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

В работе решалась задача прогнозирования числа продаж в сети магазинов Эквадора. 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 последнюю четверть тренировочного периода. Выбор параметров для ARIMA модели осуществлялся на основе значений функций автокорреляции и частичной автокорреляции. Стационарность ряда проверялась с помощью теста Дики — Фуллера.

Кроме этого, в работе была обучена ml-модель градиентного бустинга. Для нее был использован перебор гиперпараметров пакета hyperopt, в основе которого лежит байесовская оптимизация. Результаты бустинга сравнивались с результатами экономических моделей и наивного прогноза.

Все графики и данные доступны на странице проекта.

3 Заключение

Были получены следующие значения метрик:

Таблица 1: Преимущества и недостатки недифференцированного маркетинга

Метрика	ARIMA	SARIMAX	XGBRegressor	Наивный прогноз
MSE	370425.063612	11849.563053	2761.131836	20720.685499
RMSE	608.625553	108.855698	52.546473	143.946815
MAE	600.125711	70.474641	42.322258	114.795845
Процентная ошибка	1.133873	0.117918	0.080065	0.196057

По метрикам RMSE, MAE и процентной ошибке лучше всего себя показала ML модель. Второй по качеству оказалась SARIMAX модель. Полученные результаты можно объяснить тем, что ML модель имеет другую архитектуру (более сложный и тонкий процесс обучения), является более совершенной и новой, по сравнению с эконометрической моделью.