Для построения моделей была проведена работа по предобработке данных, так как среди наблюдений по транзакциями были такие, где количество проданных товаров было меньше 0, цена по которым была отрицательной (так как компания является продавцом, был сделан вывод о том, что строки, содержащие данные значения, являются ошибками в сборе данных и должны быть удалены), таким образом было очищено 219060 наблюдений, содержащих шумовую информацию. В датасетах по ценам конкурентам и издержкам таких наблюдений обнаружено не было, а так как в датасете о погоде сумма по столбцам не всегда была равна 1, был сделан вывод о существовании 4 состояния погоды, поэтому нет необходимости удалять один из столбцов, чтобы избежать dummy-trap

Датасет по транзакциям был сгруппирован таким образом, чтобы он содержал информацию о цене продажи, выручке и суммарному объему продаж за день (а не за отдельную транзакцию). Таким образом, в данном датасете содержится 14998 наблюдений, то есть практически на каждый товар и место приходится 1000 дней, содержащих информацию о показателях.

Далее было проведено объединение датасетов по индексу в один общий, например, для погоды показателями, однозначно описывающими одно конкретное наблюдение, являются место и дата, поэтому объединение итогового датасета с ним проводилось по индексу, содержащему данную информацию. Информация об издержках и ценах конкурентов также однозначно определяется, но уже трем показателями: датой, товаром и местом.

После объединения в итоговом датасете оказались пропуски, так как для некоторых дат отсутствовала информация по издержкам и ценам конкурентов, эти пропуски были последовательно заполнены сначала предыдущими значениями, а потом последующими. Так как исходя из специфики условий установления цены, которая может меняться не чаще, чем раз в три дня, можно предположить, что в случае пропуска, информация о соседних датах будет достаточно точно описывать цену. Издержки были заполнены таким же образом из-за предположения, что в исходном датасете содержалась информация только об изменении издержек, а если для даты не указаны издержки, то в данный период они были такими же, как и в последний указанный период.

Заполнение пропусков происходило отдельно для каждого места и товара, так как информация по ним является независимой и не должна перетекать друг в друга, в дальнейшем этот датасет был разделен на 15, содержащих информацию о каждом товаре и месте, работа происходит отдельно по ним.

По показателям, содержащимся в датасетах был сделан вывод о существовании явных трендов и сезонности в части из них (прибыль, выручка, издержки) и тренда в ценах, поэтому было бы оправдано использование модели ARIMA для прогнозирования, но из-за недостатка знаний в области применения этой модели, было принято решение использовать трендовую модель, являющую собой многочлен 4 степени в зависимости от индекса наблюдения (полином 4 степени выбран из-за идеи, что данная степень позволит отследить часть сезонность), то есть произошло упрощение, которое можно было бы исправить и получить более точные прогнозы в случае большего объема времени.

Была составлена функция, которая принимала на вход датасет, содержащий информацию по отдельному товару и месту, и итоговый датасет, в который записывался прогноз. Идея данного моделирования заключалась в следующем: сначала с помощью уже названной трендовой модели во временный датасет записывались прогнозы на следующие 90 периодов по таким показателям как количество, издержки и цена конкурентов (трендовая модель обучалась на 1000 наблюдений), далее была обучена модель случайного леса (так как она показала наилучший результат на тестовой выборке в сравнении с линейной и ridge регрессиями), в которой на основании погоды и указанных ранее факторов предсказывался уровень цены, выставленный фирмой. На основании полученной значимости для факторов был составлен прогноз цены на следующие 90 периодов, в качестве объясняющих переменных использовались ранее предсказанные на 90 периодов вперед показатели, при этом для прогнозирования погодных условий использовалось случайное заполнение с учетом их частоты на горизонте в 1000 наблюдений.

Таким образом, был получен прогноз цены по 5 отдельным местам и 3 товарам, который в значительной степени соответствует требованию о том, что цена должна изменяться не чаще, чем раз в три дня, и полностью – требованию о том, что единоразовый рост цены не должен превышать 1 золотой, но моделирование можно улучшить, если использовать модель ARIMA.