РОЛЬ В ПРОЕКТЕ

Роль в проекте

Анализ и подготовка данных для модели машинного обучения.

Основные исследовательские проблемы:

- Исследовать, какие данные и какого качества доступны;
- Проанализировать возможности получения нужных целевых переменных для модели;
- Проверить гипотезы о возможности построения модели на имеющихся данных;
- Сконструировать и отобрать оптимальные признаки для модели.

Сложности:

- Много источников данных (23 таблицы в MySQL и Clickhouse, 3 файла csv с историческими данными);
- Высокая разреженность и низкое качество отдельных признаков;
- Проблемы на уровне сбора данных (например, отсутствуют отдельные временные периоды).

Выполнены задачи в рамках роли

2 семестр

1. Исследование теоретических аспектов темы динамического ценообразования и опыта применения машинного обучения в области
2. Подготовка обзора литературы по теме (~70 источников)
3. Определение собственной роли в проекте, темы исследования, структуры будущей научной работы
4. С командой разработана первичная гипотеза об архитектуре модели

3 семестр

1. Проведен EDA основных данных по клиентам, заказам и продуктам
2. Проведена очистка от дубликатов
3. Обработаны отсутствующие значения
4. Обработаны выбросы и аномалии
5. Сделаны аналитические выводы для компании
6. Обнаружены проблемы с основной целевой переменной – конверсией продаж
7. Компании представлены рекомендации по сбору данных

4 семестр

1. Разработаны альтернативные варианты модели
2. Проведен EDA всех предоставленных компанией данных, произведены необходимые слияния и преобразования, сгенерированы новые признаки и целевые переменные
3. Выбрана альтернативная модель динамического ценообразования
4. Произведено ранжирование и отбор признаков для финальной модели
5. Произведен расчет экономического эффекта от внедрения модели
6. Написан автономный блок подготовки данных для модели на Python

Использованы библиотеки Python

Связь с базами данных:
- mysql.connector, sqlalchemy, requests, clickhouse_driver, clickhouse_sqlalchemy

Работа с массивами:
- pandas, numpy

Расчет статистик:
- statistics, math

Работа с временными признаками:
- datetime, matplotlib.dates, workalendar.europe (календарь праздников)

Обработка текстовых признаков:
- re, transliterate, ast (преобразование строк в списки)

Анализ пропусков:
- missingno

Проверка дубликатов:
- Recordlinkage

Визуализация данных:
- Matplotlib.pyplot, seaborn, plotly

Отбор и инженерия признаков:
- phik, phik.report, featuretools, featuretools.selection, catboost, sklearn.model_selection, sklearn.metrics, sklearn.inspection, shap

Исследование данных

Сырые данные:

Название таблицы	Число признаков
MySQL (17 таблиц) :	
all_prices_listing	5
all_prices_with_competitors	7
mindbox_all_clients	81
mindbox_applied_promotions	16
mindbox_categories_names	6
mindbox_actions	6
mindbox_history_categories	5
mindbox_history_products	5
mindbox_order_lines	18
mindbox_orders	14
mindbox_products	41
mindbox_tracking_categories	5
mindbox_tracking_products	5
pricing	7
products_impressions_all	7
products_visits_count_by_datetime	6
yandex_client_ids	6
Clickhouse (6 таблиц) :	
Hits	5
metrica_goals	40
metrica_	46
mindbox_clients	5
page_views_materialized	9
visits	49
CSV :	
customers	89
orders	131
products	41
ВСЕГО ПРИЗНАКОВ	**655**

После EDA:

Название таблицы	Число признаков	Число пропусков, %
MySQL (7 таблиц):		
mindbox_all_clients	24	57%
mindbox_order_lines	10	< 1%
mindbox_orders	5	3%
mindbox_products	4	< 1%
pricing	6	6%
products_visits_count_by_datetime	2	0
yandex_client_ids	4	< 1%
Clickhouse :		
visits	16	19%
CSV :		
customers	25	58%
orders	18	22%
products	15	23%
Внешние данные:		
female_names_rus	1	
male_names_rus	1	
cities_ru_en	2	
city	9	
urov_10subg-nm	2	
ВСЕГО ПРИЗНАКОВ	**144**	

Преобразование данных

График исключен согласно NDA

После исследования и обработки аномалий

Обрезаны данные до августа 2022г. (большое число аномалий из-за тестов на сайте)

После слияний, обработки и генерации признаков:

Число признаков
105 (из них 76 созданных)

После отбора признаков под финальную модель:

Число признаков
25 (из них 4 категориальных)

Целевая переменная - Первичная идея

Целевая: конверсия в продажу
Модель должна минимизировать скидку на товар так, чтобы при этом конверсия не снижалась.

Необходимые компоненты:

- Статистика по ценам и скидкам на каждый товар в каждый момент времени
- ~~Статистика просмотров каждого товара в каждый момент времени~~
  Аномалии:
  - не зафиксированы просмотры в период, когда были продажи
  - на некоторых товарах просмотров меньше, чем продаж
  - разные данные об активности в Яндекс Метрике и Mindbox
- Статистика покупок каждого товара в каждый момент времени

Дополнительные компоненты:

- Индивидуальные признаки клиента, позволяющие отнести его к какому-то сегменту и подстроить цену еще более точно

Целевая переменная – Финальная версия

Средняя "склонность к скидкам" для каждого клиента – измеритель того, как часто клиент покупал товары со скидкой, выходящей за границы межквартильного размаха скидок по конкретному товару.

График исключен согласно NDA

Целевая переменная – Финальная версия

Целевая: склонность клиента к скидкам
Модель должна предложить клиенту скидку максимально близкую к его персональным ожиданиям

Необходимые компоненты:
- Индивидуальные признаки клиента, позволяющие отнести его к какому-то сегменту

Дополнительные компоненты:
- Нужны допуски максимума и минимумы цены на каждый товар, в пределах которых модель будет подстраиваться (требуются внутренние данные компании по себестоимости)
- Конверсия позволила бы отследить изменения в поведении, у нас пока есть только общие данные по сайту
- Анализ отклонений общих значений продаж по ARIMA

Риски:
- Все захотят максимальных скидок

Целевая переменная – Финальная версия

Средняя "склонность к скидкам": чем ближе к -1, тем менее чувствителен человек к скидкам

График исключен согласно NDA

График исключен согласно NDA

Отбор признаков

График исключен согласно NDA

Ручное исключение:
- Признаки-ключи,
- Признаки, появляющиеся только по факту оформления заказа

Статистические методы:
- Линейная корреляция
- Phi_K корреляции

Значимость признаков:
- Permutation Importance (scikit-learn)
- Feature Importance (Catboost)

Отбор признаков:
- featuretools.selection
- CatBoost Feature Selection
- Shap

Финальные признаки для модели

График исключен согласно NDA

Источники формирования признаков:	
mindbox_all_clients	12
mindbox_order_lines	10
внешние данные	5
mindbox_orders	4
yandex_client_ids	3
visits	2

Тестовый прогон модели

Тестовая модель: CatBoostRegressor (без подбора гиперпараметров)

Метрики качества	**До обработки**	**После обработки**	**Только новые клиенты**
MSE			
R2			
MAE			

Тестовая модель: CatBoostClassifier (без подбора гиперпараметров)

	precision	recall	f1-score
0			
1			
2			

accuracy