Общий вывод

Целю исследования являлась - разработка рекомендаций для интернет-магазина "В один клик", которые позволят персонализировать предложения постоянным клиентам, чтобы увеличить их покупательскую активность

I. Обзор предоставленных для исследования данных

A. Для исследования предоставлено 4 датасета

1) market\_file содержит данные о поведении покупателя на сайте, о коммуникациях с покупателем и его продуктовом поведении

1300 записи, 13 колонок, дубикатов записей нет, пропусков данных нет

id — номер покупателя в корпоративной базе данных,

Покупательская активность — рассчитанный класс покупательской активности

Тип сервиса — уровень сервиса, например «премиум» и «стандарт»

Разрешить сообщать — информация о том, можно ли присылать покупателю дополнительные предложения о товаре

Маркет\_актив\_6\_мес — среднемесячное значение маркетинговых коммуникаций компании, которое приходилось на покупателя за последние 6 месяцев

Маркет\_актив\_тек\_мес — количество маркетинговых коммуникаций в текущем месяце

Длительность — значение, которое показывает, сколько дней прошло с момента регистрации покупателя на сайте

Акционные\_покупки — среднемесячная доля покупок по акции от общего числа покупок за последние 6 месяцев

Популярная\_категория — самая популярная категория товаров у покупателя за последние 6 месяцев

Средний\_просмотр\_категорий\_за\_визит — показывает, сколько в среднем категорий покупатель просмотрел за визит в течение последнего месяца

Неоплаченные\_продукты\_штук\_квартал — общее число неоплаченных товаров в корзине за последние 3 месяца

Ошибка\_сервиса — число сбоев, которые коснулись покупателя во время посещения сайта

Страниц\_за\_визит — среднее количество страниц, которые просмотрел покупатель за один визит на сайт за последние 3 месяца

2) market\_money содеражит данные о выручке, которую получает магазин с покупателя

3900 записи, 3 колонки, дубикатов записей нет, пропусков данных нет

id — номер покупателя в корпоративной базе данных,

Период - название периода, во время которого зафиксирова выручка

Выручка - сумма выручки, полученная магазином за период

3) market\_time содеражит данные о времени (в минутах), которое покупатель провёл на сайте в течение периода

2600 записи, 3 колонки, дубикатов записей нет, пропусков данных нет

id — номер покупателя в корпоративной базе данных,

Период — название периода, во время которого зафиксировано общее время

минут — значение времени, проведённого на сайте, в минутах

4) money содержит данные о среднемесячной прибыли покупателя за последние 3 месяца

1300 записи, 2 колонки, дубикатов записей нет, пропусков данных нет

id — номер покупателя в корпоративной базе данных,

Прибыль — значение прибыли

B. В рамках предподготовки данных

- изменен тип данных в колонках market\_file('Маркет\_актив\_6\_мес', 'Акционные\_покупки') и market\_money('Выручка')

- стандартизированы названия колонок

- удалены опечатки в значениях текстовых колонок

II. Анализ предоставленных для исследования данных

✔ Количественные показатели :

- market\_activ\_6mnth(среднемесячная маркетинговая активность за последние 6 мес.), как и market\_activ\_current (текущая маркетинговая активность): у 75% покупателей находится в диапазоне от 3 до 5 раз, медиана - 4 контакта.

- duration (длительность с момента регистрации на сайте): распределение значений показателя нормальное, у 75% покупателей срок регистрации находится в диапазоне от 284 до 912 дней, среднее и медианное количество несильно отличаются и составляют 601 и 606 дней соответственно

- promo\_buy(среднемесячная доля покупок по акции от общего числа покупок за последние 6 мес): на гистограмме видны две четкие области концентрации значений - в диапазоне от 0.1 до 0.4, где сконцентрировано наибольшее количество значений, и в диапазоне от 0.8 до 0.9, средний показатель составляет 0.32.

- avg\_cat\_views (среднее количество просмотренных категорий товаров): распределение значений показателя нормальное, у 75% покупателей количество просмотренных категорий колблется от 2 до 5, среднее и медианное количество составляет 3 категории товара

- not\_paid\_item\_per\_q (количество неоплаченных товаров в корзине за квартал): ассиметричное распределение значений показателя - положительная ассиметрия (большинство значений показателя расположена справа от моды), у 75% покупателей количество неоплаченных товаров находится в диапазоне от 1 до 5, среднее и медианное количество - 3 товара

- servis\_error (ошибки сервиса): распределение значений показателя нормальное, у 75% покупателей число сбоев находится в диапазоне от 2 до 7 раз, среднее и медианное значения равны и составляют 4 сбоя

- pages\_per\_visit (число страниц, просмотренных за визит за последние 3 месяца): ассиметричное распределение значений показателя - положительная ассиметрия, у 75% покупателей число страниц находится в диапазоне от 4 до 13 страниц, среднее и медианное значения равны и составляют 8 страниц

- revenue(выручка от покупателя за период): распределение значений показателя нормально, у 75% покупателей ежемесячные расходы находятся в диапазоне от 4,353.95 до 5,755.12, значения медианы и среднего близки и составляют 4,957.5 и 5,025.7 соответственно

- minutes(минуты, проведенные покупателем за месяц): распределение значений показателя нормальное, 75% покупателей проводят на сайте от 9 до 19 минут в месяц, значения медианы и среднего близки и составляют 13.00 и 13.34 соответственно

- profit (прибыль от покупателя) нормально распределен, 75% покупателей в среднем за месяц приносят прибыль в диапазоне от 2.83 до 5.13. Максимальная среднемесячная прибыль составляет 7.43, среднее и медианное значения среднемесячной прибыли практически равны - 4.00 и 4.04 соответственно.

✔ Категорийные показатели

- покупательская активность (целевой признак) - две категории: 'прежний уровень' и 'снизилась', преобладает категория 'прежний уровень' - 61.7%

- тип сервиса(уровень сервиса) - две категории: 'премиум' и 'стандарт', преобладает категория 'стандарт' - 71.1%)

- разрешене на отправку сообщений - две категории: 'да' и 'нет', преобладает категория 'да' - 74.0%

- популярная категория - шесть категорий, преобладает категория 'товары для детей'(25.4%), наименее популярная - категория 'кухонная посуда' (10.6%)

- периоды - три категории: 'препредыдущий месяц', 'предыдущий месяц' и 'текущий месяц', расходы покупателей по месяцам распределены практичечски равномерно.

- периоды - две категории: 'предыдущий месяц' и 'текущий месяц', количество минут, проведенных покупателями на сайте как в предыдущем, так и в текущем практически равно.

Нулевые значения не являются аномалией и означают, что покупатель не делал покупок водном из анализируемых периодов,

а вот расходы на 106,862.2 явно похожи на аномалию. Посмотрим на количество таких транзакций (она одна)

Одна транзакцияя покупателя, обычно делающего покупки на 5-6 тыс.

Повидимому, в данные зазкралась ошибка, которую надо исправить, удалим из данных аномальные покупки свыше 10,000

✔ Количественные и категорийные показатели в разрезе покупательской активности

a) у покупателей с прежним уровнем покупательской активности выше уровень:

- маркетинговых коммуникаций за последние 6 мес,

- среднее количество просмотренных категорий товаров,

- количество просмотренных страниц за визит,

- выручка за текущий месяц (при этом за прошлый и предпредыдущий - одинакова для обеих групп),

- время, проведенное на сайте как в текущем, так и в прошлом месяце,

- количество сбоев сервиса также превалирует у этой группы (при этом медиана - 4 сбоя - на одном уровне у обеих групп)

b) у покупателей со сниженной покупательской активностью выше уровень

- доли акционных покупок

- количества неоплаченных товаров в корзине

c) примерно в одинаковых диапазонах находится длительность с момента регистрации на сайте у покупателей обеих групп распределена в примерно в одинаковых диапазонах, с небольшим преимуществом в длительности по группе, чья покупательская активность снизилась

d) доля преобладающего целевого признака по категориям идентична: по уровню сервиса (в обеих покупательских группах преобладает стандартный уровень), по согласию на отправку сообщений (преобладает акцепт отправки) и по по наиболее популярной товарной категории (мелкая бытовая техника)

✔ В результате корреляционного анализа объединенных данных выявлен высокий уровень корреляции между данными о выручке в текущем и предыдущем месяце, а также данными о периоде с момента регистрации на сайте и количестве неоплаченных покупок в корзине. Для избежания влияния мультиколлинеарности, данные о выручке за текущий месяц и о периоде с момента регистрации были удалены

Есть две пары признаков с высокой корреляцией:

- `duration`(срок с момента регистрации на сайте) и `not\_paid\_item\_per\_q` (общее число неоплаченных товаров в корзине за последние 3 месяца)

- `revenue\_current` и `revenue\_per` - выручка за текущий и прошлый месяца

Удалим из датасета признаки, которые меньше влияют на покупательскую активность: `duration` из первой пары и `revenue\_current` из второй.

III. Выбор оптимальной модели для прогноза покупательской активности и определение значимых признаков покупательского поведения, влияющих на изменение покупательской активности

Перед началом подготовки данных для дальнейшего кодирования, изменим показатель 'promo\_buy' на категориальный.

Как раннее при анализе распределения признака было отмечено, что на гистограмме распределения признака четко прослеживается два пика,

т.е. перед нами ярковыраженное бимодальное распределение с границей межде ними в районе значения 0.5. В связи с этим преобразуем этот признак в категорийный

с порогом 0.5: все что выше порога 0.5 - 1, ниже порога - 0

Как было выявлено раннее, в данных целевого параметра - покупательская активность - есть преобладание категории 'прежний уровень' (61.7%), поэтому для оценки качества модели будем использовать метрику ROC\_AUC, которая не зависит от дисбаланса классов и не зависит от значения порога отнесения к тому или иному классу.

✔ В результате перебора моделей, их гиперпараметров и методов подготовки данных было определено, что наилучшей моделью для прогнозирования покупательской активности является модель опорных векторов - SVC() с масштабированием методом StandardScaler()

При такой конфигурации методов и модели метрика ROC-AUC на тренировочных составляет 0.9137, на тестовых - 0.9134

✔ Наиболее значимыми признаками модели являются

- pages\_per\_visit - количество просмотренных страниц: значение выше - вероятность принадлежности к классу 1 ниже

- minutes\_pre и minutes\_current - количество минут, проведенных покупателем на сайте: значение выше - вероятность принадлежности к классу 1 ниже

- avg\_cat\_viewse - среднее количество просмотренных категорий: значение выше - вероятность принадлежности к классу 1 ниже

- not\_paid\_item\_per\_q - количество неоплаченных товаров в корзине: прямая зависимость - значение выше - вероятность принадлежности к классу 1 выше

- revenue\_prepre - выручка от покупателя за пердыдущий месяц: значение выше - вероятность принадлежности к классу 1 ниже<

- market\_activ\_6mnth - среднее количество маркетинговых контактов за последние 6 мес: значение выше - вероятность принадлежности к классу 1 ниже

Прямая зависимость: рост значения признака приводит к росту вероятности принадлежности объекта к классу 1 наблюдается у признаков not\_paid\_item (неоплаченные товары в корзине) и promo\_1 (доля товаров, приобретенных по скидке, более 50%)

про матрицу корреляций. Работает с помощью метода пирсона

https://mindthegraph.com/blog/ru/pearson-correlation/

если что вставить сразу после y\_test = label\_encoder.transform(y\_test)

y\_1 = y\_train.sum() + y\_test.sum()

if y\_1 < 500:

print('Класс\_1 присвоен категории покупателей со сниженной покупательской активностью')

else:

print('Класс\_1 присвоен категории покупателей с прежним уровнем покупательской активности')

KNeighborsClassifier

Алгоритм K-ближайших соседей основан на принципе близости объектов в пространстве признаков. Он состоит в том, что для каждого объекта из тестовой выборки находят k ближайших соседей из обучающей выборки и классифицируют объект на основе классов его соседей. Класс, который наиболее часто встречается среди соседей, и будет классом, к которому относится исходный объект.

DecisionTreeClassifier — это тип алгоритма контролируемого обучения, который использует древовидную модель для классификации данных по различным категориям. 1

Алгоритм работает путём рекурсивного разбиения данных на более мелкие подмножества на основе значений входных объектов. Каждый внутренний узел в дереве представляет объект или атрибут, а каждый конечный узел представляет метку класса. 1

Процесс классификации включает в себя перемещение по дереву от корневого узла к конечному, при этом каждый узел принимает решение на основе входных признаков

SVC (Support Vector Classifier) — это алгоритм машинного обучения, который широко применяется для решения задач классификации. 1

Идея алгоритма заключается в поиске оптимальной разделяющей гиперплоскости в признаковом пространстве. Эта гиперплоскость должна максимально отделять объекты разных классов и поддерживать опорные векторы — точки данных, ближайшие к гиперплоскости. 1

SVC стремится максимизировать зазор между классами, то есть расстояние от опорных векторов до гиперплоскости. Это позволяет методу быть устойчивым к выбросам и хорошо обобщать на новые данные. 1

LogisticRegression

Идея алгоритма логистической регрессии заключается в том, что с помощью набора входных переменных он моделирует вероятность конкретного исхода. 1

Алгоритм работает, моделируя связь между независимыми переменными (предикторами) и зависимой переменной (целевой) с помощью сигмовидной функции, которая выводит вероятность в диапазоне от 0 до 1. 1 Эта вероятность указывает на то, вероятно ли, что данный вход соответствует одной из двух предопределённых категорий. 2

ROC\_AUC

Метрика ROC AUC простыми словами — это мера способности классификатора различать классы. 2 Она позволяет суммировать производительность модели одним числом, измеряя площадь под кривой ROC (рабочей характеристики приёмника). 12

Чем выше показатель AUC, тем качественнее классификатор. 3 AUC колеблется от 0 до 1, где более высокое значение указывает на более высокую производительность модели. AUC равный 0,5 указывает на отсутствие дискриминационной способности модели, тогда как AUC равный 1,0 означает идеальное различие классов. 1

DummyClassifier makes predictions that ignore the input features.

GridSearchCV — это функция из пакета model\_selection Scikit-learn, которая позволяет провести решётчатый поиск по заданным наборам гиперпараметров. 12

RandomizedSearchCV — это метод для гиперпараметрического поиска в библиотеке Scikit-learn для машинного обучения в Python. 1

Он вместо перебора всех возможных комбинаций случайным образом выбирает заданное количество наборов гиперпараметров

#для перебора гиперпараметров моделей создадим список словарей,

# в котором каждый словарь — это модель с гиперпараметрами и методами подготовки данных

param\_distributions = [

# словарь для модели KNeighborsClassifier()

# {

# # название модели

# 'models': [KNeighborsClassifier()],

# # указываем гиперпараметр модели n\_neighbors

# 'models\_\_n\_neighbors': range(1, 20),

# # указываем список методов масштабирования

# 'preprocessor\_\_num': [StandardScaler(), MinMaxScaler(), 'passthrough']

# },

# # словарь для модели DecisionTreeClassifier()

# {

# 'models': [DecisionTreeClassifier(random\_state=RANDOM\_STATE)],

# 'models\_\_max\_depth': range(2, 11),

# 'preprocessor\_\_num': [StandardScaler(), MinMaxScaler(), 'passthrough']

# },

# словарь для модели SVC()

{

'models': [SVC(random\_state=RANDOM\_STATE, probability=True)],

'preprocessor\_\_num': [StandardScaler(), MinMaxScaler(), 'passthrough']

},

# словарь для модели LogisticRegression()

# {

# 'models': [LogisticRegression(

# random\_state=RANDOM\_STATE,

# solver='liblinear',

# penalty='l1'

# )],

# 'models\_\_C': range(1, 5),

# 'preprocessor\_\_num': [StandardScaler(), MinMaxScaler(), 'passthrough']

# }

]