Проект: Обучение с учителем: качество модели

Описание проекта

Интернет-магазин «В один клик» продаёт разные товары: для детей, для дома, мелкую бытовую технику, косметику и даже продукты. Отчёт магазина за прошлый период показал, что активность покупателей начала снижаться. Привлекать новых клиентов уже не так эффективно: о магазине и так знает большая часть целевой аудитории. Возможный выход — удерживать активность постоянных клиентов. Сделать это можно с помощью персонализированных предложений.

«В один клик» — современная компания, поэтому её руководство не хочет принимать решения просто так — только на основе анализа данных и бизнес-моделирования.

Цель исследования - разработать решение, которое позволит персонализировать предложения постоянным клиентам, чтобы увеличить их покупательскую активность.

Описание данных

market_file.csv - таблица, которая содержит данные о поведении покупателя на сайте, о коммуникациях с покупателем и его продуктовом поведении.

market_money.csv - таблица с данными о выручке, которую получает магазин с покупателя, то есть сколько покупатель всего потратил за период взаимодействия с сайтом.

market_time.csv - таблица с данными о времени (в минутах), которое покупатель провёл на сайте в течение периода.

money.csv - таблица с данными о среднемесячной прибыли продавца за последние 3 месяца: какую прибыль получает магазин от продаж каждому покупателю.

Этапы проекта:

1. Загрузка данных
2. Предобработка данных
3. Исследовательский анализ
4. Объединение таблиц
5. Корреляционный анализ
6. Использоваине пайплайнов
7. Анализ важности признаков
8. Сегментация покупателей
9. Общий вывод

В работу поступило 3 датасета

- market_file - 13 столбцов, 1300 строк, 0 пропусков, 0 дубликатов;
- market_money - 3 столбцов, 3900 строк, 0 пропусков, 0 дубликатов;
- market_time - 3 столбцов, 2600 строк, 0 пропусков, 0 дубликатов;
- money - 2 столбцов, 1300 строк, 0 пропусков, 0 дубликатов.

В market_file и market_time отредактированы название столбцов. В money произведено разделение столбцов и проведена замена тип столбца "Прибыль" на float64.

Во всех четырёх датасетах название столбцов соотвествуют описанию.

Было выполнено:

- market_file - Исправлена опечатка в обозначении данных на "стандарт". В столбце "Длительность" данные без выбросов, в столбцах "Акционные_покупки" и "Маркет_актив6мес" имеются выбросы, принято решение их не удалять т.к. их достаточно много их влияние оценим на следующем шаге.
- market_money - Были удаленны два значения 106862.2 - это явный выброс и 0 - это значение удаленно по условиям исследования. Так же большое колличество выбросов которые не будут удаленны.
- market_time - Исправленна опечатка в данных на 'предыдущий_месяц'. Данные без выбросов.
- money - Значения прибыли имеют выбросы их удалять не будем.

Обобщая информацию по всем 4 датафреймам можно увидеть, что даже при снижении маркетинговой активности время пользователей в магазине не изменилось, выручка при этом расла последнии 3 месяца.

Пользователи со сниженной покупательской активностью просматривают страниц и категорий товаров одинаково, что даёт нам понять, что они ищут товары только по акции. Они же получают всех больше рекламных уведомлений от магазина.

Пользователи с прежней активностью просматривают большое количество страниц в магазине и различные категории, и отправляют товары в корзину и оплачивают их когда в корзине становится 6 и более товаров.

По итогу 1296 клиентов сохраняло свою активность на протяжении не менее трёх месяцев.

Итоговая таблица market_result состоящая из market_file, market_money, market_time:

- 18 столбцов,
- 1296 записей,
- пропуски отсуствуют,
- дубликаты отсуствуют.

О мультиколлинеарности можно говорить если значение корреляции выше 0.9. Такой корреляции для целевого признака, покупательская активность, нет.

Проведён дополнительный анализ по VIF метрике. По её условиям о мультиколлинеарности можно говорить при значении метрики выше 5, не один параметр не отвечает данному условию.

Поэтому можно сделать вывод об отсуствии мультиколлиарности.

Способ для поиска модели и гиперпараметров был выбран RandomizedSearchCV потому, что перед нами стоит задача найти лучшую модель идля экономии времени был выбран этот способ при следующих параметрах:

- cv=5 стандарное количество раз тестирования модели,
- scoring - стратегия оценки производительности модели была выбрана 'roc_auc' т.к. эта метрика работает при оценки сразу 4 моделей и отсуствует необходимость выберать определённые пороги отсечения и учитывает возможность дисбаланса классов.

В результате работы поплайна были обработаны 4 модели: KNeighborsClassifier(), DecisionTreeClassifier(), LogisticRegression() и SVC(). Лучшей стала с гиперпараметрами SVC:

- C - параметр регуляризации = 1.668,
- kernel - ядро linear,
- degree - степень полимиального ядра = 10,
- random_state=100)

Метрики:

- метрика по кросс валидационной выборке: 0.897
- ROC-AUC на тестовых данных: 0.906
- F1 на тестовых данных: 0.882

Модель выдаёт довольно хорошие метрики, что даёт нам уверенность в её работоспособности.

Для модели важны признаки (интрепритация по beeswarm):

- Самыми важными признаками для сохранения покупательской активности просмотр страниц_за_визит, минут_текущий_месяц, минут_предыдущий_месяц, средний_просмотр_категорий_за_визит, маркет_актив6мес;
- Признаки неоплаченные_продукты_штук_квартал, разрешить_сообщать_нет, ошибка_сервиса, маркет_актив_тек_мес отрицательно влияют на сохранение покупательской активности, но их значение настолько невелико, что можно предположить, что нам необходимо искать другую причину влияющую на снижение покупательской активности.

Было принято решение выделить клинтов которые покупают в основном по акции и вероятностью снижения покупательской активности т.к. за эту категорию клиетов следует поборотся и она более тругих подвержена уходу.

В ходе анализа было выявлено, что такие клиенты активно пользуются премиумной подпиской 35%. Чаще покупают товары для детей и плохо различную бытовую технику. Просматривают только нужные категории и страницы в магазине. Накапливают неоплачиваемые товары в корзине.

Можно предположить, что такими клиентами являются клиенты с детьми и находящийся матерью в декрете. И клиенты с небольшим доходом которым нужны конкретные товары и нет средств на излишки.

Можно выдать рекомендации направить магазину больше рекламы об акциях, для повышения просмотров категорий и страниц в магазине. При возможности предлогать рассрочку в оплате товаров, для повышения лояльности клиентов, повышения товаро оборота и снижения тем самым скапливаемости товаров на складах и снижение складских расходов.