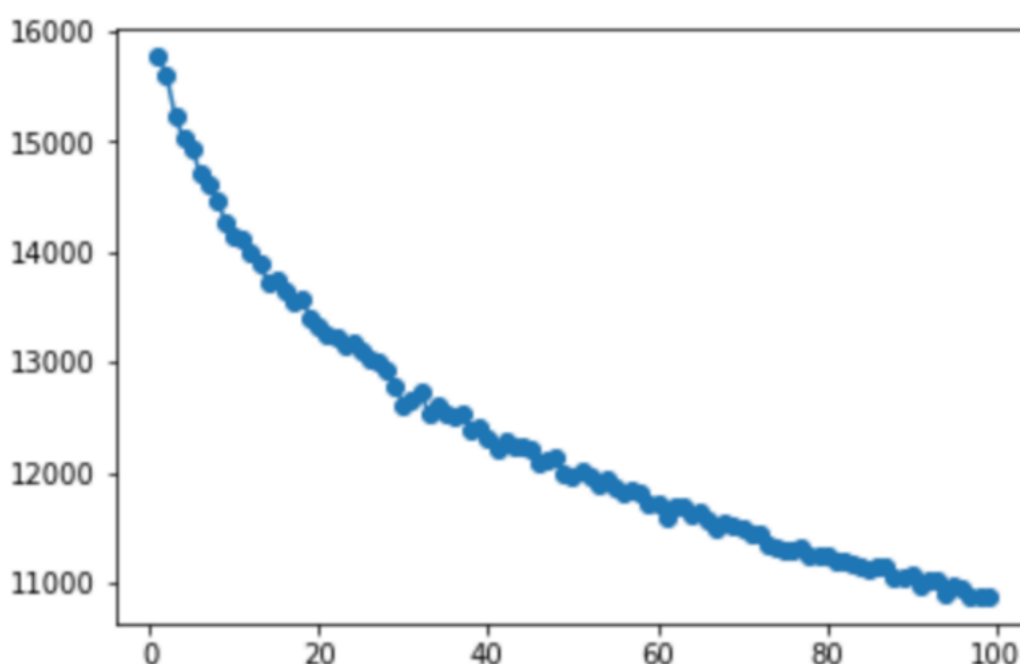


Рисунок 2 – График зависимости величины инерции от количества кластеров

При помощи алгоритма k-средних все наименования товаров были разбиты на 40 непересекающихся категорий. 25 из 40 категорий получились осмысленными, в них вошли, например, раздел новогодних товаров, кружки и спортивные товары.

Заключение

Был разработан алгоритм, который на основании данных о прошлых покупках клиента предсказывает стоимость следующей покупки и категорию, к которой может принадлежать купленный товар. Исходя из этих данных могут быть отобраны товары, подходящие для рекомендации клиенту.

Список использованных источников

1. Online Retail Data Set from UCI ML repo [Выборка данных] – Режим доступа: <https://www.kaggle.com/jihyeseo/online-retail-data-set-from-uci-ml-repo>.
2. Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study / K. Mahboubeh [et al.] // Procedia Computer Science. – 2011. – № 3. – P. 53-67.
3. Вариации регрессии [Электронный ресурс] // Университет ИТМО. – 2020. – Режим доступа: https://neerc.ifmo.ru/wiki/index.php?title=%C2%E0%F0%E8%E0%F6%E8%E8_%F0%E5%E3%F0%E5%F1%F1%E8%E8.
4. Islam, M. Investigating the Performance of Naive- Bayes Classifiers and K- Nearest Neighbor Classifiers / M. Islam, M. Ahmadi, M. Sid-Ahmed // 2007 International Conference on Convergence Information Technology (ICCIT 2007). – 2007. – ISSN 0- 7695-3038-9