Анализ маркетинговой базы данных на основе методов машинного обучения

Э. Ф. Миниахметов, email: emilpriemil72@gmail.com
В. В. Мокшин, email: vladimir_kgtu@mail.ru

Казанский национальный исследовательский технический университет имени А. Н. Туполева

Аннотация. *Проанализирована маркетинговая база данных на предмет эффективности рекламных кампаний.*

Ключевые слова: *маркетинг, исследовательский анализ данных, статистический анализ, бизнес-данные.*

Введение

Широкое распространение технологий автоматизированной обработки информации и накопление в компьютерных системах больших объемов данных, сделали очень актуальной задачу поиска неявных взаимосвязей, имеющихся в наборах данных. Для её решения используются методы математической статистики, теории баз данных и ряда других областей [1].

Используя программное обеспечение для поиска закономерностей в больших пакетах данных, предприятия могут выстраивать маркетинговые стратегии, управлять кредитными рисками, обнаруживать мошенничество, фильтровать спам или даже выявлять настроения пользователей. В данной работе проанализирована маркетинговая база данных на основе методов машинного обучения.

1. База данных

В данной работе будет использоваться маркетинговая база данных, предоставленная студентам для финального проекта в рамках программы магистра в области бизнес-аналитики.

Данная база данных состоит из 2240 наблюдений (клиентов) с 28 переменными, связанными маркетинговыми данными, такими как: характеристики клиентов, приобретённые продукты, успехи и неудачи проведённых кампаний [2].

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 28 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   2240 non-null   int64
 1   Year_Birth           2240 non-null   int64
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4    Income              2216 non-null   object
 5   Kidhome              2240 non-null   int64
 6   Teenhome             2240 non-null   int64
 7   Dt_Customer          2240 non-null   object
 8   Recency              2240 non-null   int64
 9   MntWines             2240 non-null   int64
 10  MntFruits            2240 non-null   int64
 11  MntMeatProducts      2240 non-null   int64
 12  MntFishProducts      2240 non-null   int64
 13  MntSweetProducts     2240 non-null   int64
 14  MntGoldProds         2240 non-null   int64
 15  NumDealsPurchases    2240 non-null   int64
 16  NumWebPurchases      2240 non-null   int64
 17  NumCatalogPurchases  2240 non-null   int64
 18  NumStorePurchases    2240 non-null   int64
 19  NumWebVisitsMonth    2240 non-null   int64
 20  AcceptedCmp3         2240 non-null   int64
 21  AcceptedCmp4         2240 non-null   int64
 22  AcceptedCmp5         2240 non-null   int64
 23  AcceptedCmp1         2240 non-null   int64
 24  AcceptedCmp2         2240 non-null   int64
 25  Response             2240 non-null   int64
 26  Complain             2240 non-null   int64
 27  Country              2240 non-null   object
```

Рис. 1. База данных Marketing Analytics

2. Исследовательский анализ данных

Приступая к исследовательскому анализу данных, необходимо ответить на следующие вопросы:

Существуют ли какие-либо полезные переменные, которые можно спроектировать с учетом данных? Существуют ли закономерности или аномалии в данных? Можно ли составить их график?

На основе имеющихся данных можно создать и использовать более практичные:

- Общее количество неработающих ("Dependents") может быть рассчитано из суммы "Kidhome" и "Teenhome";
- Год становления клиентом ("Year_Customer"), может быть спроектирован с помощью "Dt_Customer";

– Общая потраченная сумма ("TotalMnt") может быть рассчитана на основе суммы всех функций, содержащих ключевое слово "Mnt";
– Общее количество покупок ("TotalPurchases") может быть рассчитано на основе суммы всех функций, содержащих ключевое слово "Purchaches";
– Общее количество рекламных кампаний, в которых принял участие клиент ("TotalCampaignsAcc") может быть рассчитано на основе суммы всех функций, содержащих ключевые слова "Cmp" и "Response".

	ID	Dependents	Year_Customer	TotalMnt	TotalPurchases	TotalCampaignsAcc
0	1826	0	2014	1190	15	1
1	1	0	2014	577	18	2
2	10476	1	2014	251	11	0
3	1386	2	2014	11	4	0
4	5371	1	2014	91	8	2

Рис. 2. Новые используемые переменные

Чтобы выявить закономерности, необходимо определить корреляции признаков. Положительные корреляции между объектами отображаются красным цветом, отрицательные корреляции отображаются синим цветом, а отсутствие корреляции отображается серым цветом на карте корреляционного анализа ниже. Также необходимо разделить эти закономерности на кластеры.

Кластер "Высокий доход":

– Потраченная сумма ("Totalmnt" и другие функции "Mnt") и количество покупок ("TotalPurchases" и другие функции "Num...Purchases") положительно коррелируют с "Income".

– Покупки в магазине, в Интернете или через каталог ("NumStorePurchases", "NumWebPurchases", "NumCatalogPurchases") положительно коррелируют с "Income".

Кластер "Наличие детей и подростков":

– Потраченная сумма ("TotalMnt" и другие функции "Mnt") и количество покупок ("TotalPurchases" и другие функции "Num...Purchases") отрицательно коррелируют с "Dependents".

– Покупки по акции ("NumDealsPurchases") положительно коррелируют с "Dependents" (детьми и/или подростками) и отрицательно коррелируют с "Income".

Кластер "Рекламные кампании":

– Принятие рекламных кампаний ("AcceptedCmp" и "Response") сильно положительно коррелирует друг с другом.

– Слабая положительная корреляция рекламных кампаний наблюдается с кластером "Высокий доход", а слабая отрицательная корреляция наблюдается с кластером "Наличие детей и подростков".

Количество посещений веб-сайта за последний месяц ("NumWebVisitsMonth") не коррелирует с увеличением числа покупок в Интернете ("NumWebPurchases"). Вместо этого "NumWebVisitsMonth" положительно коррелирует с количеством покупок по акции ("NumDealsPurchases"), что позволяет предположить, что предложения являются эффективным способом стимулирования покупок на веб-сайте.

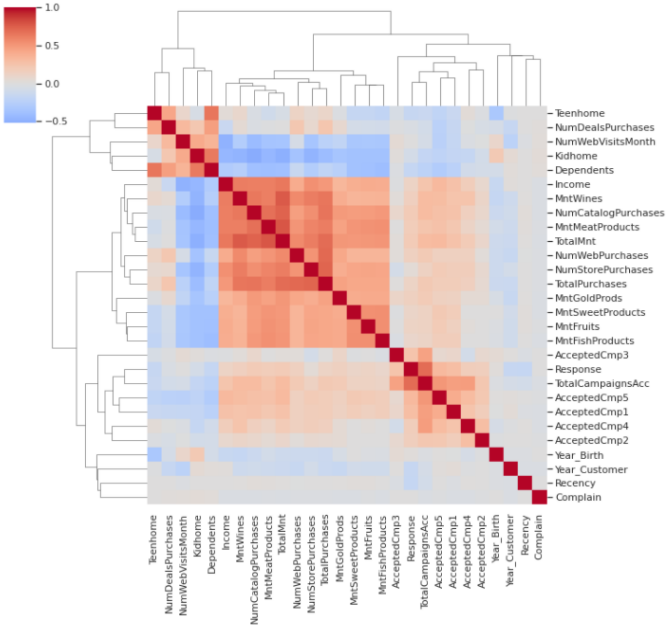

Рис. 3. Корреляционный анализ базы данных

Важно также проиллюстрировать корреляцию между количеством принятых рекламных кампаний ('TotalCampaignsAcc') и доходом клиентов ('Income'), также между ('TotalCampaignsAcc') и наличием детей/подростков в семье ('Dependents'). По графикам можно сказать,

что принятие рекламной кампании положительно коррелирует с доходом и отрицательно коррелирует с наличием детей/подростков.

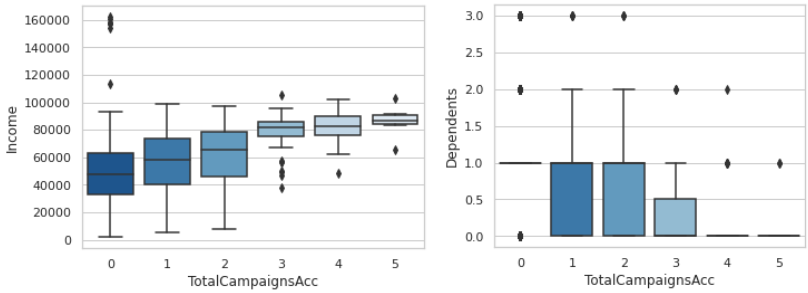

Рис. 4. Зависимость между ('TotalCampaignsAcc') и ('Income'), ('Dependents')

3. Статистический анализ

Для начала необходимо определить, какие факторы связаны с продажами товара в магазине.

С помощью SHAP (SHapley Additive exPlanations) строится график линейной зависимости числа продаж в магазинах с общим числом продаж, числом продаж из каталога, сайта, акций и.т.д.

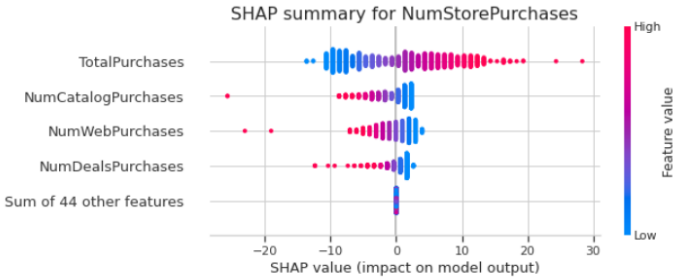

Рис. 5. Зависимость (' NumStorePurchases ')

По данному графику можно сказать, что:

– Количество покупок в магазине увеличивается с увеличением общего количества покупок ("TotalPurchases").

– Количество покупок в магазине уменьшается с увеличением количества покупок по каталогу, через Интернет или по акционным

продажам ('NumCatalogPurchases', 'NumWebPurchases', 'NumDealsPurchases').

Можно сделать вывод, что клиенты, покупающие товар в магазинах, зачастую не покупают их через каталог, сайт или по акции.

Для общей статистики необходимо определить общее количество продаж в разных странах.

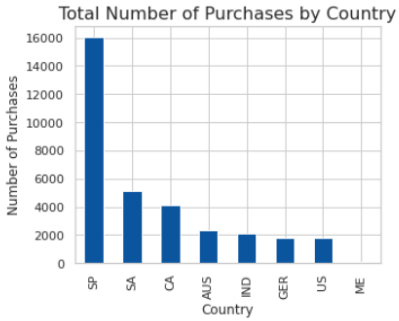

Рис. 6. Зависимость (' NumStorePurchases ')

Для анализа эффективности рекламных кампаний необходимо определить процент принятия рекламных кампаний в разных странах.

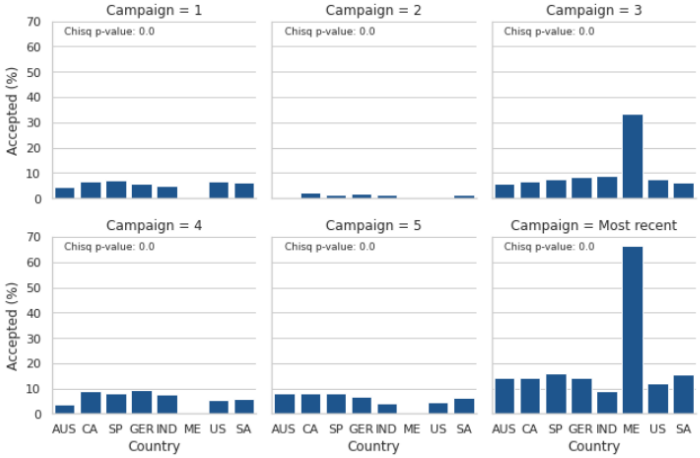

Рис. 7. Принятие рекламных кампаний в разных странах

По полученным данным можно сказать, что показатели принятия в целом низкие. Кампанией с наибольшим показателем принятия является последняя, а страной с самым высоким уровнем принятия является Мексика.

4. Визуализация данных

После анализа данных переходим к их визуализации. Определим среднестатистического клиента:

- родился в 1969 году;
- стал клиентом в 2013 году;
- имеет доход примерно в 52 000 долларов в год;
- имеет 1 неработающего в семье;
- совершил последнюю покупку 49 дней назад.

Среднестатистический клиент тратит: 25-50 долларов на фрукты, сладости, рыбу или золотые украшения; более 160 долларов на мясные продукты; более 300 долларов на вино – всего более 600 долларов. Самые популярные продукты: вино, мясная продукция.

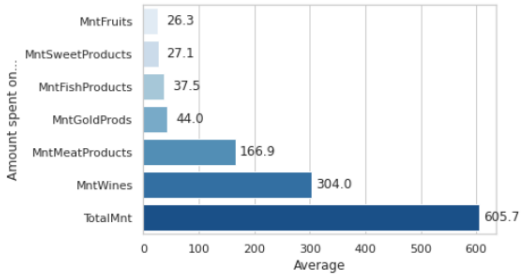

Рис. 8. Среднестатистические траты покупателя

Также, среднестатистический клиент:

- участвовал менее, чем в одной рекламной кампании;
- совершил 2 покупки по сделкам, 2 покупки по каталогу, 4 покупки в Интернете и 5 покупок в магазине;
- общее количество покупок составило 14;
- заходил на сайт 5 раз.

Неэффективные каналы: каталог, акции, рекламные кампании.

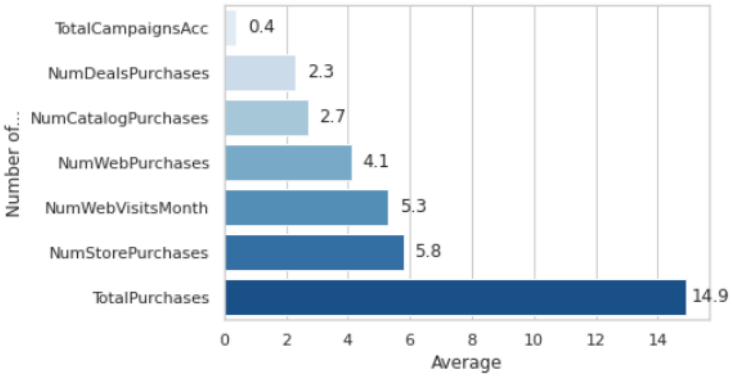

Рис. 9. Среднестатистические активности покупателя

По проанализированным данным можно сделать следующие выводы:

– Самой успешной рекламной кампанией была самая последняя кампания (название колонки: Ответ), и она была особенно успешной в Мексике (коэффициент принятия >60%!). Рекомендуется проводить будущие рекламные кампании с использованием той же модели, которая недавно была внедрена в Мексике.

– Принятие рекламной кампании положительно коррелирует с доходом и отрицательно коррелирует с наличием детей/подростков. Рекомендуется создать два потока целевых рекламных кампаний, один из которых нацелен на людей с высоким уровнем дохода без детей/подростков в семье, а другой - на людей с низким уровнем дохода с детьми/подростками.

– Наиболее успешными продуктами являются вина и мясо (т.е. в среднем покупатель тратит больше всего денег на эти товары). Рекомендуется сосредоточить рекламные кампании на увеличении продаж менее популярных товаров.

– Неэффективными каналами являются сделки и покупки по каталогу (т.е. средний клиент совершил наименьшее количество покупок по этим каналам).

– Наиболее эффективными каналами являются покупки в Интернете и магазинах (т.е. средний клиент совершил наибольшее количество покупок по этим каналам). Рекомендуется сосредоточить рекламные кампании на более успешных каналах, чтобы охватить больше клиентов.

Заключение

В данной статье был проведён анализ маркетинговой базы данных на основе машинного обучения. Были осуществлены исследовательский и статистический анализы, а также визуализация данных. В итоге проделанной работы мы выявили недостатки рекламных кампаний и предложили вариант повышения её эффективности.

Литература

1. Нестеров, С. А. Базы данных. Интеллектуальный анализ данных : учеб. пособие / С. А. Нестеров – СПб. : Изд-во Политехн. ун-та, 2011. – 272 с.
2. Маркетинговая аналитика [Электронный ресурс]. – Режим доступа : https://www.kaggle.com/jackdaoud/marketing-data
3. Мокшин, В. В. Метод формирования модели анализа сложной системы / В.В. Мокшин, И.М. Якимов // Информационные технологии. – 2011. – № 5. – С. 46-51.
4. Мокшин, В. В. Рекурсивный алгоритм построения регрессионных моделей сложных вероятностных объектов / В. В. Мокшин, И. Р. Сайфудинов, А. П. Кирпичников // Вестник Технологического университета. – 2017. – Т. 20. – №9. – С. 112-116.