РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Мария (github.com/krasarma)

СОДЕРЖАНИЕ

АНАЛИЗ ДАННЫХ

Структурный анализ

- Проведена проверка заполняемости таргета и экзогенных признаков. Наблюдения без таргета из обучающей выборки исключены. Заполняемость экзогенных признаков на обучающей - 5% и более.
- Проведена проверка корректности типов данных, проверка доли уникальных значений для признаков, уникальность временных меток.

Статистический анализ

- Произведен рассчет статистик для экзогенных признаков: эксцесс, ассиметрия, пик, среднее, медиана, перцентили, распределение. Сделан вывод о типах распределений. На основе полученных данных приняты решения о потребности и методах нормализации, стандартизации и обработки выбрасов для признаков. (Статистики и визуализация в ноутбуке)

АНАЛИЗ ДАННЫХ

Статистический анализ

- Провизведено исследование на наличие выбросов методом межквартильного размаха (IQR, Interquartile Range). IQR рассчитывается как разница между Q3 и Q1: IQR=Q3−Q1. По итогам исследования было принято решение об обработке выбросов для признаков. Результаты визуализированы в ноутбуке с помощью boxplot.

Статистический анализ

- Проведен тест Дики-Фулера в целях определения стационарности ряда. С учетом полученных результатов, считаем ряд стационарным.

ADF статистика: -4.470501502413692 5

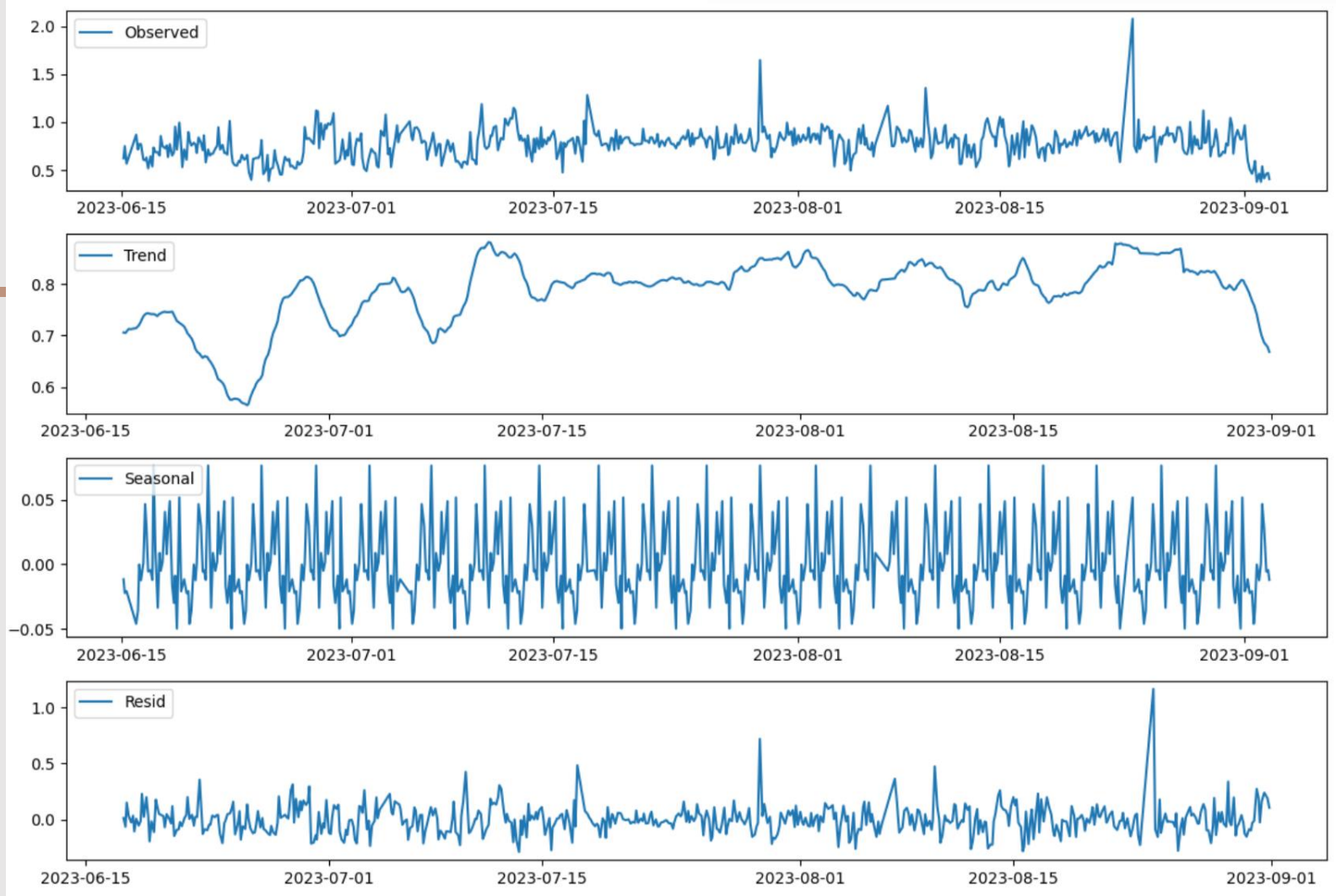
p-значение: 0.00022223504507362 13

Критические значения:

1%: -3.441204979288887, 5%: -2.86632910370007, 10%: -2.56932048425654

АНАЛИЗ ДАННЫХ

Декомпозиция временного ряда



ПОДГОТОВКА ДАННЫХ

Нормализация данных

- Для экзогенных признаков была проведена нормализация. С учетом результатов проведенного анализа:
  median: 0, 4, 5, 6, 7, 8, 9, 12, 13, 14
- mean: 3, 10, 11

Обработка выбрасов

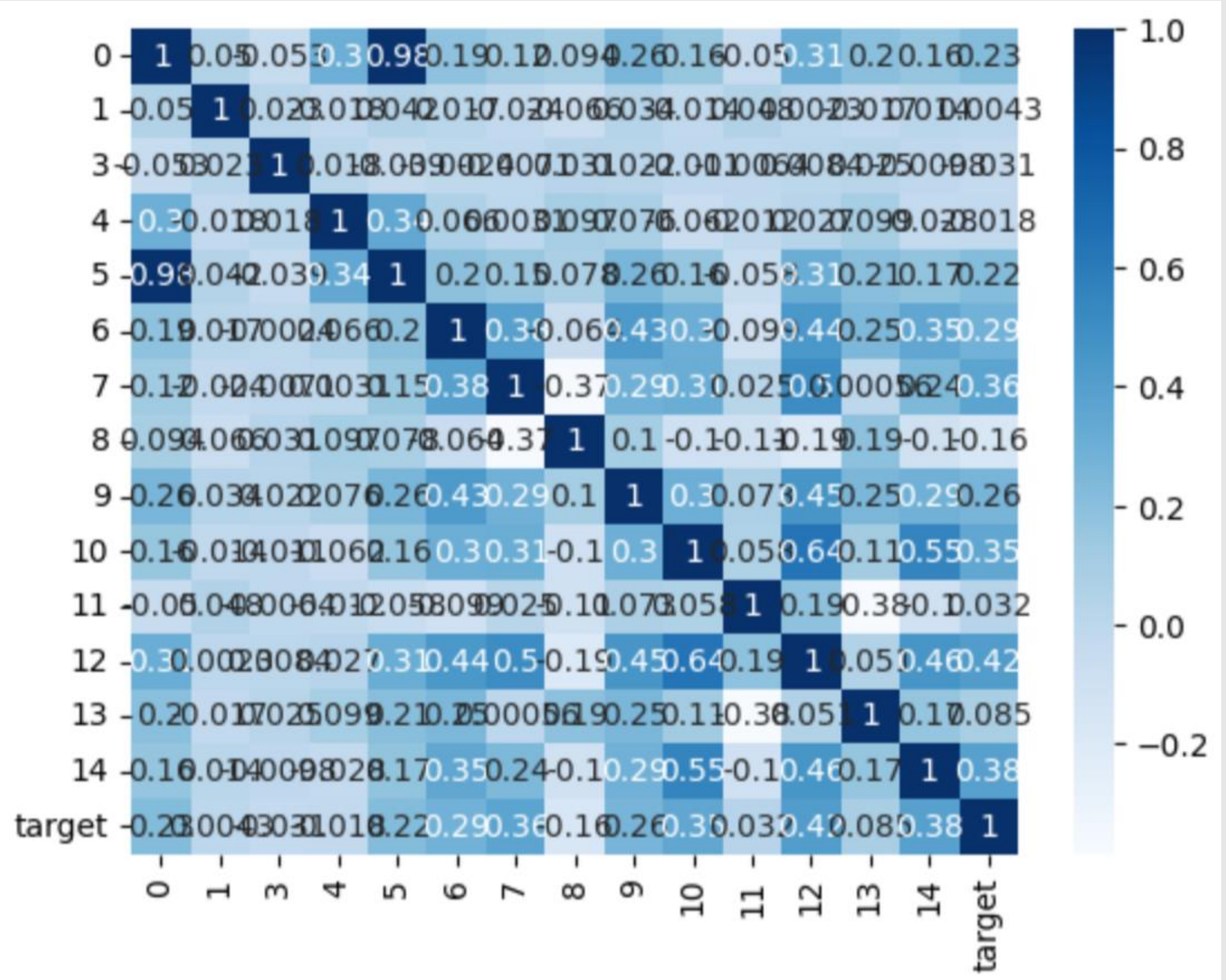
- В соответствии с резульатами исслебования обработаны выбросы в признаках: 0, 4, 5, 7, 12, 13, 14 на уровне 0.99 квантиля.

Стандартизация данных

- Нормализованны с помощью StandardScaler признаки 3, 6, 8, 10, 11

ОТБОР ПРИЗНАКОВ

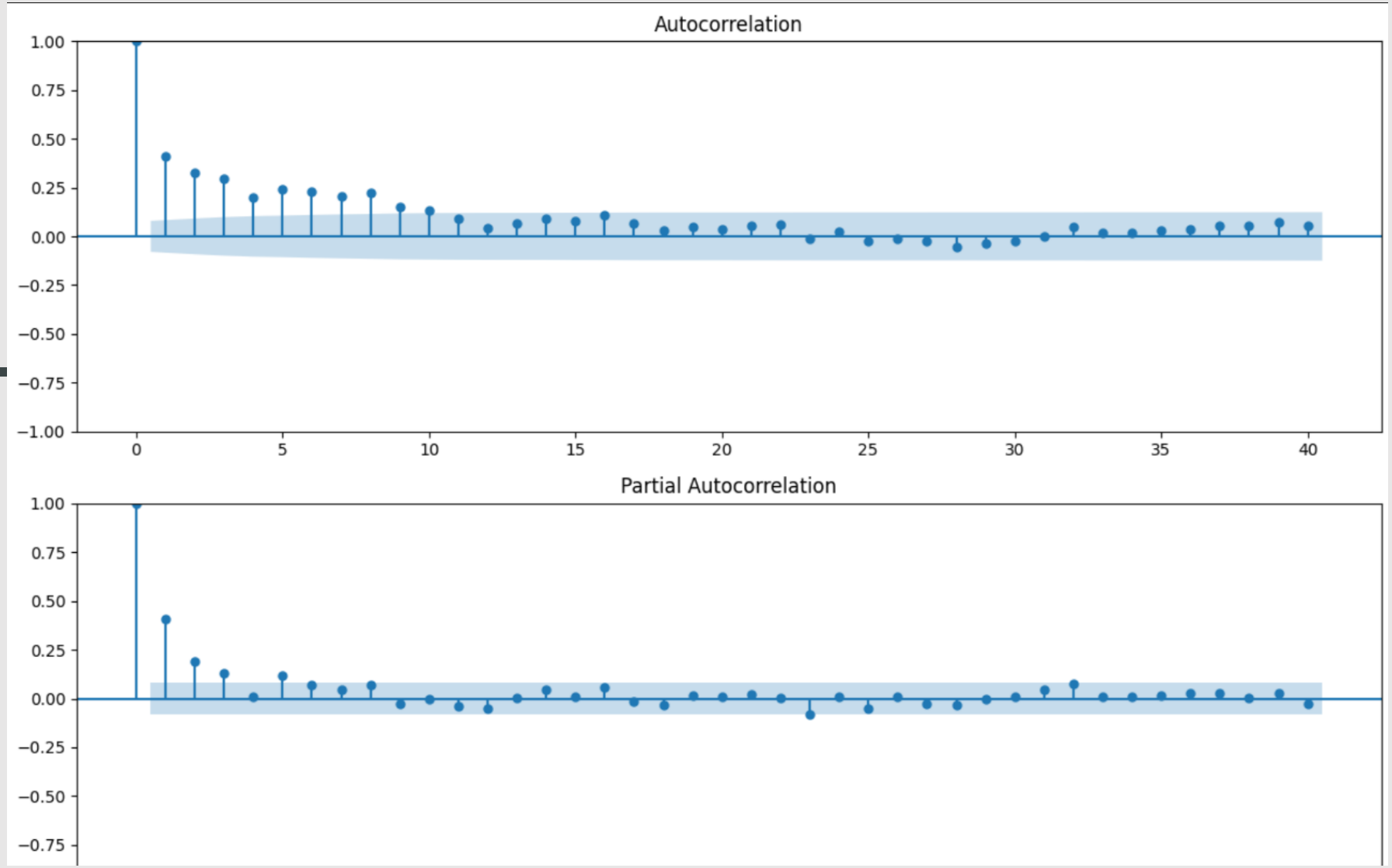
Для отбора экзогенных признаков была проведена оценка корреляции, а также оценка PSI. По результатам анализа из выборки были исключены признаки 0, 1, 3, 4, 11. Визуализация PSI и рассчеты в ноутбуках.



ЛАГОВЫЕ ПРИЗНАКИ

Анализ автокорреляции и частичной автокорреляции

По результатам анализо было принято решение о создание пяти лаговых признаков для наблюденияю



СОЗДАНИЕ ЛАГОВЫХ ПРИЗНАКОВ

Основные критически важные результаты

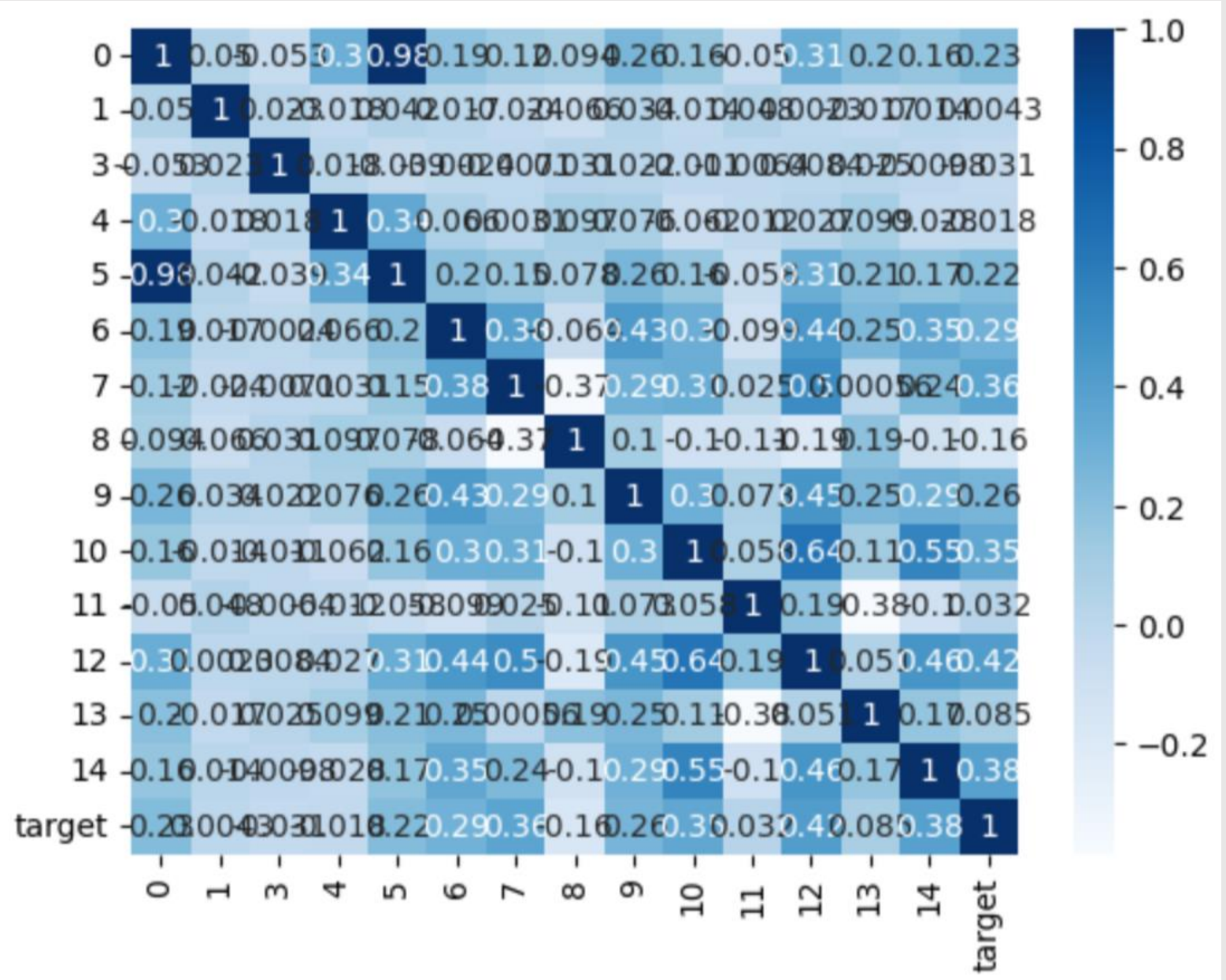
- Запуск продукта
- Обновления программного обеспечения
- Пресс-релиз
- Печатные материалы

Оценка уверенности

- Уверенность 5/5, что проект будет завершен по расписанию

ОТБОР ПРИЗНАКОВ

Для отбора экзогенных признаков была проведена оценка корреляции, а также оценка PSI. По результатам анализа из выборки были исключены признаки 0, 1, 3, 4, 11. Визуализация PSI и рассчеты в ноутбуках.



МОДЕЛИРОВАНИЕ

XGBoost

Превой для эксперемента была выбрана модель XGBoost - алгоритм машинного обучения, основанный на методе градиентного бустинга. Это решение показало более низкое качество в сравнении с альтернативным подходом.

ARIMA

Второй была выбрана ARIMA. Эта статистическая модель хорошо подходит для стационарных временных рядов и позволяет эффективно учитывать автокорреляцию. Она показала наилушие результаты и для прогнозирования будем использовать именно ее.

ПОДБОР МАКРОПАРАМЕТРОВ

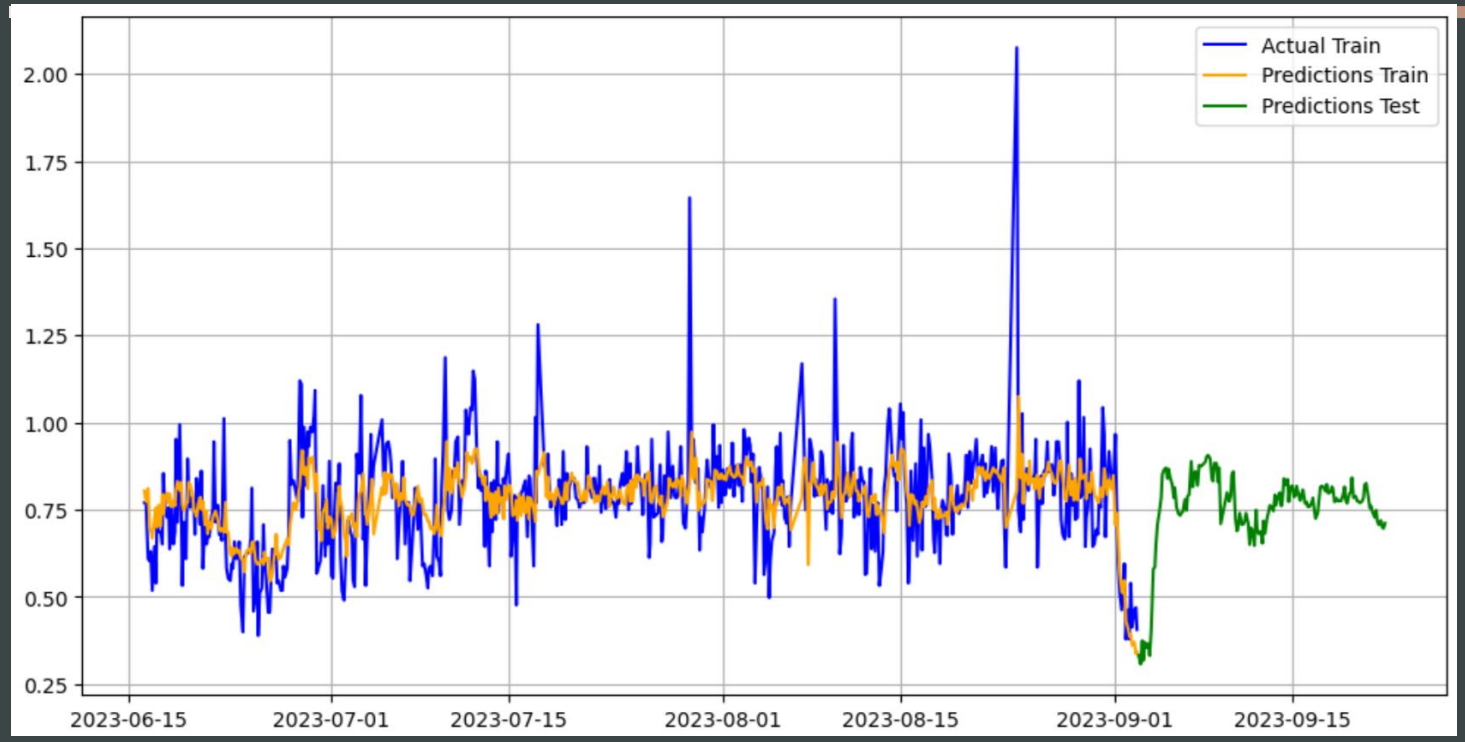
XGBoost

Для подбора макропараметров использована optuna, для оценки качества RMSE и R^2, кроссвалидация на 5 сплитов. Регуляризация - l2, т.к. мы располагаем коротким списком отобранных и предобработанных фичей. Результаты: objective='reg:squarederror', alpha=0.7301647364174587, learning_rate=0.0 8531620322319103, n_estimators=326

ARIMA

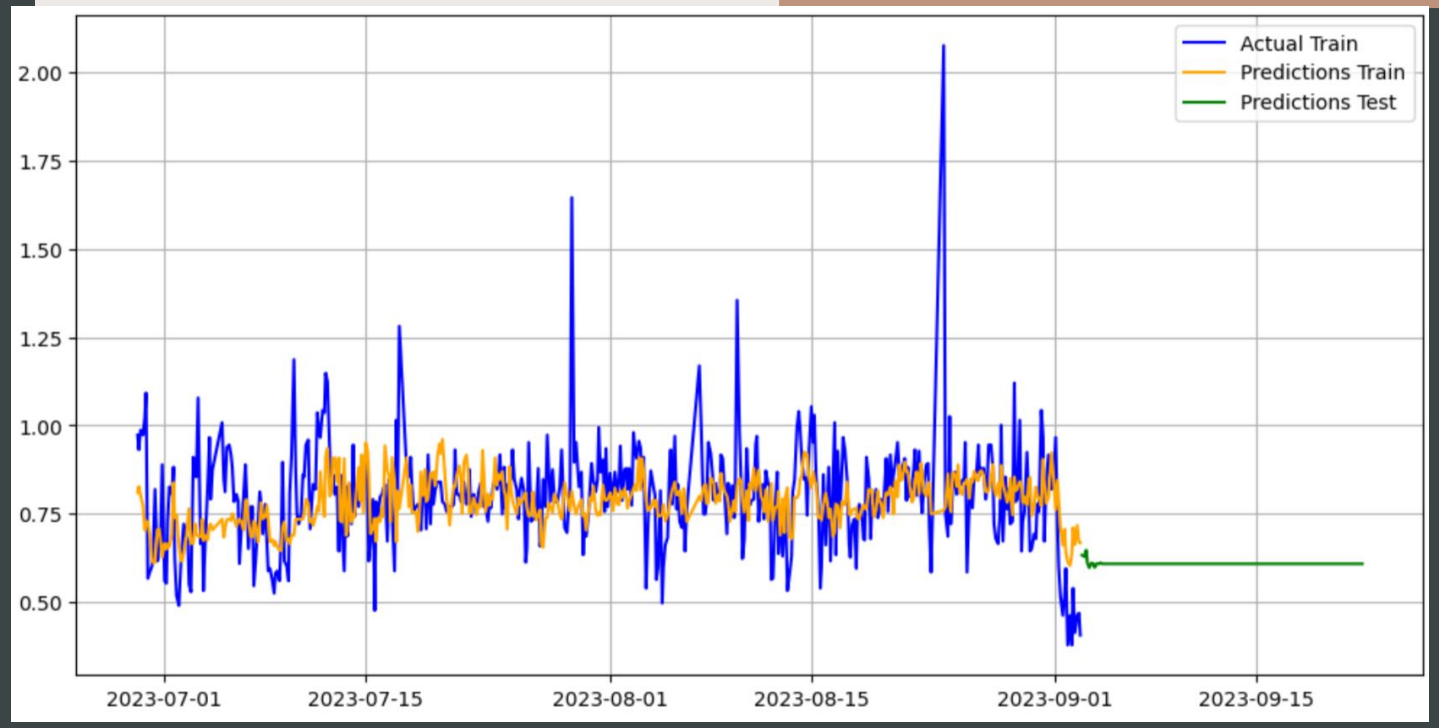
Для отбора также использованы optuna и приведенные метрики. Результаты подбора: 'p': 4, 'd': 0, 'q': 4. Модели не передаются лаговые признаки.

РЕЗУЛЬТАТЫ ARIMA



RMSE: 0.12841
R^2: 0.30943

РЕЗУЛЬТАТЫ XGBOOST



Mean RMSE: 0.1436
Mean R^2: -0.0227