DANO

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

Т БАНК

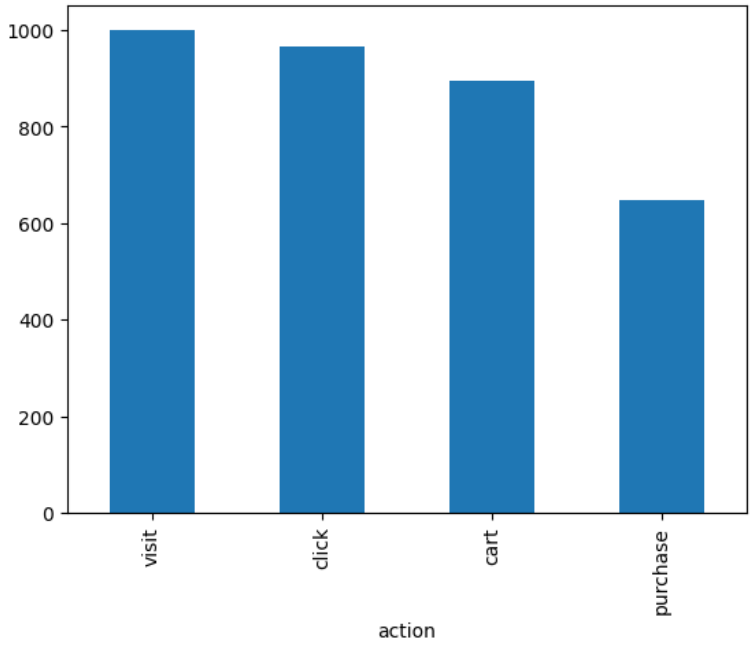
Задача 1 (12 баллов)

1. $E = \frac{ln(\frac{Q}{Q_0}) \times P_0}{P_0 - P} = 0{,}51$
2. При E = 0,51 и P = 22000 Q = 190 055
3. При текущих моделях ножей продастся 1000, сковородок 5000 -> Продавец заработает 1000 * 2000 + 5000 * 1000= 7_000_000 рублей

Задача 2 (23 балла)

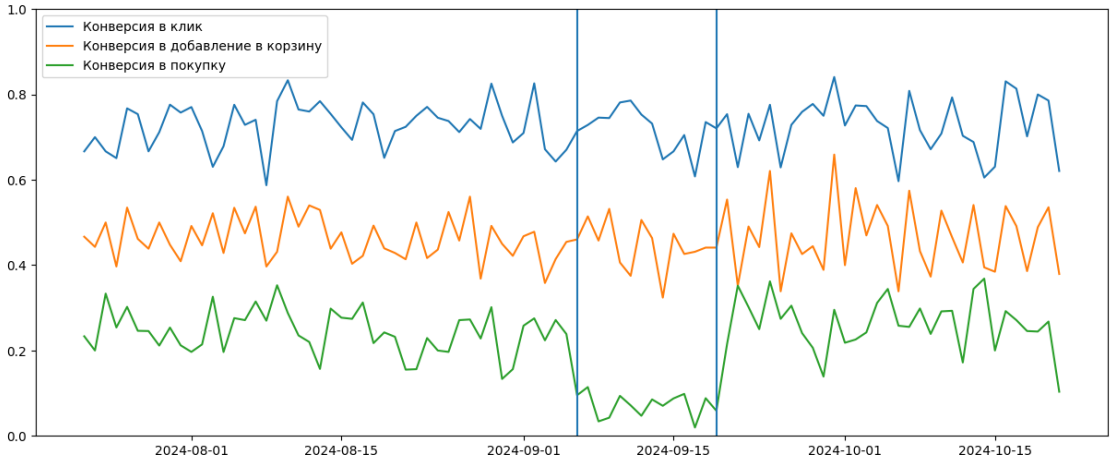
1. Находим кол-во уникальных пользователей на каждом шагу воронки. Это:
   - cart 894
   - click 967
   - purchase 647
   - visit 1000

   Получаем:
   - Из визита на сайт в клик на карточку товара: 0.967=967/1000
   - Из клика в добавление в корзину: 0.925=894/967
   - Из добавления в корзину в покупку: 0.724=647/894
   - Из визита на сайт в добавление в корзину: 0.894=894/1000
   - Из визита на сайт в покупку: 0.647=647/1000
2. Используем числа из 1 пункта, чтобы построить график:



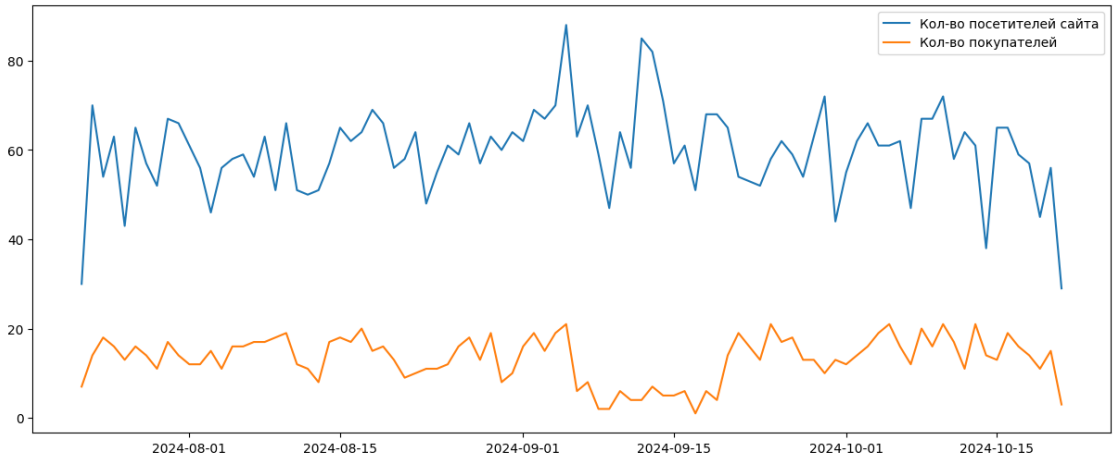
3. По 1 пункту задачи видим, что конверсия из предыдущего шага ниже всего у конверсии из добавления в корзину в покупку: 0.724. Гипотеза: трата денег является самым сложным шагом в пути пользователя и поэтому конверсия из добавления товара в корзину в покупку самая низкая. Механизм: для пользователя деньги это его ограниченный ресурс — остальные шаги (посещение, клик, добавление в корзину) являются бесплатными для пользователя — трата денег останавливает большую часть пользователей, чем остальные этапы — конверсия из добавления товара в корзину в покупку самая низкая

4. Считаем кол-во уникальных пользователей для каждого типа действия и календарного дня. Для каждого дня находим значение конверсии, аналогично 1 пункту. Строим график



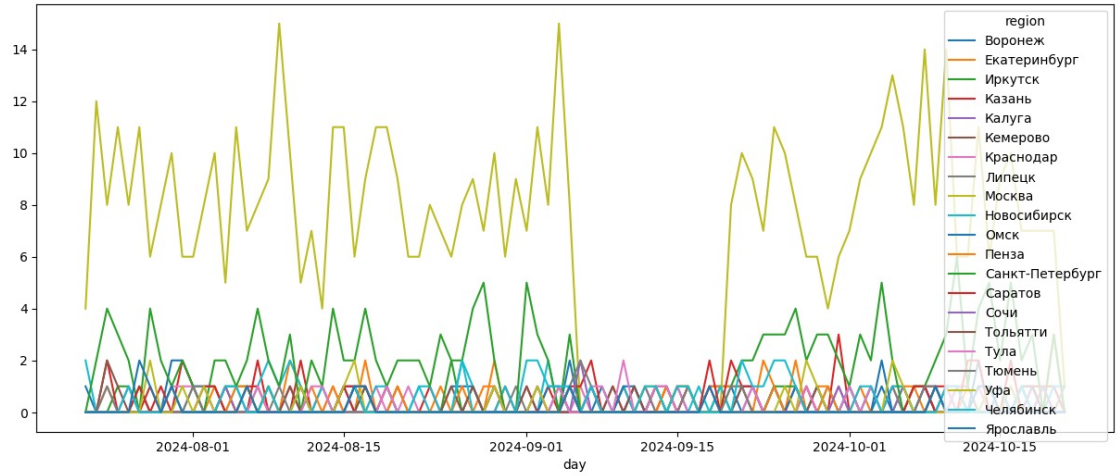
5. Видим, что просадка конверсии в покупку была примерно в период 6-19 сентября.

6. Проверяем, отдельно числитель и знаменатель этой конверсии:



Видим, что просадка объясняется снижением кол-ва покупателей — кол-во посетителей колеблется на одном и том же уровне.

Возьмем середину периода просадки (даты примерно с 9 сентября по 15 сентября — неделю), чтобы наиболее явно видеть изменения, и аналогичный период сильно до просадки (чтобы случайно не захватить аномалию). Посмотрим разбивки по основным столбцам: product_id, delivery_price, sex, region. Для всех категорий, кроме региона, видим, что разбивка не выявляет какой-то закономерности — кол-во покупок падает у всех видов товаров, у разных типов доставок, у разных полов. Однако в случае регионов видим, что покупки просели только у половины. Смотрим динамику с разбивкой по регионам и видим, что в 0 ушли покупки в Москве и Питере — это может быть нашей технической причиной:



Получили:

— снижение конверсии было вызвано снижением кол-ва покупателей в основных регионах: Москва и Санкт-Петербург

— причиной этому могли послужить логистические проблемыс этими регионами, которые не позволяли пользователю оформить доставку и соответственно покупку

Задача 3 (25 баллов)

Ошибки:

1. Отсутствие взаимосвязи между возрастом и удовлетворенностью.

   Комментарий. Используемая модель множественной регрессии учитывает только линейные взаимосвязи между переменными. На графиках видна квадратичная зависимость между возрастом и удовлетворенностью, поэтому оценивая в модели только саму переменную возраста мы увидим, что коэффициент незначим в данной модели. Отсутствие линейной взаимосвязи видно по коэффициенту корреляции (из матрицы корреляций) между возрастом и удовлетворенностью (-0.000892 0), но это не означает отсутствие взаимосвязи как таковой (потому что она может быть нелинейной). Корректно было бы добавить в модель две отдельные переменные для возраста: сам показатель и возраст в квадрате. В выводе неверно указано про отсутствие взаимосвязи между переменными. Корректнее было бы указать, что отсутствует линейная взаимосвязь между этими переменными. Корректный вывод об отсутствии или наличии значимой квадратичной взаимосвязи можно сделать только переоценив модель, добавив в нее переменную возраста в квадрате.

2. «Средняя оценка купленных пользователем товаров в 13,6 раза сильнее влияет на удовлетворенность пользователя, чем среднее время доставки заказов за последний год.»

   Комментарий. Данные, с которыми работает стажер, имеют разный масштаб. По гистограмме распределения возраста видно, что переменная среднего времени доставки заказа распределена примерно от 0 до 90, а переменная средней оценки заказов распределена от 1 до 5. Чтобы корректно делать вывод об относительной силе взаимосвязей каждой из двух переменных с третьей, необходимо предварительно отнормировать данные, то есть применить преобразование вида x_new_i = x_i — mean(x) / std(x).

3. «Среднее время доставки товаров за последний месяц не влияет на удовлетворенность клиента на маркетплейсе.»

   Комментарий. Так как стажер рассматривает только новых клиентов, две переменные (среднее время доставки за последний месяц и за последний год) для них одинаковые (с некоторой погрешностью в данных). Это следует из логики сбора данных для исследования, но также подтверждается, если посмотреть на коэффициент корреляции (он очень близок к 1, видимо, из-за погрешностей или особенностей сбора данных). Мы видим, что при добавлении только одной из этих переменных (orders_delivery_new) в модель, коэффициент при ней является статистически значимым. При добавлении сразу двух этих переменных мы сталкиваемся с проблемой частичной мультиколлинеарности (сильной линейной взаимосвязи между предикторами в модели). При частичной мультиколлинеарности увеличиваются стандартные ошибки коэффициентов, из-за чего мы получаем широкие доверительные интервалы для оценок коэффициентов. Следовательно, чаще делаем вывод о незначимости коэффициента. Исправить модель можно, исключив одну из этих переменных из фильной модели: никакой дополнительной информации из-за этого потеряно не будет, поскольку переменные по смыслу дублируют друг друга.

Задача 4 (20 баллов)

1. В этом пункте ставится по 3 балла за комментарий про каждый из трёх рядов. Необходимо учесть, что RMSE (из-за возведения остатков в квадрат) более чувствителен к выбросам, чем MAE и MAPE, а MAPE проще интерпретируется, что две другие метрики, но неустойчив, когда изучаемая переменная принимает околонулевые значения (вплоть до того, что может понадобиться делить остаток на ноль).

   Ряд A: подходят все три метрики, потому что ряд стабильный и далёк от нуля

   Ряд B: есть очень сильные выбросы, RMSE к ним чувствительнее, он явно выигрывает у MAE. MAPE для этого ряда можно использовать, потому что ряд далёк от нуля

   Ряд C: он большую часть времени нулевой, MAPE нельзя использовать, потому что знаменатель при расчёте MAPE может оказаться нулевым. Больших выбросов нет, MAE и RMSE оба подходят

2. Ошибка: R^2 имеет смысл и лежит в границах от 0 до 1 только для линейной регрессии с константой внутри выборки, для вневыборочного прогноза нелинейной моделью он не подходит. Более того, есть две разные формулы его расчёта (та, которой пользуются консультанты, и 1 — дисперсия ошибки, делённая на дисперсию факта), которые эквивалентны только для линейной регрессии с константой.

   Более того, в рассматриваемой ситуации можно домножить прогноз на константу, его дисперсия увеличится и R^2 станет ещё выше, что очевидно бессмысленно.

   За рассуждения про R^2: +2 балла за указание на то, что границы работают только для регрессии с константой in-sample, +2 балла за указание на то, что мы используем другую модель и оцениваем прогноз out-of-sample.

   Простая модель: наивный прогноз (реплицировать последнее наблюдение), прогноз средним по последним нескольким точкам, регрессия на линейный тренд и т.д. (Достаточно любого из этих вариантов). Рассуждения об оценке ещё более сложной модели или о лучшей настройке текущей модели в этом пункте не принимались, вопрос был именно в простом подходе!

3. Недочёт: прогноз суммарных продаж не равен сумме прогнозов отдельных товаров

   Простой способ исправить: либо прогноз суммарных пересчитать как сумму отдельных, либо отнормировать отдельные продажи, чтобы их сумма билась с прогнозом суммарных

   Сложный способ исправить: Что-то среднее между двумя «простыми» пунктами, взвешенное в зависимости от точности отдельных моделей (чем точнее — тем больше вес)

Задача 5 (20 баллов)

Имея количество закупленных товаров и реальный спрос в эти дни можно посчитать итоговую прибыль по формуле:

$$\sum_{day=1}^{7} \sum_{i=1}^{10} \min(Y_i, Ypred_i) \times revenue_i - Ypred_i \times cost_i$$

где Y_i — фактические спрос i-го товара, $Ypred_i$ — количество закупленного товара i, $revenue_i$ — доход с продажи одного товара i, $cost_i$ — цена закупки одной единицы товара.

Максимальная прибыль, которую можно получить сделав идеальный прогноз закупки, который повторит спрос, 122055 у.е. Итоговая оценка выставляется по формуле:

20баллов * (X-50000)/(122055-50000)

где X- итоговая прибыль

и округляется до целых по математическим правилам округления.

Фактический спрос на товары с 64го по 70й дни.

date	item1	item2	item3	item4	item5	item6	item7	item8	item9	item10
63	126	148	312	45	100	68	294	16	225	200
64	128	144	312	45	125	69	306	27	231	220
65	130	140	313	49	130	66	297	17	234	229
66	132	136	318	46	117	63	296	24	225	227
67	134	132	321	51	102	49	280	5	225	205
68	136	128	323	45	85	62	302	28	231	180
69	138	124	325	47	95	56	301	26	234	195

Разберем каждый ряд по отдельности:

1й ряд линейно растет каждый день на 2 единицы товара. Предсказываем продлевая данную линейную зависимость на оставшиеся дни.

2й ряд линейно падает на 4 единицы товара. Предсказываем продлевая данную линейную зависимость на оставшиеся дни.

3й ряд линейно растет каждый день на 3 единицы товара. Это можно выяснить применив сглаживание ряда. Шум сам себя нивелирует и останется только зависимость тренда +3 единицы товара. При этом вычтя тренд можно понять, что шум лежит в диапазоне ±3. Предсказываем продлевая данную линейную зависимость на оставшиеся дни ничего не делая с шумом.

4й ряд колеблется около значения 50 с интервалом ±5. Можно предсказать значением 50. Но оптимальнее будет взять значение 45, так как мы будем равномерно ошибаться в обе стороны, если оставим прогноз в 50 товаров. А закупка товара обходится куда дороже (150 у.е.), чем прибыль от его продажи (20 у.е.).

5й ряд имеет положительный тренд 1 и сезонность, которая напоминаю синусоидальные колебания, это можно увидеть вычтя тренд. А так же шум с интервалом ±5.

6й ряд имеет отрицательный тренд -2. А так же легко заметить недельную сезонность вида 10,18,13,12,0,15,13, убрав тренд из ряда. Так же получится, что есть шум ±3. Легко сделать прогноз учитывая тренд и сезонность и не обращать внимание на шум. Так как он

незначительный.

7й ряд имеет положительный тренд равный 1. Две сезонности одна недельная, а одна длинной 5 дней вида 10,18,13,12,0,15,13 и 2,3,-1,4,7 соответственно. Так же получится, что есть шум ±3. Легко сделать прогноз учитывая тренд и сезонность и не обращать внимание на шум. Так как он незначительный.

8й ряд ряд колеблется около значения 5 с шумом ±5 и имеет две сезонности одна недельная, а одна длинной 5 дней вида 10,18,13,12,0,15,13 и 2,3,-1,4,7 соответственно.

9й ряд состоит из двух частей: первая половина колеблется вокруг 190 с шумом ±5, вторая половина колеблется вокруг 230 с шумом ±5. Для верного прогноза стоит учитывать только вторую половину. Оптимальным будет предсказание 225 на каждый день по аналогии с рядом товара 4.

10й ряд. Для начала стоит найти выбросы в точках с индексами 15,47,54 значений выбросов в них 300, 312, 0 соответственно их можно заменить средним двух значений справа и слева от этих дней. Далее по аналогии с рядом 5 легко заметить, что ряд имеет положительную сезонность 1, синусоидальную сезонность и шум с интервалом ±10. Можно продлить ряд со значениям по низу шума, по аналогии с рядом товара 4.

В итоге получим следующее предсказания

date	item1	item2	item3	item4	item5	item6	item7	item8	item9	item10
63	126	148	309	45	105	66	289	14	225	195
64	128	144	312	45	122	72	298	26	225	219
65	130	140	315	45	126	65	295	15	225	226
66	132	136	318	45	117	62	293	15	225	211
67	134	132	321	45	100	48	283	0	225	186
68	136	128	324	45	90	61	301	19	225	171
69	138	124	327	45	95	57	298	20	225	178