Репозиторий проекта: проекта.

1 Постановка задачи

В работе было необходимо построить прогноз значений временного ряда с помощью моделей эконометрики.

Модель ARIMA (Autoregressive Integrated Moving Average) используется для прогнозирования временных рядов. Формула модели ARIMA(p, d, q):

$$\Delta^d y_t = \alpha_1 \Delta^d y_{t-1} + ... + \alpha_p \Delta^d y_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + ... + \beta_q \varepsilon_{t-q},$$

y_t - значение временного ряда в момент времени t, $\Delta = (1 - L)^d$, ε_t - белый шум. L - лаговый оператор.

Модель SARIMA – расширение для рядов с сезонной составляющей. SARIMAX – расширение, включающее внешнюю регрессионную составляющую

2 Ход работы

2.1 Выбранные модели и их спецификация

В работе решалась задача прогнозирования числа продаж в сети магазинов Эквадора за 2016 год. Было рассмотрено 3 способа прогнозирования:

- ARIMA модель с ручной настройкой параметров p, d и q
- SARIMAX модель
- Модель машинного обучения (градиентный бустинг)

Качество оценивалось с помощью следующих метрик: MSE, RMSE, MAE, абсолютная процентная ошибка. Результаты сравнивались с наивным прогнозом – средним числом продаж за последние 3 месяца.

Для моделей SARIMAX и бустинга рассматривались два подхода. В первом случае в качестве переменных в модель подавались временные признаки: номера месяца, квартала, дня в году и недели, а также день недели. Кроме этого, использовались «лаговые переменные»: число продаж в эту же дату год, два и три назад (если есть данные за этот период). Такой подход позволяет

получать прогноз модели на любое число дней вперед. Во втором случае к этим переменным добавлялись переменные из датасета: цена на топливо за выбранную дату, число акций в магазинах и число транзакций. Трудность второго подхода заключается в том, что мы не знаем будущие значений добавленных драйверов заранее. Если мы хотим получить прогноз на даты, которых нет в нашем датасете, – величины драйверов тоже придется предсказывать.

2.2 Предварительная обработка данных

Переменные датасета требовали предварительной предобработки. Например, в значениях цены на газ были пропуски, которые были заполнены с помощью интерполяции квадратичными сплайнами. На рисунках 1 и 2 приведены графики цены до интерполяции и после соответственно.

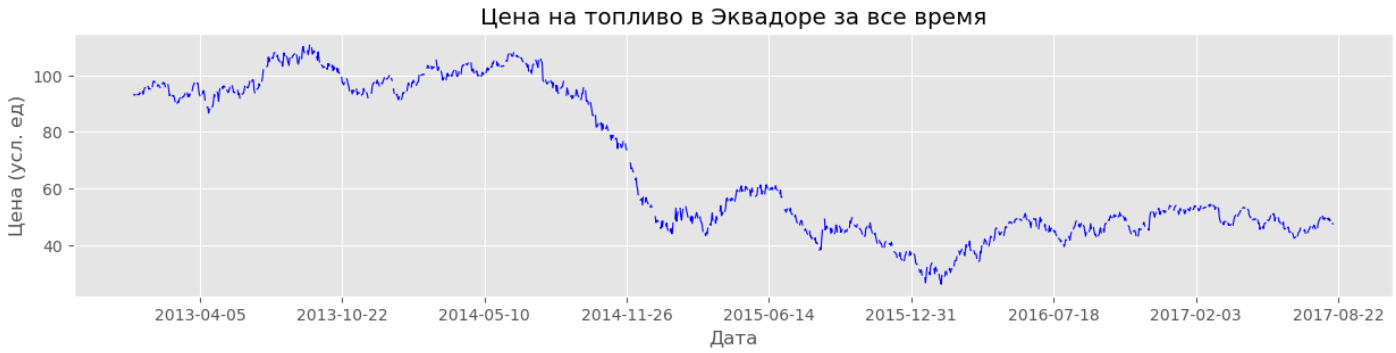

Рис. 1: Цена за литр топлива в Эквадоре до интерполяции.

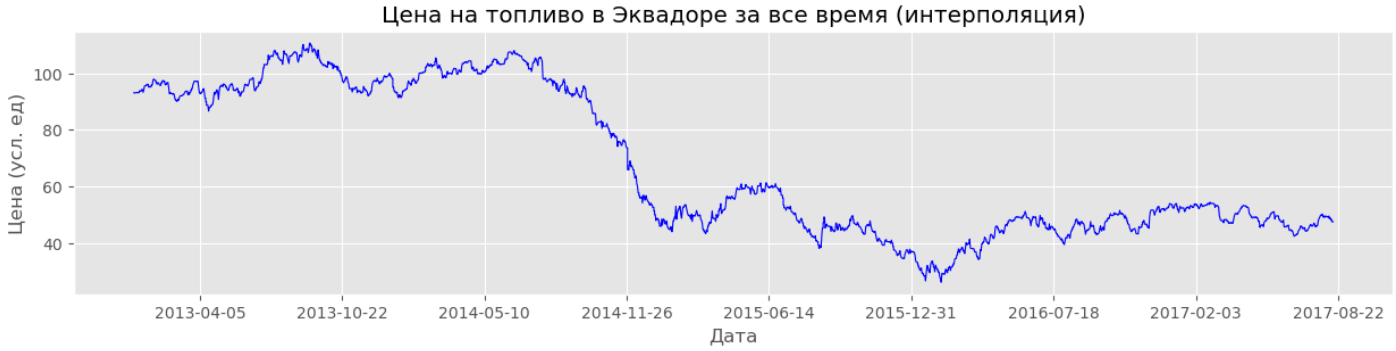

Рис. 2: Цена за литр топлива в Эквадоре после интерполяции.

Кроме этого, категориальные переменные, такие как тип магазина, его локация и флаг праздничного дня, были закодированы числовыми значениями. Однако далее все эти признаки были исключены из модели из-за маленькой дисперсии (весь признак мог быть заполнен одним значе-

нием) или из-за их мальтиколлинеарности. Кроме того, для этих признаков наблюдалось большая доля пропусков, порядка 80 процентов.

2.3 Обучение моделей

2.3.1 Наивный прогноз

Как уже было описано ранее, для наивного прогноза использовалось среднее значение продаж в Эквадоре за последние 3 месяца 2016 года. График прогноза приведен на рисунке 3.

Рис. 3: Наивный прогноз для продаж в Эквадоре.

Для прогноза были измерены основные метрики. Их значения приведены в таблице 1.

Таблица 1: Значения метрик на прогнозе модели градиентного бустинга.

	MSE	RMSE	MAE	Процентная ошибка
Временные переменные	20703.85	143.88	114.75	0.2

2.3.2 ARIMA модель

Выбор параметров для ARIMA модели осуществлялся на основе значений функций автокорреляции и частичной автокорреляции. Стационарность ряда проверялась с помощью теста Дики — Фуллера.

Было установлено, что исходный ряд не является стационарным ($p - value : 0.68$). После однократного дифференцирования ряд стал стационарным ($p - value : 6.33 \cdot 10^{-14}$). Кроме того, для ошибок был построен график их распределения, оно оказалось близко к нормальному.

Результаты работы ARIMA модели проведены на рисунке 4.

Рис. 4: Прогноз ARIMA модели для числа продаж в Эквадоре.

Из графика видно, что модель не улавливает динамику ряда и очень скоро прогноз превращается в константный. Значения метрик для ARIMA модели приведены в таблице 2.

Таблица 2: Значения метрик на прогнозе ARIMA модели.

	MSE	RMSE	MAE	Процентная ошибка
Временные переменные	11908.6	109.13	93.53	1.491

2.3.3 SARIMAX модель

Главное отличие SARIMAX модели от ARIMA модели – ее способность учитывать сезонность в предсказании, а также дополнительные параметры, помимо значений самого ряда. Как уже было сказано выше, в качестве дополнительных параметров в модель подавались значения числа продаж в дату, цена топлива и числа акций в магазинах. Обучение модели проходило автоматически с помощью пакета *statsmodels* языка Python. Кроме того, модель обучалась два раза: в первый раз – только на временных и лаговых переменных, во второй – с дополнительными переменными. Результаты предсказания модели приведены на рисунке 5. Значения метрик представлены в таблице 3.

Таблица 3: Значения метрик на прогнозе SARIMAX модели.

	MSE	RMSE	MAE	Процентная ошибка
Временные переменные	24258.07	155.75	125.47	0.22
Временные переменные + дополнительные	19785.57	140.661	88.76	0.15

Рис. 5: Прогноз SARIMAX модели для числа продаж в Эквадоре.

Из графика и значений метрик видно, что учет в модели дополнительных параметров дает значительный прирост в качетсве предсказания.

2.3.4 Boosting

В качестве дополнения к работе, нам было интересно посмотреть, какие результаты покажет на тех же входных данных бустинг. Принцип работы бустинга принципиально отличается от эконометрических моделей, рассмотренных выше.

В работе рассматривался XGBRegressor из библиотеки *sklearn*. Обучение также как и для SARIMAX модели осуществлялось двумя подходами. Перебор гиперпараметров осуществлялся с помощью библиотеки *hyperopt*. В отличии от обычного перебора по сетке, данный подход использует средства Байесовской оптимизации.

Результаты предсказания модели приведены на рисунке 6. Значения метрик представлены в таблице 4.

Рис. 6: Прогноз модели градиентного бустинга для числа продаж в Эквадоре.

Из графика и значений метрик видно, что точность прогноза для спецификации с дополнитель-

ными переменными близка к реальным значениям числа продаж.

Таблица 4: Значения метрик на прогнозе модели градиентного бустинга.

	MSE	RMSE	MAE	Процентная ошибка
Временные переменные	14971.71	122.36	86.83	0.15
Временные переменные + дополнительные	2745.55	52.4	39.19	0.07

3 Заключение

В работе была решена задача прогнозирования числа продаж в сети магазинов Эквадора за 2016 год. Было рассмотрено 3 способа прогнозирования:

- ARIMA модель с ручной настройкой параметров p, d и q
- SARIMAX модель
- Модель машинного обучения (градиентный бустинг)

Было установлено, что среди эконометрических моделей лучше всего себя показала модель SARIMAX. Этот результат, в частности, обусловлен ее способностью учета внешних параметров, а не только прошлые значения ряда.

Кроме того, в работе дополнительно был рассмотрен способ прогнозирования с помощью градиентного бустинга. Этот метод превзошел все эконометрические модели по качеству предсказаний.

Все подробности по отбору и обработке признаков, обучению и валидации моделей можно найти на странице проекта.