**НЕТОЛОГИЯ**

**Дипломный проект**

**«** Прогноз динамики продаж товаров бытовой химии

и эффективности проведения маркетинговых акций **»**

#### Работу выполнила : Гупта Ирина

Группа : Data Science (DS-16)

Москва, 2021

**Постановка задачи**

Компания-производитель и дистрибьютер товаров бытовой химии уже более 15 лет занимается поставками на российский рынок освежителей воздуха, чистящих и моющих средств для санузла и кухни, средств для мытья полов и проч. В данный момент ассортимент продукции насчитывает около 50 наименований.

В последнее время на рынке кроме прямых конкурентов-производителей в сегменте бытовой химии появились собственные торговые марки (СТМ) больших федеральных сетей. Как следствие, продажи падают, т.к. сети исключают из своего ассортимента продукцию сторонних производителей, выставляя на полках в основном СТМ.

Поэтому компании нужно более тщательно планировать работу с клиентами. Современные методы аналитики могут помочь предсказать тренды и предупредить заранее и возможном падении продаж и соответственно порекомендовать проведение маркетинговых акций.

Однако на данный момент какая-либо аналитика в компании отсутствует. Эта работа – первая попытка построения хоть какой-то аналитики с нуля. Поэтому появилась задача в данной работе рассмотреть динамику продаж и эффективность маркетинговых акций с использованием machine learning.

В рамках этого дипломного проекта будут рассмотрены две задачи:

1. Прогноз динамики продаж товаров бытовой химии (прогнозирование временного ряда).
2. Прогноз эффективности проведения маркетинговых акций с целью повышения уровня продаж (задача классификации).

**Анализ**

Существует много моделей и подходов к прогнозированию временных рядов и задачам классификации.

Прогнозирование временных рядов относится к тому типу задач, в которых мы должны предсказывать результат на основе зависящих от времени входных данных.

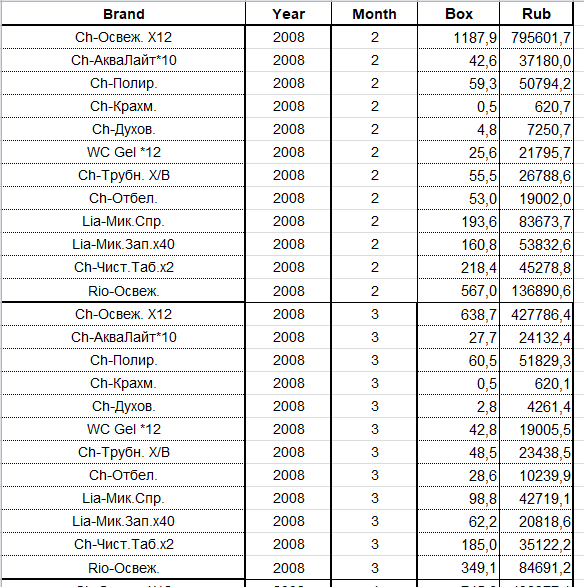
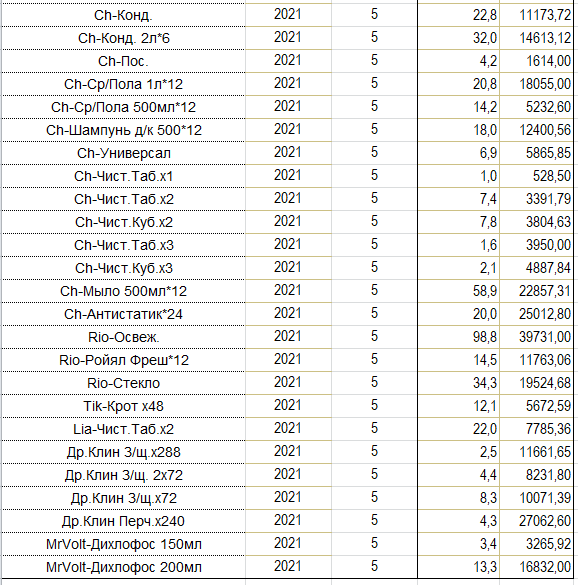
Прежде, чем приступить к решению, можно ознакомиться с аналогичными задачами на просторах интернета. Например, есть [подробно описаное решение](https://habr.com/ru/company/ozontech/blog/431950/) по прогнозу спроса OZON, где обучили 130 моделей с очень хорошим в итоге результатом средней абсолютной ошибки MAE = 1 на сайте <https://habr.com/>.

Решение задач последовательности [с помощью LSTM](https://pythobyte.com/solving-sequence-problems-with-lstm-in-keras-9348798b/) хорошо описано на сайте <https://pythobyte.com/>.

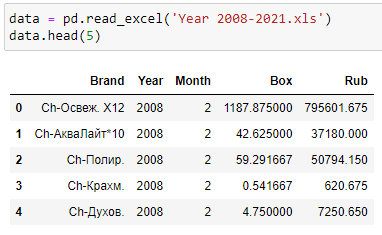
А задача классификации представляет собой одну из базовых задач прикладной статистики и машинного обучения, а также искусственного интеллекта в целом. Это связано с тем, что классификация является одной из наиболее понятных и простых для интерпретации технологий анализа данных, а классифицирующие правила могут быть сформулированы на естественном языке. Примеров решения таких задач в интернете множество. Главное, что для получения эффективного решения необходимо качественно подготовить входные данные.

Для начала исследуем полученные данные и проведем преобразования, необходимые для применения моделей и методов, а затем приступим к решению поставленных задач.

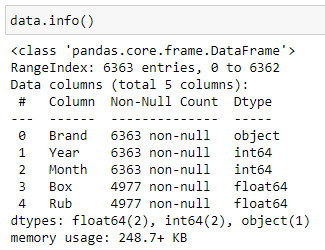
Исходные данные в виде Excel таблицы содержат информацию по закупкам одного из ключевых клиентов компании за каждый месяц в период с февраля 2008 года по май 2021 года.

Данных немного, образованный из них датасет имеет 6363 записи и 5 колонок:

Посмотрим информацию о датасете:



Brand - наименование товара, тип object

Year - отчетный год, тип int64

Month - отчетный месяц, тип int64

Box - объем продаж в коробках за данный месяц, тип float64

Rub - сумма продаж в рублях за данный месяц, тип float64

В последних двух столбцах встречаются пропущенные значения, это значит, что в эти месяцы клиент не закупал конкретный товар, следовательно, все пропущенные значения заполняем нулями.

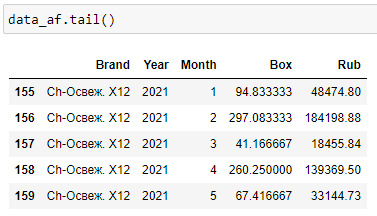
Т.к. за десять лет уровень цен много раз менялся, дальнейший анализ и прогноз будем строить на объеме продаж в коробках.

**Методика решения**

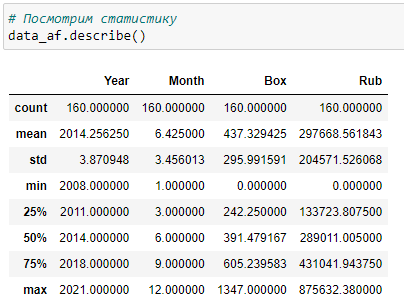
***1. Первая задача - прогноз продаж (на основе анализа временного ряда).***

* 1. *Подготовка данных*

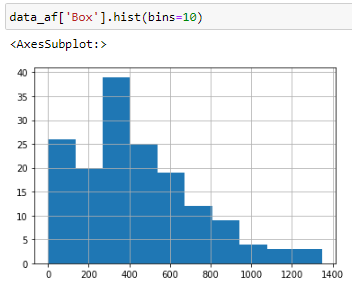
Для работы возьмем один ключевой товар – Освежитель воздуха (Ch-Освеж).



По освежителю воздуха имеем 160 записей.



Отрицательных значений нет.

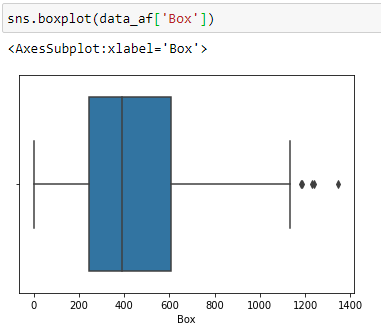


В среднем за месяц продается

около 400 коробок освежителя.

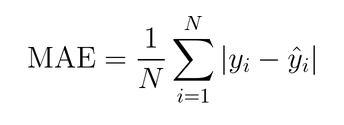
Минимально – 0 шт.

Максимально – 1347 шт.

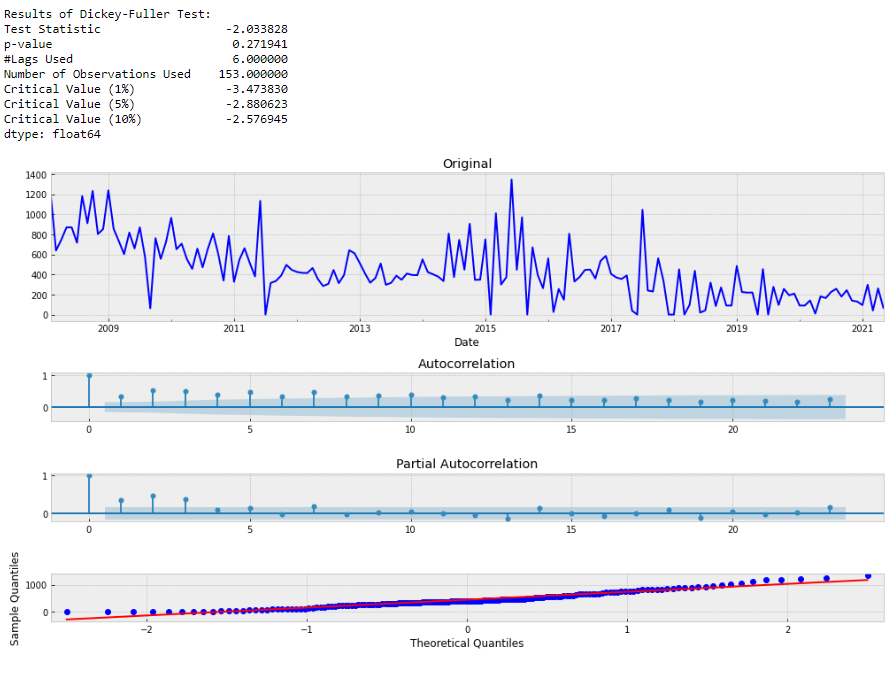


Видим выбросы значений, но данных у нас и так немного, удалять не будем.

Метрикой точности полученных моделей будет Средняя Абсолютная Ошибка (MAE), т.к. она более устойчива (менее чувствительна к выбросам).

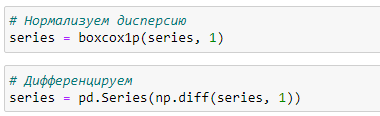


Проанализируем полученный временной ряд продаж.

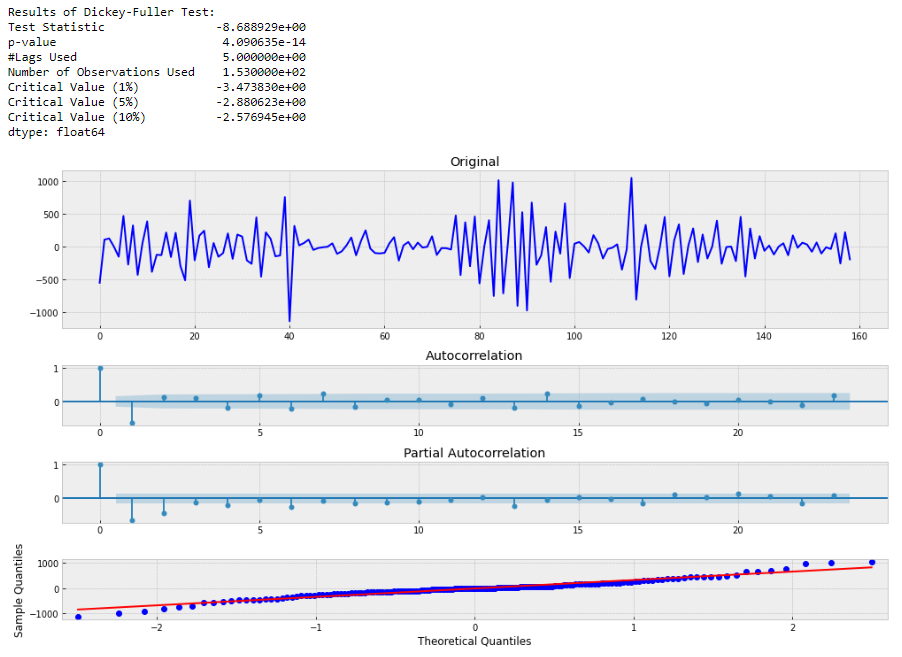


Тест Дики-Фуллера, показывает, что P-value = 0.27, это больше 0.05, значит, полученный временной ряд не стационарен и требуется применить преобразование.

После нормализации дисперсии методом Бокса-Кокса и дифференцирования



получаем стационарный временной ряд, к которому можно применять методы анализа.



При решении этой задачи будем сравнивать результаты модели SARIMA, рекуррентной нейронной сети LSTM и модели GARCH.

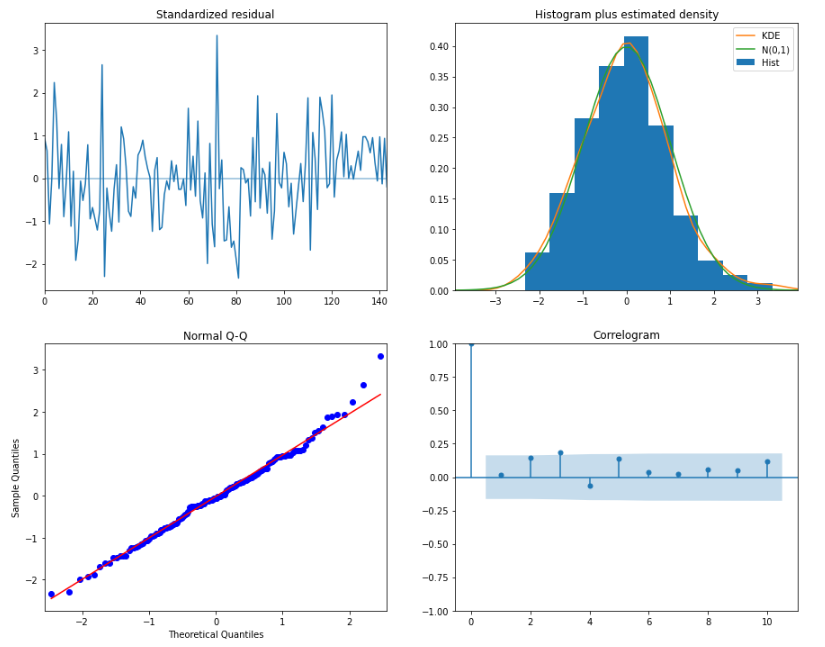
* 1. *Модель SARIMA (метод сезонного авторегрессионного интегрированного скользящего среднего)*

Параметры модели подбираем в соответствии с критерием Акаике (AIC). По графику автокорреляции временного ряда наблюдается некая сезонность в 7 периодов, поэтому сезонный компонент модели берем равный семи. По мнению отдела продаж, сезонность освежителей воздуха – полгода, к лету продажи растут, к зиме – падают. При решении были проверены так же периоды 6 и 12, но при 7 были получены наилучшие результаты.

Окончательные оптимальные параметры модели: (1, 1, 2), (0, 2, 2, 7)

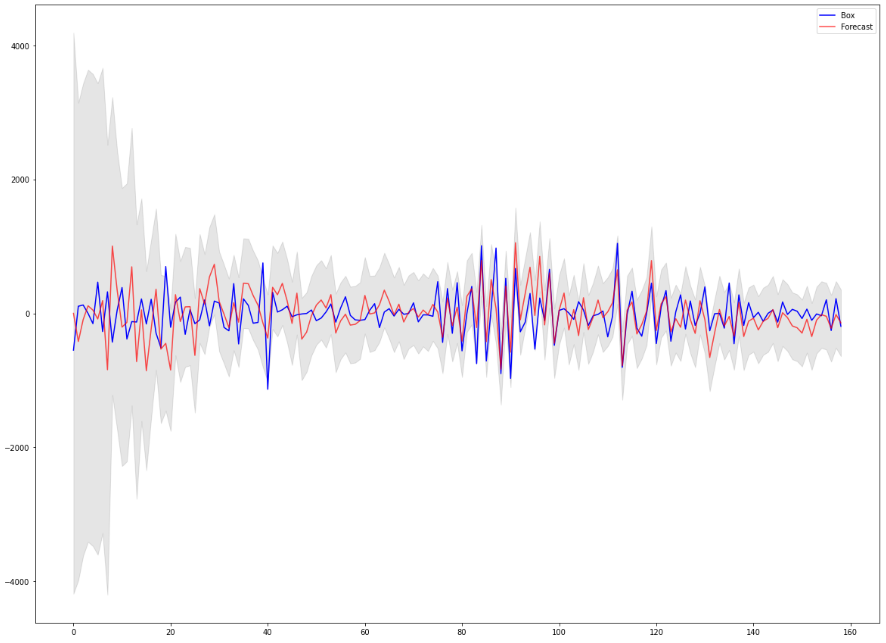


Посмотрим на распределение остатков модели.



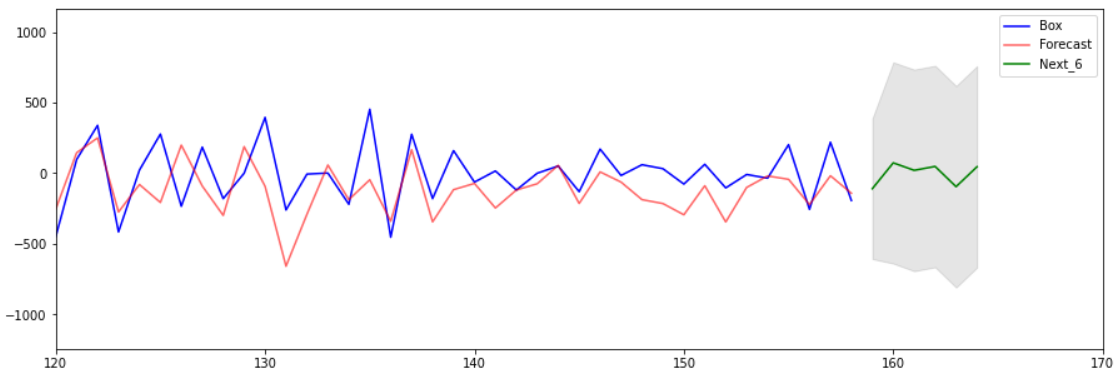
Диагностика показывает, что остатки близки к нормальному распределению, значит, модель сможет понять данные и спрогнозировать будущие значения.

По графику реальных и спрогнозированных значений видно, что предсказанные значения довольно далеки от реальных.



Средняя Абсолютная Ошибка MAE = 261.77

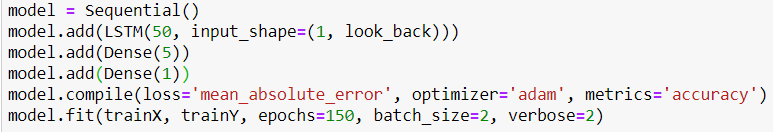
Используем модель SARIMA для прогнозирования на 6 значений вперед

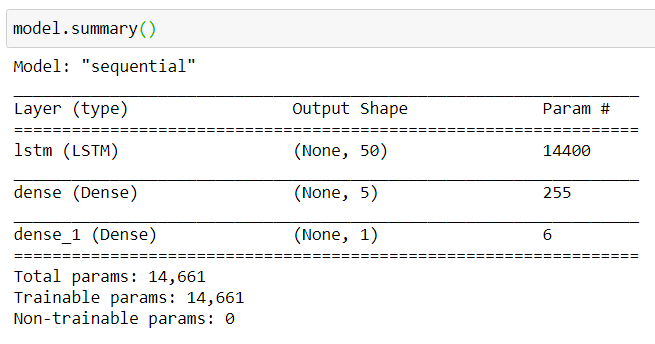


* 1. *Рекуррентная нейронная сеть LSTM (Долгая краткосрочная память)*

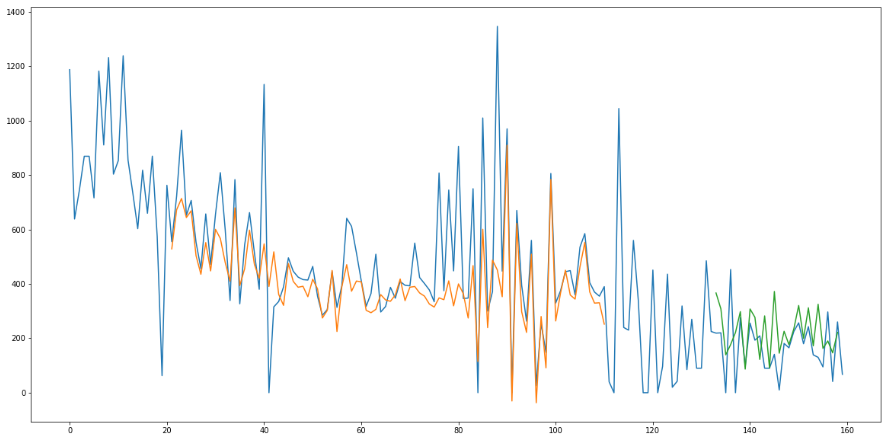
Данные нормализуем и делим на обучающую (70%) и тестовую (30%) выборки.

Создаем и обучаем следующую LSTM (50 нейронов):





Период взгляда назад (look\_back) берем равным 21. При решении были опробованы так же периоды 7 и 14, но при 21 были получены наилучшие результаты.



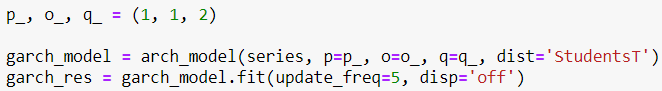
Train Score: Средняя Абсолютная Ошибка MAE = 89.88

Test Score: Средняя Абсолютная Ошибка MAE = 109.45

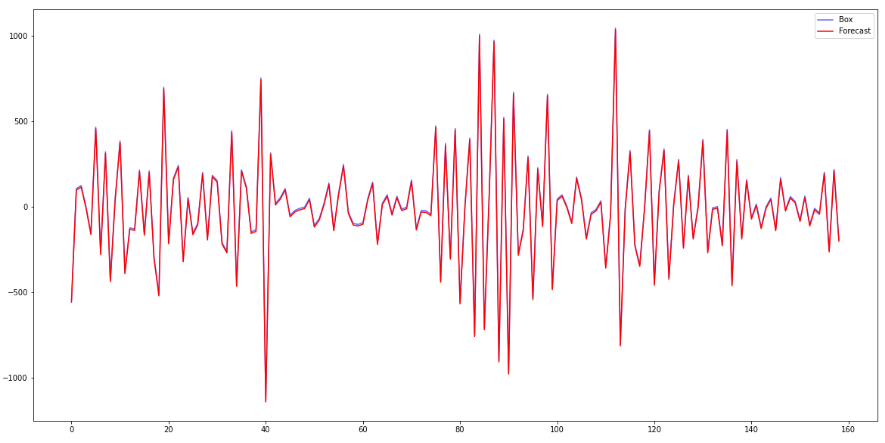
Эти показатели уже лучше, чем у модели SARIMA.

* 1. *Модель GARCH (метод авторегрессионной условной гетероскедастичности)*

Коэффициенты для модели возьмем из оптимальных параметров SARIMA (1, 1, 2). Метод распределения ошибок - T-распределение Стьюдента.



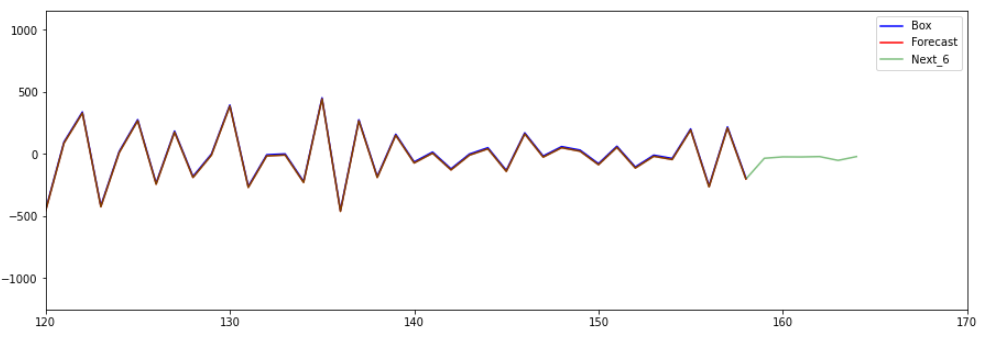
По графику реальных и спрогнозированных значений видно, что предсказанные значения практически повторяют реальные.



Средняя Абсолютная Ошибка MAE = 9.01

Модель GARCH показала наилучшие результаты с минимальной средней абсолютной ошибкой.

Используем модель GARCH для прогнозирования на 6 значений вперед



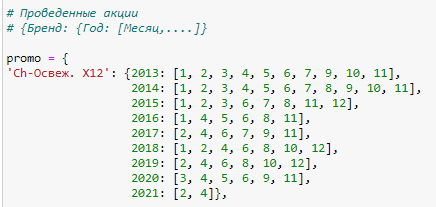
1. ***Вторая задача – прогноз эффективности проведения маркетинговых акций.***

*2.1. Подготовка данных*

В этой работе не рассматривались отдельные виды акций (% скидки, 1+1, 10+1, 12+1, один товар + другой товар того же бренда и т.д.), фиксировался только факт проведения любой из акций в конкретном месяце.

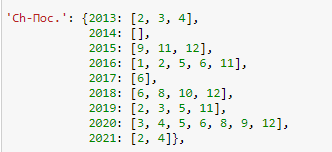
Решения об объявлении акций носили несистематический характер, обусловленный различными причинами (наличие товара, его сезонность, срок годности и прочее). Поэтому иногда складывалась ситуация, что акция проведена, цены снижены, а эффекта на уровень продаж она не имела. Компания заинтересована в том, чтобы с помощью автоматической аналитики можно было спрогнозировать, спланировать необходимость проведения акций или, наоборот, понять, что эффективности акция на конкретный товар в конкретном месяце не принесет.

Например, на освежитель воздуха довольно часто были акции:

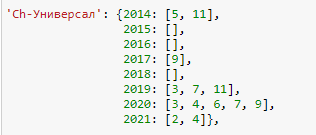


А на другую продукцию гораздо реже:

Моющее средство для посуды:



Универсальное чистящее средство:



Информация по проведенным маркетинговым акциям предоставлена только начиная с 2013 года. Поэтому для решения этой задачи, удаляем из датасета все предыдущие года, остается 4854 записи.

Добавим новые признаки:

***Incr.Box*** - Увеличение/снижение объема продаж в коробках относительно предыдущего месяца

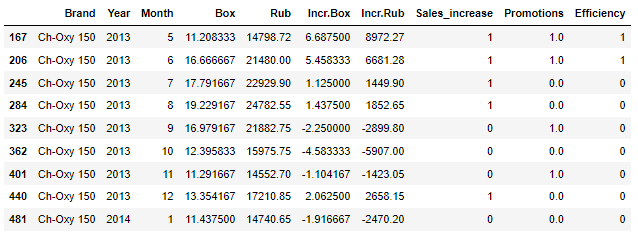
***Incr.Rub*** - Увеличение/снижение суммы продаж в рублях относительно предыдущего месяца

***Sales\_increase*** - Было ли увеличение объема, либо суммы относительно предыдущего месяца (1/0)

***Promotions*** - Проводилась ли акция на этот товар в этом месяце (1-да / 0-нет)

***Efficiency*** - Был ли эффект от акции (1-увеличился объем, либо сумма продаж при акции / 0-нет)

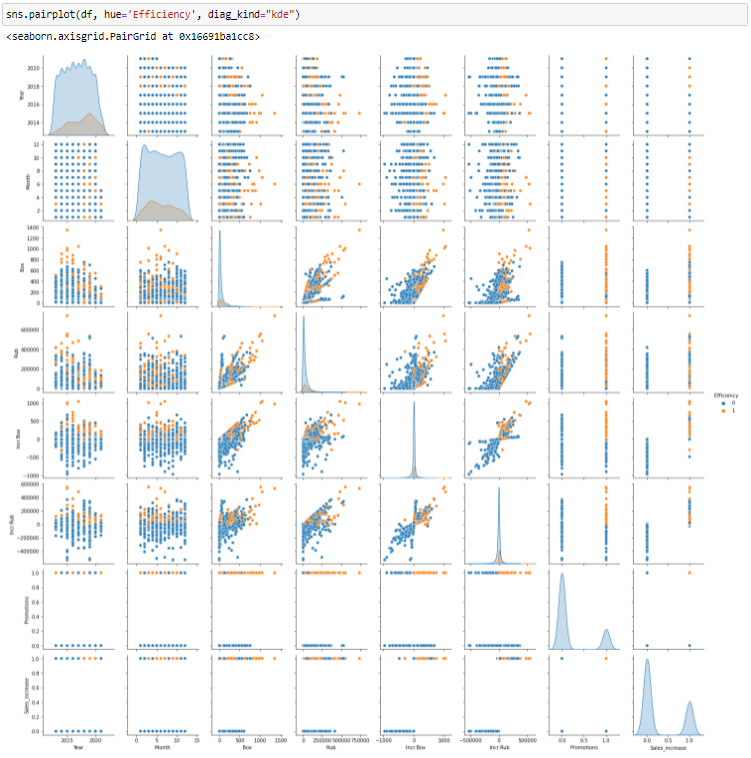
Например, для освежителя воздуха получаем следующие значения:



В итоге получается, что с 2013 года:

* были проведены 1692 различные акции на различную продукцию
* но только 860 из них имели эффект увеличения продаж
* при 832 акциях повышения продаж не было (а это 49% от всех проведенных акций!)
* в то же время без проведения акций продажи росли в 1218 случаях
* падали продажи без проведения акций в 1944 случаях

Попарное распределение фичей датасета:



По изображению видим, что некоторая зависимость признаков есть, значит, построенная модель сможет их выявить и показать результаты.

Столбец “Brand” является категориальной переменной, поэтому прежде, чем приступить к построению моделей, переведем ее в значения 0 и 1, добавив столбцы с соответствующими названиями.



Эффективность акции (*Efficiency)*будет таргетом. Построенная модель должна научиться предсказывать её значение  для более эффективного планирования бизнес-процесса.

Метрикой оценки качества работы моделей классификации будет F1 мера (гармоническое среднее между точностью и полнотой).

Все данные разделим на обучающую (70%) и тестовую (30%) выборки.

*2.2. Прогноз эффективности акций (LogisticRegression, DecisionTreeClassifier, StackingClassifier).*

Для решения задачи классификации были построены и обучены три модели:

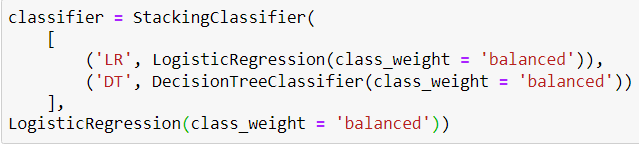
* Логистическая регрессия



* Дерево решений



* Ансамбль из этих моделей (стекинг)



Для сравнения результаты всех моделей представлены в таблице.

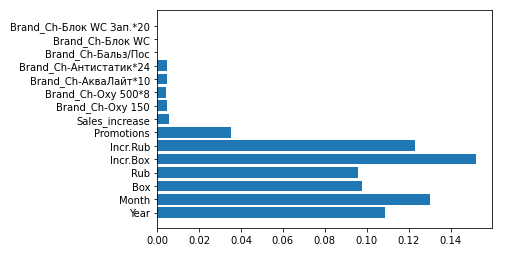
|  |  |  |  |
| --- | --- | --- | --- |
|  | *LogisticRegression* | *DecisionTreeClassifier* | *StackingClassifier* |
| Train score | 0.6603 | 0.8219 | 0.9514 |
| Test score | 0.6648 | 0.6944 | 0.7569 |
| Матрица ошибок |  |  |  |
| F1-мера | 0.4352 | 0.3995 | 0.3980 |
| Roc-Auc кривая |  |  |  |
| Roc-Auc  score | 0.7744 | 0.6719 | 0.6968 |

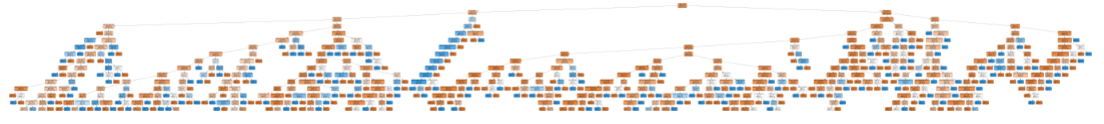
Все модели не очень хорошо обучились, однако у всех Roc-Auc score >0.5, значит предсказания моделей не случайны. F1-мера лучше других у LogisticRegression.

*2.3. Визуализация дерева решений.*

Для понимания процесса и оценки важности признаков визуализируем Дерево Решений.

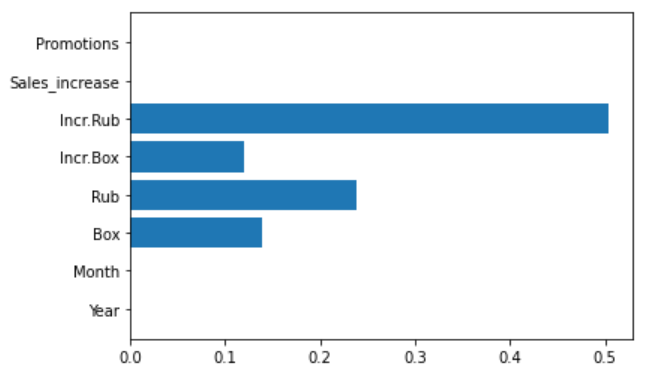
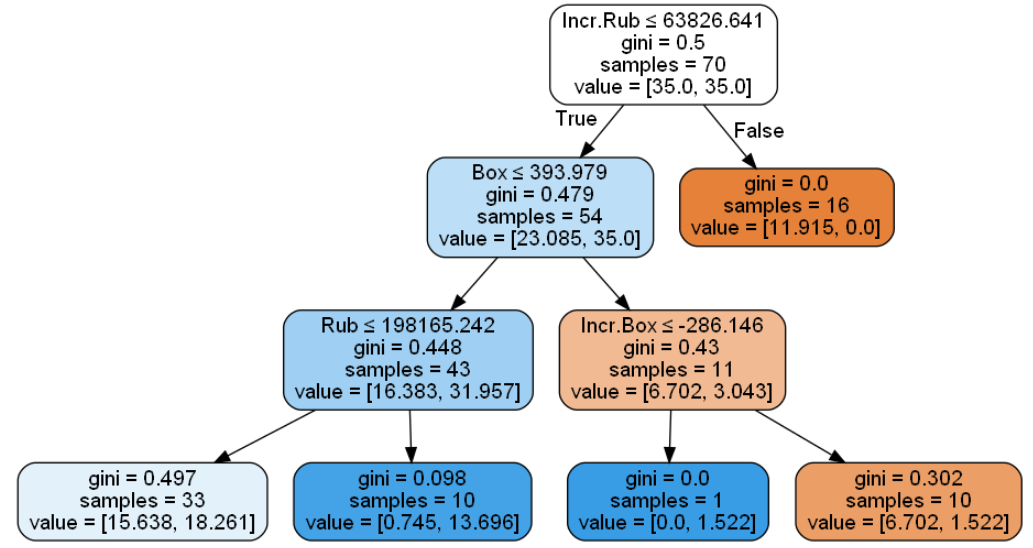
Важность признаков:





Для наглядности визуализируем часть дерева для одного бренда - Освежителя воздуха, глубиной = 3.

Важность признаков:



В основе алгоритма построения дерева решений лежит принцип жадной максимизации прироста информации – на каждом шаге выбирается тот признак, при разделении по которому прирост информации оказывается наибольшим.

Максимизацию критерия Неопределенности Джини можно интерпретировать как максимизацию числа пар объектов одного класса, оказавшихся в одном поддереве.

Видим, что в дереве задействованы сначала разбиение по возрастанию продаж в рублях (порог = 63826.641 руб.), потом по объему коробок за месяц (порог = 393.927 шт.), а затем по сумме продаж в рублях (порог = 198165.242 руб.) и возрастанию объема в коробках (порог = -286.146 шт.). А почему в дереве оказались именно эти признаки? Потому что по ним разбиения оказались лучше (по критерию Неопределенности Джини). Самая простая эвристика для обработки количественных признаков в дереве решений: количественный признак сортируется по возрастанию, и в дереве проверяются только те пороги, при которых целевой признак меняет значение.

Посчитав таким образом прирост информации для каждого разбиения, можно отобрать пороги, с которыми будет сравниваться каждый количественный признак.

В принципе дерево решений можно построить до такой глубины, чтоб в каждом листе был ровно один объект. Но на практике это не делается из-за того, что такое дерево будет переобученным – оно слишком настроится на обучающую выборку и будет плохо работать на прогноз на новых данных. Где-то внизу дерева, на большой глубине будут появляться разбиения по менее важным признакам.

**Выводы и заключение**

В данной дипломной работе было решено две задачи:

1. Прогноз динамики продаж товаров бытовой химии (прогнозирование временного ряда).
2. Прогноз эффективности проведения маркетинговых акций с целью повышения уровня продаж (задача классификации).

Задачи решались на основе датасета с информацией по закупкам одного из ключевых клиентов компании-производителя бытовой химии за каждый месяц в период с февраля 2008 года по май 2021 года.

Для решения первой задачи была выбрана одна ключевая продукция компании – освежитель воздуха. Для прогнозирования временного ряда были применены три разных метода : SARIMA, LSTM, GARCH. Метрикой точности полученных моделей была выбрана Средняя Абсолютная Ошибка (MAE).

SARIMA 🡪 MAE = 261.77, LSTM 🡪 MAE = 94.33, GARCH 🡪 MAE = 9.01

Наилучший результат MAE показал метод авторегрессионной условной гетероскедастичности GARCH.

По другой продукции компании имеющиеся данные более разрозненные, требуется дальнейший тщательный сбор и обработка.

Для решения второй задачи были построены три модели классификации: LogisticRegression, DecisionTreeClassifier и ансамбль из них StackingClassifier. Метрикой точности полученных моделей была выбрана F1 мера.

LR 🡪 F1 = 0.4352, DTClassifier 🡪 F1 = 0.3995, StackingClassifier 🡪 F1 = 0.3980

Наилучший результат F1 = 0.4352 показала LogisticRegression.

Для наглядности, понимания процесса расщепления и оценки важности признаков в работе было визуализировано Дерево Решений. Но существенных инсайдов в рассмотрении верхних узлов дерева не получено.

Хотя ни одна из моделей не показала высокого качества, полученные данные могут быть использованы для прогнозирования. В дальнейшем возможно улучшить результаты каждой из моделей, если еще тщательнее подобрать параметры к каждой из них. В задачи классификации можно попробовать добавить новых фичей, например, разбить товары бытовой химии по категориям: для кухни, для ванной, для комнат.

Так же для улучшения прогноза планируется дальнейший, более широкий сбор данных, исследование, выявление интересных зависимостей, анализ бизнес-процессов. Построенные модели хорошего качества дадут положительный бизнес-эффект, будут тщательнее и эффективнее планироваться производство, работа склада, отдела продаж, проведение стимулирующих маркетинговых акций.

**Список источников**

* [sklearn.metrics.mean\_absolute\_error](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html)
* [SARIMAX: Introduction](https://www.statsmodels.org/dev/examples/notebooks/generated/statespace_sarimax_stata.html)
* [Keras, LSTM layer](https://keras.io/api/layers/recurrent_layers/lstm/)
* [Introduction to ARCH Models](https://arch.readthedocs.io/en/latest/univariate/introduction.html)
* [Numpy and Scipy Documentation](https://docs.scipy.org/doc/)
* [sklearn.linear\_model.LogisticRegression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)
* [sklearn.tree.DecisionTreeClassifier](https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html)
* [sklearn.ensemble.StackingClassifier](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.StackingClassifier.html)