Отчет ГПН КАП

(Бойко Дмитрий Алексеевич)

Первичная обработка данных

Изначально данные о транзакциях были представлены в виде датасета с датой и временем продажи. Некоторые строки содержали пропуски, но их количество было пренебрежимо мало по сравнению с общим числом данных поэтому было решено просто их вычеркнуть. Промежутки между наблюдениями были не постоянными по длине, что затрудняет анализ временного ряда и, вообще, информации было слишком много. Было решено привести данные к виду, более удобному для изучения.

В связи с этим был совершен переход от общего датасета для всех видов товаров и мест и поминутными данными к отдельным датасетам для каждой пары товар-место с данными по дням. Цена за день высчитывалась как среднее арифметическое цен в этот день на данный товар в данном месте, а объем продаж – как сумма объемов продаж за день. При подсчете цены и объема продаж не учитывались строки таблицы, в которых встречались значения nan, что позволило довольно просто и безболезненно избавиться от пропусков в таблице.

Также в новые датасеты для каждой пары товар-место была добавлена информация о ценах конкурентов. Каждого дня из нового датасета находилось среднее арифметическое цен конкурентов в этот день, и таким образом, можно считать, что была учтена средняя по рынку цена в этот день. Стоит заметить, что данные о цене конкурентов были даны не для всех троек день-товар-место, поэтому в новых датасетах в некоторых строчках на позиции средняя цены у конкурентов образовались пропуски.

Также для каждого дня нового датасета была добавлена информация о погоде, которая была в этот день.

Наконец, для каждого дня были рассчитаны затраты на производство единицы продукции путем деления общих затрат в этот день на объем продаж. Стоит заметить, что данные о затратах были даны не для всех дней, и поэтому было решено заменить пропуски на медианные значения.

Пример датасета для товара «Целебные травы» и места «Анор Лондо» представлена на скриншоте ниже.

Изображение выглядит как текст, снимок экрана, число, Шрифт

Автоматически созданное описание

Прогнозирование

Было сделано предположение, что спрос на товар в конкретный день зависит от цены на товар нашей компании, погодных условий и цен конкурентов. Также предполагалось, что во временном ряду могли присутствовать тренд и сезонность. Пример временного ряда для спроса на товар «Целебные травы» в «Анор Лондо» представлены на скриншоте ниже.

Изображение выглядит как снимок экрана, График, линия

Автоматически созданное описание

Исходя из графика можно предположить, что наблюдается сезонность с периодом в примерно 500 дней. Также, возможно, присутствует небольшой восходящий тренд.

В качестве модели для прогнозирования спроса была выбрана линейная регрессия исходя из двух причин. Во-первых, линейные модели просты и довольно эффективны в задачах прогнозирования. Во-вторых, спрос часто линейно зависит от цены. Если построить линейную модель с несколькими факторами, включая цену, а потом предсказать все факторы, кроме цены, то можно получить линейную функцию спроса от цены, которую можно использовать при формулировке задачи квадратичного программирования с квадратичной целевой функцией и линейными ограничениями.

Для моделирования сезонных колебаний были использованы первые гармоники ряда Фурье. Позже было решено отказаться от синуса в виду малой значимости фактора. Чтобы учесть тренд, была введена переменная равная номеру наблюдения во временном ряду. Как уже упоминалось ранее, столбец «Средняя цена у конкурентов» содержит пропуски. При попытки построить модель с данным фактором выяснилось, что он оказывает очень малое влияние, поэтому от него было решено избавиться. Можно предположить, что покупатели мало ориентируются при выборе товаров на цену у конкурентов, потому что их больше интересует абсолютная цена на товар и, возможно, у нашей компании также хорошая репутация. Таким образом, удалось избавиться от пропусков в датасете и от лишнего фактора. Погодные условия выступили в качестве категориальных переменных.

Было замечено, что временной ряд спроса ведет себя похожим образом для всех товаров и мест, поэтому было решено построить 15 моделей линейной регрессии. Чтобы оценить качество моделей, для каждой из них была проведена кросс-валидация по метрике MAPE и взято среднее значение. Ниже представлены результаты кросс валидации для всех моделей. Для большинства моделей MAPE лежит в районе 7–12%, что является неплохим результатом.

Изображение выглядит как текст, Шрифт, снимок экрана, черно-белый

Автоматически созданное описание

Хуже результаты показали себя модели для товара «Эльфийская пыльца», особенно в месте «Врата Балдора», где средняя MAPE достигла 77%. На скриншоте ниже можно видеть фактический временной ряд и его аппроксимацию. Низкий показатель MAPE можно объяснить очень большой амплитудой колебаний, и их необычным поведением. Несмотря на низкий показатель MAPE было решено оставить построенные модели.

Изображение выглядит как График, снимок экрана, линия, диаграмма

Автоматически созданное описание

В построенных моделях погодные условия являются важными факторами, поэтому для прогнозирования спроса необходимо знание погоды в соответствующие дни. Погоду предсказывать не совсем понятно как, поэтому было решено в качестве погодных данных за все дни взять постоянную величину – доля наблюдений, в которых дынный погодный фактор принимал значение 1.

Еще одной важной величиной, прогноз которой необходим – затраты на производство единицы продукции. Их важно знать, так как от них зависит прибыль. Пример временного ряда для затрат на производство единицы продукции для товара «Целебные травы» и места «Анор Лондо» предствален на скриншоте ниже. Ряд напоминает стационарный. Похожая картина наблюдается и для других пар товар-место. Было решено проверить временные ряды на стационарность при помощи критерия Дики-Фуллера.

Изображение выглядит как текст, снимок экрана, линия, диаграмма

Автоматически созданное описание

При проведении теста Дики-Фуллера для всех датасетов p\_value оказалось ниже 0.05 (что можно видеть на скриншоте ниже), из чего можно сделать вывод, что следует отвергнуть гипотезу о нестационарности для все рядов. В виду стационарности, для упрощения прогноза затраты было решено положить затраты постоянными равными медиане.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Шрифт

Автоматически созданное описание

Последний фактор, который необходимо прогнозировать – это средняя цена конкурентов. Несмотря на то, что данный фактор был исключен из моделей линейной регрессии, он все равно необходим нам, так как нам нужно держать цену на уровне, не превышающем цену конкурентов более, чем на 20%. Для прогнозирования средней цены по рынку была выбрана модель Prophet из одноименной библиотеки. Плюс данной модели не только в довольно точном прогнозе, но и в том, что помимо основного прогноза она предоставляет еще и верхнюю и нижнюю границу прогноза, своеобразный доверительный интервал. Было решено в качестве прогноза цены конкурентов взять среднее арифметическое основного прогноза и нижней границы, чтобы перестраховаться от получения большого штрафа за сильное превышение средней цены по рынку, и в то же время избежать слишком сильного занижения цены. Пример прогноза по модели prophet для товара «Целебные травы» и места «Анор Лондо» представлен ниже.

Изображение выглядит как линия, График, текст, диаграмма

Автоматически созданное описание

Формулировка задачи квадратичного программирования.

Для того, чтобы задачу оптимизации можно было решить при помощи солвера, она была сформулирована на языке квадратичного программирования.

Изображение выглядит как текст, рукописный текст, Шрифт, документ

Автоматически созданное описание

Изображение выглядит как текст, рукописный текст, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, рукописный текст, Шрифт, документ

Автоматически созданное описание

Изображение выглядит как текст, рукописный текст, Шрифт, документ

Автоматически созданное описание

Стоит отметить, что наша задача содержит очень много бинарных переменных (порядка 90 \* 4 штук) что может сделать вычисления очень долгими по времени. Для этого было решено решать данную задачу в несколько этапов, а именно решить 9 задач для 10 дней в каждой. То есть сначала построить прогноз с 1 по 10 день, затем с 11 по 20 и т. д. до 90. У всех подзадач ограничения остаются такими же, как и у исходной задачи, однако в подзадачах, начиная со второй (прогнозирование с 11 дня по 20) стоит учесть один момент. Он состоит в том, что первые 4 переменные не обязательно должны быть одинаковыми, т. е., например цены с 11 по 14 день могут отличаться друг от друга, т. к. цена в день 11 могла быть такой же как в дни с 8 по 10 и тогда цена в 12 день может отличаться от цены в 11. Чтобы это учесть, необходимо добавить в подзадачу задачу переменные за последние 3 дня предыдущей подзадачи, задать им найденные значения и добавить ограничения, связанные с ними для первых 4 дней текущей подзадачи.

Реализация точной постановки и нахождение решения для нашей задачи было проведено в компьютерной среде Wolfram Mathematica при помощи встроенной функции для решения задачи квадратичного программирования QuadraticOptimization. В качестве входных данных послужили записанные нами ранее экселевские данные с датасетами.

Результаты сравнения ожидаемой прибыли по плану, составленному нами, и по бейзлайну представлены ниже. Всего в двух случаях прибыль по бейзлайну оказалась выше (в одном из них всего на 0.8%), но при этом в определенные дни бейзлайновая цена была больше, чем на 20% выше ожидаемой цены конкурентов. В остальных случаях в среднем наше решение лучше базового на 10.5%.

Изображение выглядит как текст, число, Шрифт, снимок экрана

Автоматически созданное описание