СБЕР МАРКЕТ

Прогноз спроса и планирование рабочих смен

Захаров Андрей

Senior Data Scientist

Telegram: @aizakharov94

Безопасники

Анонимизация данных:
1) Добавлен шум
2) Изменен порядок чисел
3) Убрана часть данных
4) Убраны доп. данные

Доп. данные

1) Маркетинговые компании
2) Промо акции
3) Погода
4) Фрод партнеров
5) Признаки точек и ТД
6) Город

СБЕР МАРКЕТ

Доступные данные

История заказов

delivery_area_id	date	orders_cnt
29	2021-07-06 13:00:00	1
317	2021-10-16 16:00:00	2
39	2021-10-11 20:00:00	4
458	2021-09-28 12:00:00	3
41	2021-07-31 13:00:00	10
509	2021-10-28 12:00:00	1
559	2021-09-01 20:00:00	4
411	2021-09-30 19:00:00	1
134	2021-09-02 14:00:00	3
251	2021-09-16 10:00:00	3

Выводы партнеров и их опоздания

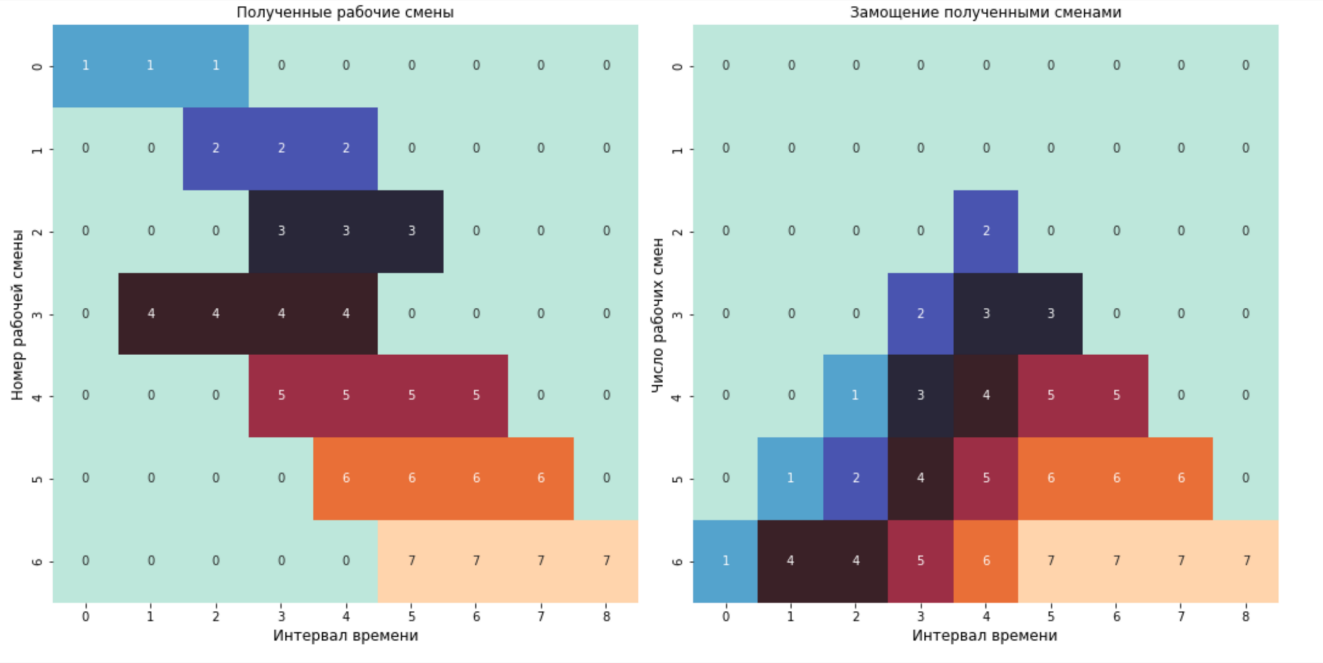
delivery_area_id	dttm	partners_cnt	delay_rate
591	2021-11-27 17:00:00	3.0	0.000000
150	2021-11-08 16:00:00	2.0	0.000000
419	2021-04-02 17:00:00	1.0	0.000000
59	2021-06-14 20:00:00	1.0	0.000000
342	2021-06-22 20:00:00	1.0	0.000000
31	2021-06-27 20:00:00	3.0	0.166667
27	2021-06-26 12:00:00	2.0	0.250000
431	2021-10-13 10:00:00	3.0	0.000000
224	2021-09-20 19:00:00	1.0	0.000000
265	2021-07-01 15:00:00	1.0	0.000000

Что надо сделать

Коротко: Запланировать рабочие смены для курьеров на 7 дней вперед.

СБЕР МАРКЕТ

Что надо сделать



Что надо сделать

shift_name	2022-09-12	2022-09-13	2022-09-14
shift_1-13	7	5	0
shift_10-22	9	3	8
shift_11-23	16	20	15
shift_12-0	0	6	6
shift_13-1	16	6	9
shift_2-14	0	2	0
shift_20-7	0	0	0

Что надо сделать

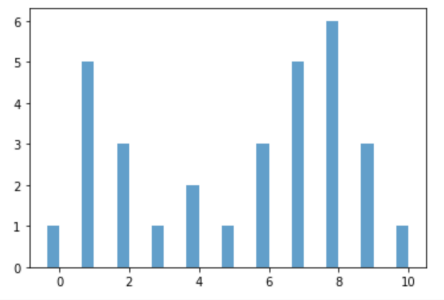
Более строго: Создать продукт (код), который на ежедневной основе «тригерит» функцию (F) для подсчета рабочих смен.

Функция F:

1) Принимает
   - Таблицу с актуальной историей по заказам
   - Таблицу с опозданиями
2) Возвращает
   - Таблицу с расписанием рабочих смен

Как достичь

1) Спрогнозировать количество заказов по часам на 7 дней вперед
2) В зависимости, сколько будет в i-ый час заказов, рассчитать, сколько надо курьеров
3) Потребность в курьерах разбить на смены (от 4-ех до 8-ми сменами)



Как достичь (baseline)

1) Спрогнозировать количество заказов по часам на 7 дней вперед **(Алгоритм без ML)**
2) В зависимости, сколько будет в i-ый час заказов, рассчитать, сколько надо курьеров **(Константа)**
3) Потребность в курьерах разбить на смены (от 4-ех до 8-ми сменами) **(Жадный перебор)**

Мб есть какой-то другой путь?

Как достичь (хотелось бы)

1) Спрогнозировать количество заказов по часам на 7 дней вперед **(Алгоритм ML)**
2) В зависимости, сколько будет в i-ый час заказов, рассчитать, сколько надо курьеров **(Алгоритм ML)**
3) Потребность в курьерах разбить на смены (от 4-ех до 8-ми сменами) **(Оптимизационный алгоритм)**

Мб есть какой-то другой путь?

Для чего

1) Понять, как работает планирование спроса
2) Послушать интересные идеи и реализовать у нас
3) Хорошая практика на реальной задаче

С чего начнем

Написать дизайн решения (решение без кода):

1) Как по вашему мнению нужно решать задачу (каждую из частей)
2) Какие признаки вы бы использовали в алгоритме ML
3) Какие подходы вы бы использовали *(самый большой пункт)*
4) Природа алгоритма (NN, деревья, бустинг, LR, ...)
5) Приемлемость простого решения (константа). Надо будет обосновать.

Итог: небольшой отчет (текстовый файлик)

СБЕР МАРКЕТ

Первая задача (прогноз заказов)

Идеи к реализации

1) Авторегрессионные модели (ARIMA, SAR, ...)
2) Prophet от Facebook, Orbit от Uber, ...
3) Свою модельку (тут подробнее)

Своя модель

Первый подход:

Одна модель. Предсказывает количество заказов в конкретный час в конкретный день. Как признак вставить час, день предсказания.

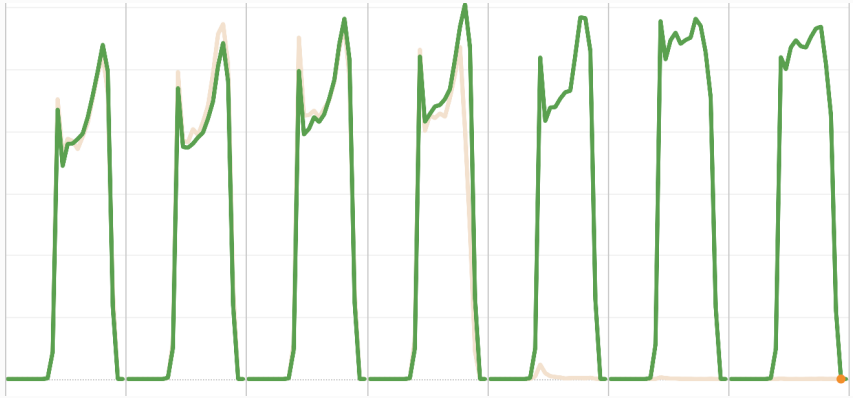
- ❑ Плюсы: одна модель
- ❑ Минусы: Сложно работать с признаками
  - ➢ Сложно вставить признак: «какое значение было неделю назад»
  - ➢ На первый день предсказывать легче, чем на седьмой, так как вы не знаете информацию за предыдущие 6 дней (если предсказываете 7-ой день)

Нюанс номер 1

Распределение заказов внутри дня.
На графике: распределение по часам внутри дня для каждого дня недели (с понедельника по воскресенье)
Как можете заметить, что распределение по часам в **будни** практически не отличаются:

- ❑ Пик заказов утром
- ❑ Пик заказов вечером (после работы)

В выходные выглядит более равномерно (людям не надо на работу и они могут заказывать в любое время)



Хорошая идея: попытаться предсказывать дневные заказы и потом разбить внутри дня, согласно распределению.

Своя модель

Второй подход:

Семь моделей. Предсказываем каждый день отдельно (кол-во заказов завтра, кол-во заказов послезавтра, … , кол-во заказов на 7-ой день)

- ❑ Плюсы: Можно вставить признак: «какое значение было неделю назад», «процентное изменение с тем, что было неделю назад» и подобные признаки.
- ❑ Минусы: Семь моделей

Выбор модели

Можно реализовывать, что душе угодно.
Я подробнее расскажу про второй подход и подводные камни.
Можете попробовать реализовать его, можете взять идеи из него и применить к своему подходу.

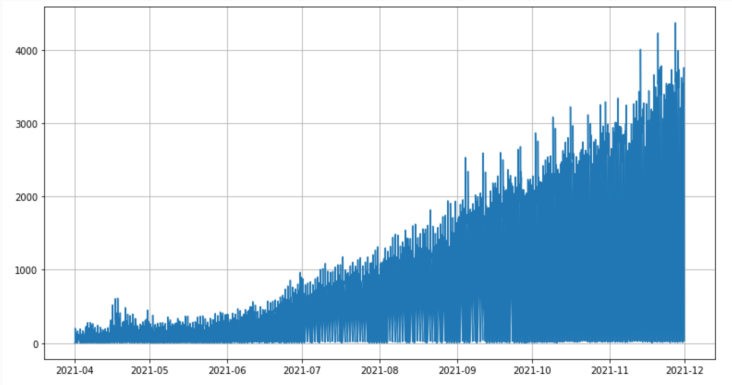
Нюанс номер 2

Растущий рынок

На графике: Суммарное количество заказов в час по всем ТД.

Как можете заметить, что СберМаркет постоянно растет. Поэтому:

- ❑ Лучше брать признаки из ближайшего прошлого (2-3 недели)
- ❑ Предсказывать не абсолютную, а относительную величину.



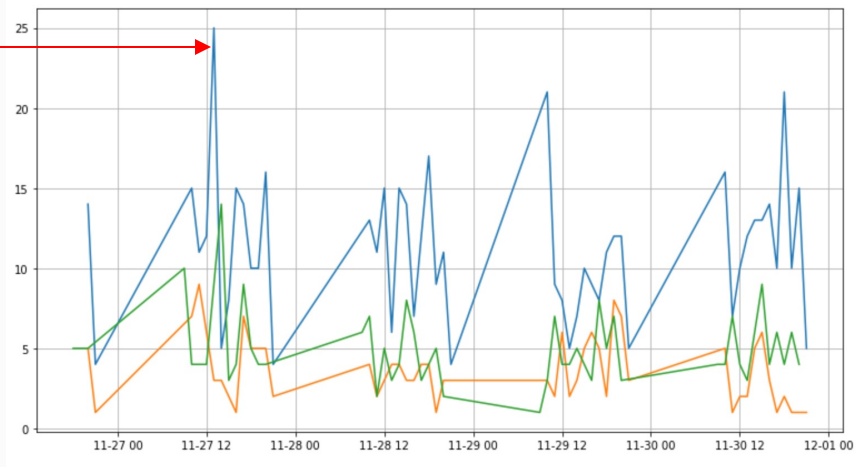
Хорошая идея: попытаться предсказывать относительное изменение по сравнению с прошлой неделей. Далее опишу подробнее.

Нюанс номер 3

Разный порядок кол-ва заказов

На графике: Количество заказов для трёх разных территорий доставки. Как можете заметить, у одной ТД диапазон заказов больше, чем у других. Поэтому:

- ❑ Хорошо бы сделать нормировку, чтобы сделать выборку максимально приближенной к однородной.
- ❑ Нормировать хорошо на устойчивую величину, а то **выбросы** могут испортить картину.



Хорошая идея: нормировать на медиану или на квантильное среднее значение прошлой недели.

СБЕР МАРКЕТ

Второй подход

prev_i – значение i дней назад
future_i – таргет i дней вперед

Генерация обучающей выборки:

	date	delivery_area_id	target
0	2021-04-01	0	24
1	2021-04-02	0	21
2	2021-04-03	0	33
3	2021-04-04	0	18
4	2021-04-05	0	32
...	...	...	...
95181	2021-11-19	592	53
95182	2021-11-20	592	56
95183	2021-11-21	592	47
95184	2021-11-22	592	52
95185	2021-11-23	592	49

	date	delivery_area_id	prev_7	prev_6	prev_5	prev_4	prev_3	prev_2	prev_1	future_1	future_2	future_3	future_4	future_5	future_6	future_7
0	2021-04-02	0	NaN	NaN	NaN	NaN	NaN	NaN	24.0	21	33.0	18.0	32.0	29.0	36.0	23.0
1	2021-04-03	0	NaN	NaN	NaN	NaN	NaN	24.0	21.0	33	18.0	32.0	29.0	36.0	23.0	28.0
2	2021-04-04	0	NaN	NaN	NaN	NaN	24.0	21.0	33.0	18	32.0	29.0	36.0	23.0	28.0	22.0
3	2021-04-05	0	NaN	NaN	NaN	24.0	21.0	33.0	18.0	32	29.0	36.0	23.0	28.0	22.0	17.0
4	2021-04-06	0	NaN	NaN	24.0	21.0	33.0	18.0	32.0	29	36.0	23.0	28.0	22.0	17.0	20.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
94588	2021-11-19	592	43.0	42.0	37.0	51.0	49.0	50.0	43.0	53	56.0	47.0	52.0	49.0	NaN	NaN
94589	2021-11-20	592	42.0	37.0	51.0	49.0	50.0	43.0	53.0	56	47.0	52.0	49.0	NaN	NaN	NaN
94590	2021-11-21	592	37.0	51.0	49.0	50.0	43.0	53.0	56.0	47	52.0	49.0	NaN	NaN	NaN	NaN
94591	2021-11-22	592	51.0	49.0	50.0	43.0	53.0	56.0	47.0	52	49.0	NaN	NaN	NaN	NaN	NaN
94592	2021-11-23	592	49.0	50.0	43.0	53.0	56.0	47.0	52.0	49	NaN	NaN	NaN	NaN	NaN	NaN

СБЕР МАРКЕТ

Второй подход

Нормировка на медиану

prev_7	prev_6	prev_5	prev_4	prev_3	prev_2	prev_1	future_1	future_2	future_3	future_4	future_5	future_6	future_7
NaN	NaN	NaN	NaN	NaN	NaN	24.0	21	33.0	18.0	32.0	29.0	36.0	23.0
NaN	NaN	NaN	NaN	NaN	24.0	21.0	33	18.0	32.0	29.0	36.0	23.0	28.0
NaN	NaN	NaN	NaN	24.0	21.0	33.0	18	32.0	29.0	36.0	23.0	28.0	22.0
NaN	NaN	NaN	24.0	21.0	33.0	18.0	32	29.0	36.0	23.0	28.0	22.0	17.0
NaN	NaN	24.0	21.0	33.0	18.0	32.0	29	36.0	23.0	28.0	22.0	17.0	20.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
43.0	42.0	37.0	51.0	49.0	50.0	43.0	53	56.0	47.0	52.0	49.0	NaN	NaN
42.0	37.0	51.0	49.0	50.0	43.0	53.0	56	47.0	52.0	49.0	NaN	NaN	NaN
37.0	51.0	49.0	50.0	43.0	53.0	56.0	47	52.0	49.0	NaN	NaN	NaN	NaN
51.0	49.0	50.0	43.0	53.0	56.0	47.0	52	49.0	NaN	NaN	NaN	NaN	NaN
49.0	50.0	43.0	53.0	56.0	47.0	52.0	49	NaN	NaN	NaN	NaN	NaN	NaN

prev_7	prev_6	prev_5	prev_4	prev_3	prev_2	prev_1	future_1	future_2	future_3	future_4	future_5	future_6	future_7	1_week_median
NaN	NaN	NaN	NaN	NaN	NaN	1.000000	0.875000	1.375000	0.750000	1.333333	1.208333	1.500000	0.958333	24.0
NaN	NaN	NaN	NaN	NaN	1.066667	0.933333	1.466667	0.800000	1.422222	1.288889	1.600000	1.022222	1.244444	22.5
NaN	NaN	NaN	NaN	1.000000	0.875000	1.375000	0.750000	1.333333	1.208333	1.500000	0.958333	1.166667	0.916667	24.0
NaN	NaN	NaN	1.066667	0.933333	1.466667	0.800000	1.422222	1.288889	1.600000	1.022222	1.244444	0.977778	0.755556	22.5
NaN	NaN	1.000000	0.875000	1.375000	0.750000	1.333333	1.208333	1.500000	0.958333	1.166667	0.916667	0.708333	0.833333	24.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1.000000	0.976744	0.860465	1.186047	1.139535	1.162791	1.000000	1.232558	1.302326	1.093023	1.209302	1.139535	NaN	NaN	43.0
0.857143	0.755102	1.040816	1.000000	1.020408	0.877551	1.081633	1.142857	0.959184	1.061224	1.000000	NaN	NaN	NaN	49.0
0.740000	1.020000	0.980000	1.000000	0.860000	1.060000	1.120000	0.940000	1.040000	0.980000	NaN	NaN	NaN	NaN	50.0
1.020000	0.980000	1.000000	0.860000	1.060000	1.120000	0.940000	1.040000	0.980000	NaN	NaN	NaN	NaN	NaN	50.0
0.980000	1.000000	0.860000	1.060000	1.120000	0.940000	1.040000	0.980000	NaN	NaN	NaN	NaN	NaN	NaN	50.0

Второй подход

Генерация признаков:

1) Различные статистики из прошлого (mean, std, median, ...)
2) Закодировать delivery_area_id (мб поможет)
3) Временные признаки:

- День недели
- Праздник или нет
- Сколько дней до праздника
- Сколько дней с предыдущего прздника
- ...

4) Еще какие-то

Второй подход

Модель:

1) Нужно обучить 7 моделей для предсказания future_1, ... , future_7
2) Умножить предсказания на медиану
3) Разбить дневные заказы по часам на основе распределения из истории

Второй подход

Pipeline обучения:

1) df = get_data(df) (получаем обучающую выборку)
2) df = get_scaled_data(df) (нормировка)
3) df = get_features(df) (получаем признаки)
4) model = your_model.fit(df) (обучение семи моделей на каждый день)

Pipeline предсказания:

1) test = get_data(test) (получаем выборку)
2) test = get_scaled_data(test) (нормировка)
3) test = get_features(test) (получаем признаки)
4) predictions = your_model.predict(test) (получаем ответы)
5) final_df = get_hours_distribution(test, predictions) распределяем заказы по часам)

Нюанс 4

1) Будем смотреть на средний показатель MAPEпо каждому из 7-ми дней (https://en.wikipedia.org/wiki/Mean_absolute_percentage_error)
2) Можете сравнивать свою модель с baseline. Лучший бэйзлайн: прошлая неделя как ответ. Можно немного улучшить и умножить на тренд.
3) Будут вопросы: пишите.

Удачи