Решение от Хоменко Ивана Михайловича\_Направление\_Data Science

Содержание

[Подготовка данных к анализу 2](#_Toc150712023)

[Анализ инфляции на примере себестоимости 2](#_Toc150712024)

[Анализ себестоимости и цены продажи 3](#_Toc150712025)

[Сравнение цен конкурентов и цен нашей компании 3](#_Toc150712026)

[Анализ влияния погоды на продажи 3](#_Toc150712027)

[Экспорт результата 4](#_Toc150712028)

[Небольшие комментарии и мысли по поводу задания 4](#_Toc150712029)

# Подготовка данных к анализу

Во время предобработки данных выявились следующие аномалии:

1. В данных о транзакциях были пустые значения в столбце ‘place’ в количестве 432 штук. Учитывая объем самой базы данных (875 036), было принято решение не пытаться восстанавливать данные, а избавиться от них.  
   Явных дубликатов выявлено не было.  
   Из аномалий были найдены отрицательные значения в столбцах ‘price’ и ‘amount’ *(Видимо, мурлоки решили подшутить и дописали несколько записей в учетную книгу предыдущего управляющего)*. Хоть это и значительный пласт данных, но удаление этих строк является целесообразным для корректного последующего анализа.
2. В данных с постоянными ценами (baseline) для удобства было принято решение исправить индексы.
3. В данных о конкурентах также были исправлены индексы.
4. В данных о погоде аномалий не было обнаружено.
5. В данных о себестоимости также аномалий не было выявлено.

Кроме того, были проверены уникальные значения в столбцах ‘product’, ‘place’, ‘competitors’ в тех датафреймах, где они присутствовали, чтобы убедиться, что в указанных столбцах нет опечаток или других проблем.

# Анализ инфляции на примере себестоимости

Была выдвинута гипотеза о том, что себестоимость товаров для продажи (эстуса, эльфийской пыльцы и целебных трав) в реальном выражении остается одинаковой на протяжении всего рассматриваемого периода *(что тебе в 2216 году нужно было бегать до костра, чтобы набрать фляжку эстуса, что в 2218. Иными словами, затраты на производство этого волшебного напитка в реальном выражении остаются прежними)*

Таким образом, исходя из выдвинутой гипотезы, можно предположить, что изменение цены товаров может быть обусловлено только инфляцией в каждом конкретном городе.

В этом же блоке была предпринята первая попытка написания модели для прогноза цены. Конечно, для полноценной модели проанализированных данных крайне мало, но, думаю, впоследствии можно было бы на основе этого спрогнозировать цены с учетом инфляции. Т.к. эта модель работает не до конца корректно, было принято решение оставить модель в виде «Raw NBConvert», чтобы она не создавала помехи при воспроизведении остального кода. *(Конечно, можно было просто удалить модель, но решил оставить в таком виде, чтобы Древние Боги видели, что я старался…)*

# Анализ себестоимости и цены продажи

Изначально было намерение проанализировать, какую наценку делал старый управляющий сетью «Главное Пристанище Надежды» на продаваемую продукцию.

После этого снова были предприняты попытки построения модели для прогнозирования с помощью регрессионного анализа. Но, опять же, данная модель не может рассматриваться как полноценная, т.к. здесь не учитываются изначальные условия задачи и другие предоставленные данные. Иными словами, это можно считать черновиком, который в будущем можно будет доработать.

# Сравнение цен конкурентов и цен нашей компании

В данном блоке хотелось сравнить исторические цены нашей компании и цены конкурентов.

Как оказалось, некоторые цены на товары могли почти в 2 раза превышать таковые у конкурентов. Например, Целебные травы в Анор Лондо. *(Похоже, теперь ясно, почему пропал предыдущий управляющий. Все-таки Древние Боги не дремлют)*

После сравнения цен была предпринята последняя попытка построения модели. Здесь я пытался уже включить в модель, как минимум, ограничение на превышение цен конкурентов. И все же, т.к. модель исправно не работает, было принято решение оставить ее в том же формате, что и первую попытку создания модели.

# Анализ влияния погоды на продажи

Для отображения корреляции между количеством продаж товаров и погодой были построены тепловые карты для каждого товара в каждом городе.

Как мы видим из графиков, между объемом продаж и погодными условиями наблюдается слабая прямая связь.

Хотя, возможно, некорректно сравнивать таким образом количественные и качественные характеристики…

В любом случае я не смог понять, как использовать эти данные для построения более точной модели.

# Экспорт результата

Т.к. в результате решения задачи я не пришел к какому-либо конкретному результату, то, по сути, экспортировать нечего. Но меня обуяли сомнения, вдруг из-за отсутствия расписания цен как одного из атрибутов предоставления работы мое решение и отчет не рассмотрят, а очень хотелось бы получить обратную связь.

Поэтому было принято решение экспортировать хоть какой-то результат, который получился в результате выполнения задания, хоть и он не удовлетворяет изначальным условиям задания.

# Небольшие комментарии и мысли по поводу задания

Хочу выразить огромную благодарность за возможность поучаствовать в подобном кейсе. В процессе его решения я узнал много для себя нового, больше углубился в машинное обучение в попытках построения моделей.

Во многом для меня это было как один большой эксперимент. Хотелось понять, насколько далеко я смогу продвинуться. И те часы, что я провел, корпя над этим заданиям, были очень приятными и, главное, результативными. Как минимум, узнал про такие модели, как ARIMA и SARIMAX.

Чувствую, что не хватает ещё некоторых знаний и практики, т.к. до окончательного ответа к заданию я так и не дошел. Но это более чем поправимо, и благодаря этому опыту стало ясно, где мои слабые места и что именно нужно развивать.

Если же говорить про то, что бы я сделал ещё, будь у меня больше свободного времени, то:

1. Предположу, что для более качественного анализа стоило бы проверить изначальные данные на выбросы, используя, например, boxplot, а также проверить данные на нормальность распределения
2. Более глубоко углубился в тему машинного обучения для построения полноценной модели. Или же по методу trial and error просто больше тыкался, пытаясь допилить модель, пока Древние Боги бы не ниспослали на меня озарение :)

Отдельно хочу отметить, что я был удивлен подобному оформлению задания. Берсерк, Baldur’s Gate, Diablo, Dark Souls, Elden Ring, World of Warcraft (хотя, по большей части, с этой вселенной я знаком через Hearthstone) – я безумно люблю эти франшизы, и, когда я увидел такой сеттинг в задании, гик внутри меня был в полном восторге. Поэтому в данном отчете я позволил себе вольность разбавить аналитику и выводы небольшими шуточными вставками, отмеченные курсивом и помещенные в скобки.

Подводя итог, хочется еще раз сказать, что это задание оставило у меня большое количество позитивных эмоций (Конечно, были моменты фрустрации, когда было непонятно, что делать, но post factum могу сказать, что мне понравилось). Спасибо за предоставленную возможность попробовать себя в подобном кейсе. Good luck!

P.S. Не понял, что именно нужно назвать в формате «Решение от \*ФИО\*\_Направление\_Data Science», поэтому назвал так всё, что можно…