ALIF ACADEMY

Прогнозирование продаж в розничной сети: анализ временных рядов

Джаводов Санджар

Требования к датасету и описание задачи

Требования к датасету:

- Наличие временной составляющей: ежедневные данные о продажах

Описание задачи:

- Прогнозирование продаж для тысяч семейств продуктов в магазинах Favorita в Эквадоре
- Горизонт прогнозирования: 15 дней после последней даты в обучающем наборе

Бизнес-цели проекта

1. Оптимизация управления запасами
2. Улучшение планирования поставок
3. Повышение эффективности ценообразования
4. Адаптация к сезонным колебаниям и праздникам
5. Учет влияния экономических факторов (например, цен на нефть) на продажи

Описание данных

Основные файлы:

- train.csv: Обучающие данные
- test.csv: Тестовые данные
- stores.csv: Метаданные магазинов
- oil.csv: Ежедневные цены на нефть
- holidays_events.csv: Праздники и события

Ключевые признаки:

- store_nbr: идентификатор магазина
- family: тип продукта
- sales: целевая переменная (объем продаж)
- onpromotion: количество товаров по акции
- Дополнительные метаданные: город, штат, тип магазина, кластер

Схема данных в реальной жизни

Архитектура хранения данных в ClickHouse:

- Колоночное хранение для эффективной работы с временными рядами
- Оптимизированное сжатие данных
- Распределенная архитектура с возможностью масштабирования

Основные таблицы и их структура:

- sales: основная таблица продаж (date, store_nbr, item_nbr, unit_sales, onpromotion)
- products: справочник продуктов (item_nbr, family, class, perishable)
- stores: информация о магазинах (store_nbr, city, state, type, cluster)
- transactions: данные о транзакциях (date, store_nbr, transactions)
- holidays: календарь праздников и событий
- oil_prices: данные о ценах на нефть

Схема данных в реальной жизни

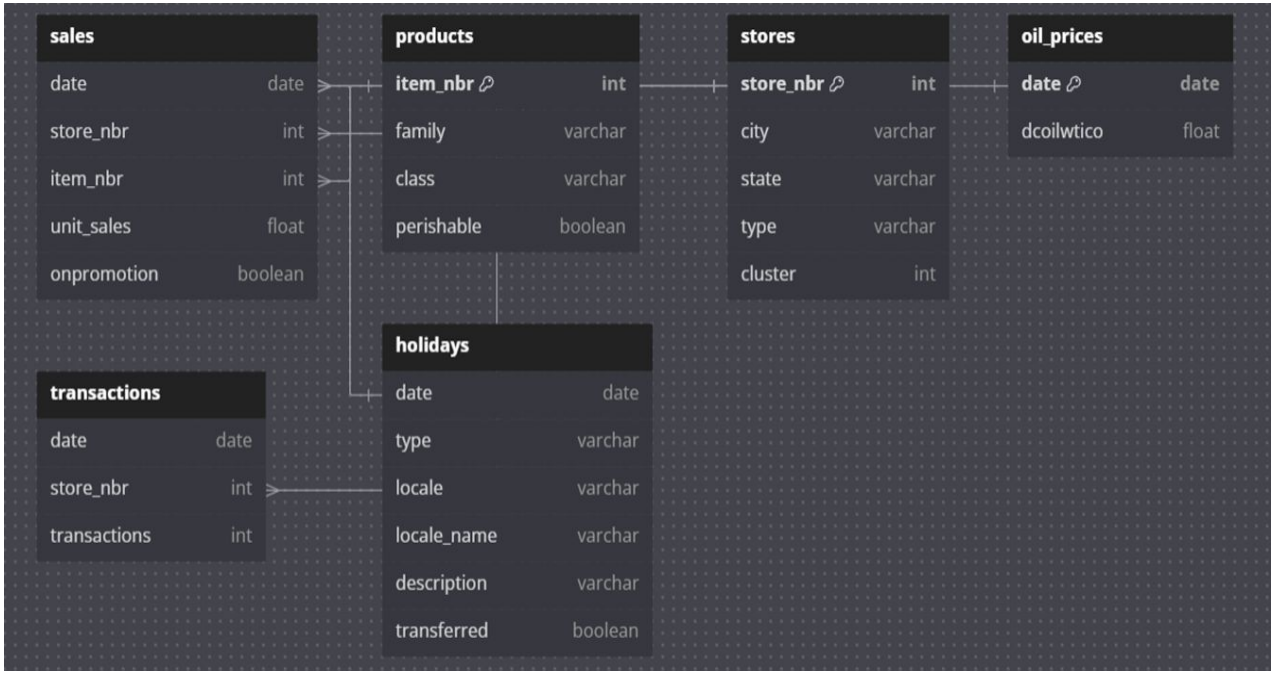
Ключевые особенности реализации:

- Использование первичных и вторичных индексов для ускорения запросов
- Оптимизация для аналитических запросов
- Эффективное хранение временных рядов
- Поддержка интеграции с ML-инструментами

Связи между таблицами:

- store_nbr как ключ связи между sales, stores и transactions
- item_nbr связывает sales и products
- date как временной идентификатор для всех таблиц

Схема данных в реальной жизни



ML Систем Дизайн(Реализованные компоненты)

Сбор и хранение данных

- Источники данных:
  - Продажи и информация о продуктах (sales, products)
  - Данные о магазинах (stores)
  - Цены на нефть (oil_prices)
  - Календарь праздников (holidays)
- Хранение: ClickHouse (как показано на схеме БД)

Подготовка данных

- Преобразование временных рядов
- Обработка статических ковариат
- Создание признаков на основе календаря и промо-акций

Метрики и валидация:

- Основная метрика: RMSLE
- Бэктестинг на исторических данных
- Валидация на отложенной выборке (16 дней)

Обучение моделей

- Базовые модели:
  - Наивная сезонная модель
  - Экспоненциальное сглаживание
  - Facebook Prophet
- Продвинутые модели:
  - LSTM
  - N-HiTS
  - TFT
  - Бустинг (LightGBM, CatBoost)

ML Систем Дизайн(Планируемые улучшения)

Мониторинг и обновление

- Система отслеживания дрейфа данных
- Автоматическое переобучение моделей
- A/B тестирование

Интерпретация

- Внедрение SHAP для объяснения прогнозов
- Создание интерактивных дашбордов

Инфраструктура

- Внедрение Kafka для потоковой обработки
- Контейнеризация с Docker
- Оркестрация через Kubernetes

Анализ и предобработка данных

Предварительная обработка

Создание объектов TimeSeries в Darts

Формирование статических ковариат:

- Номер магазина
- Семейство продуктов
- Город
- Регион
- Тип и кластер магазина

Период данных: 2013-2017 годы

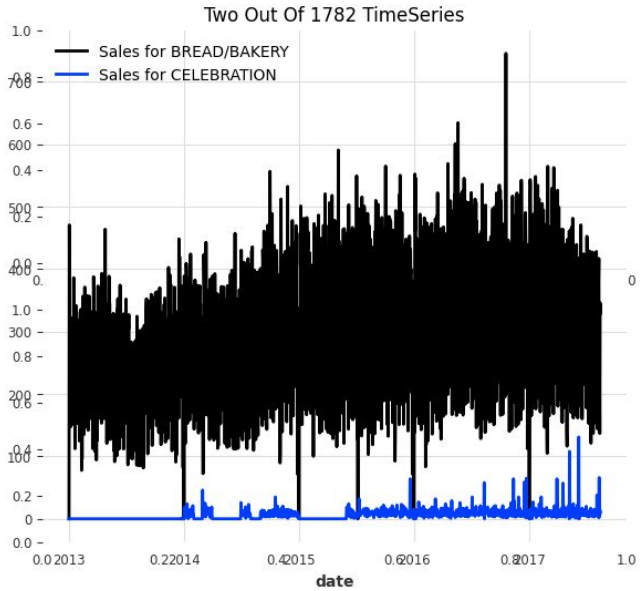
Временные признаки

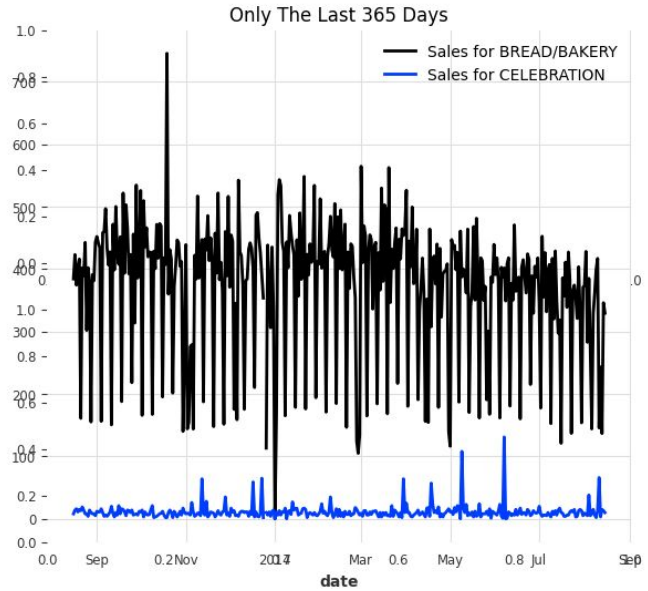
- День недели
- Месяц
- Год
- Праздничные дни

Трансформация данных

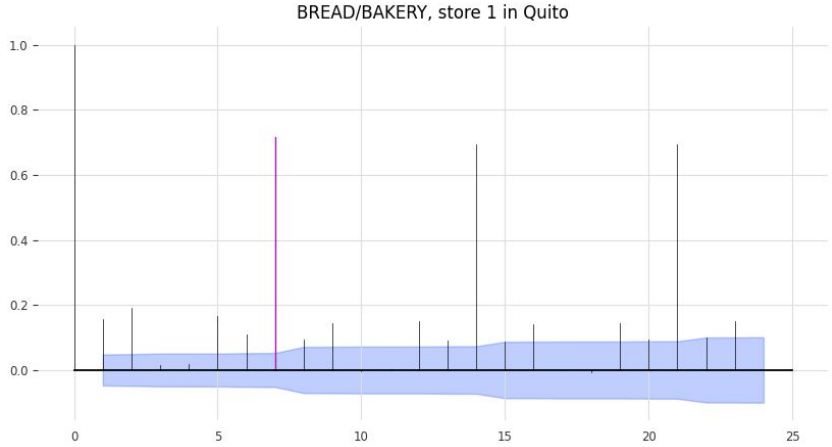
- Масштабирование всех рядов в диапазон [0,1]
- Логарифмическая трансформация
- Объединение всех ковариат в единый набор

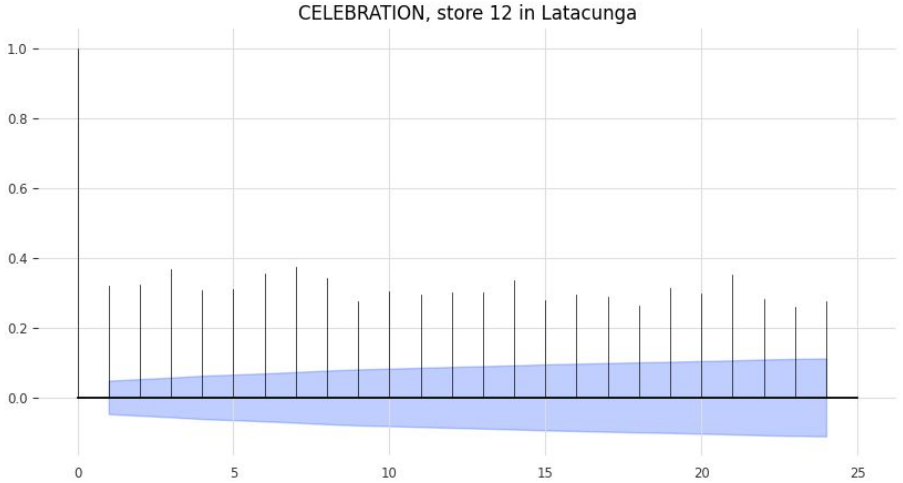
EDA анализ


Two Out Of 1782 TimeSeries
Sales for BREAD/BAKERY
Sales for CELEBRATION
date

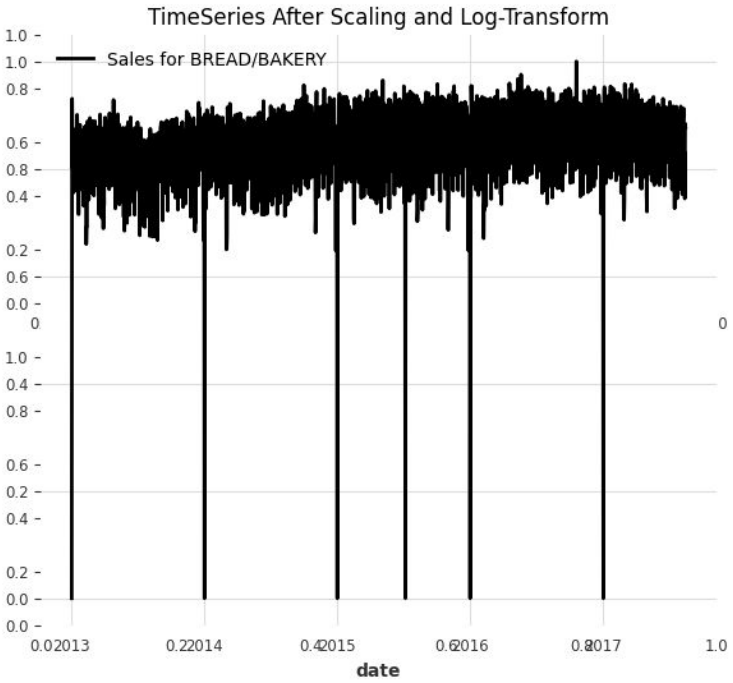

Only The Last 365 Days
Sales for BREAD/BAKERY
Sales for CELEBRATION
date

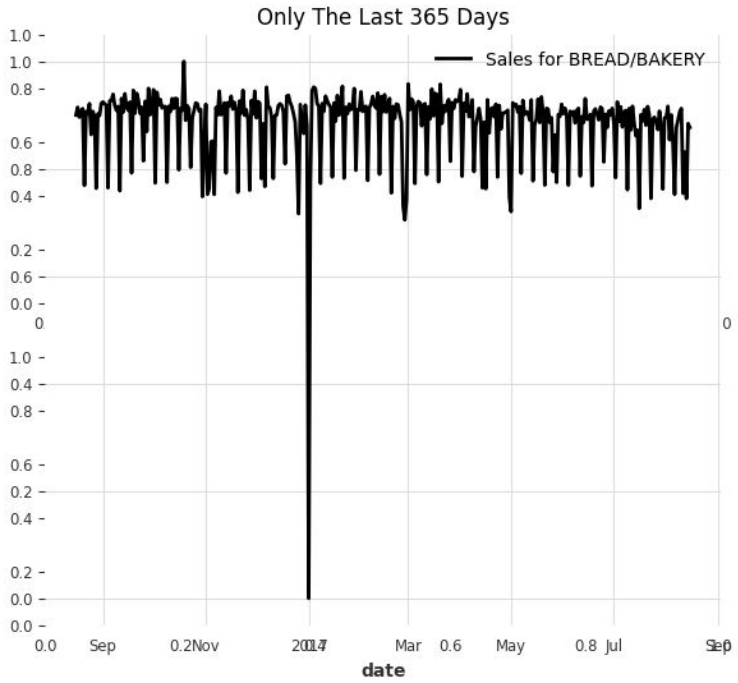
EDA анализ


BREAD/BAKERY, store 1 in Quito
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25

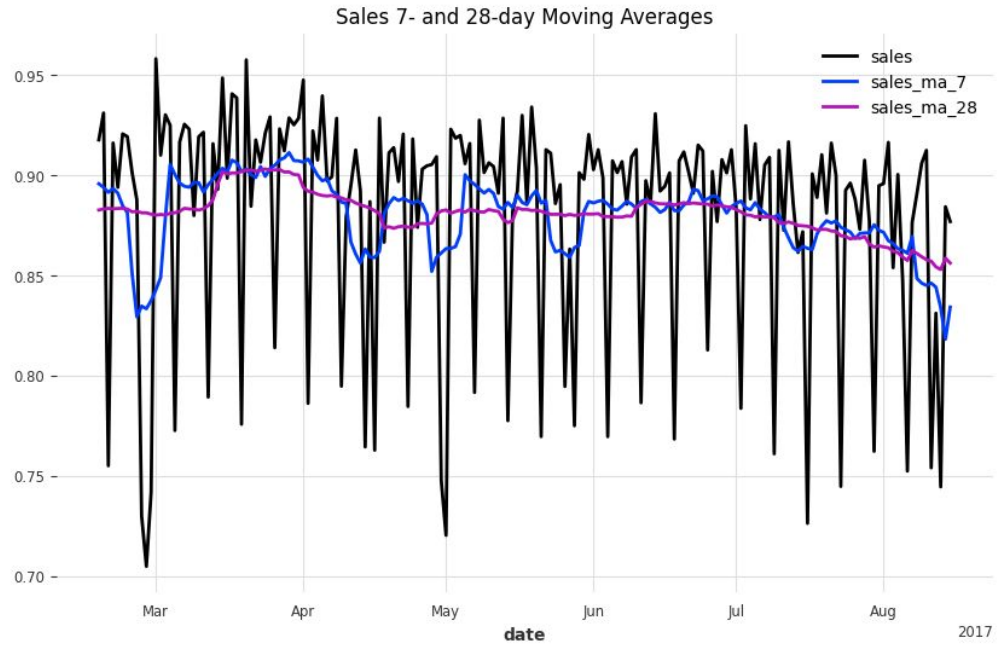

CELEBRATION, store 12 in Latacunga
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25

EDA анализ


TimeSeries After Scaling and Log-Transform
Sales for BREAD/BAKERY
date

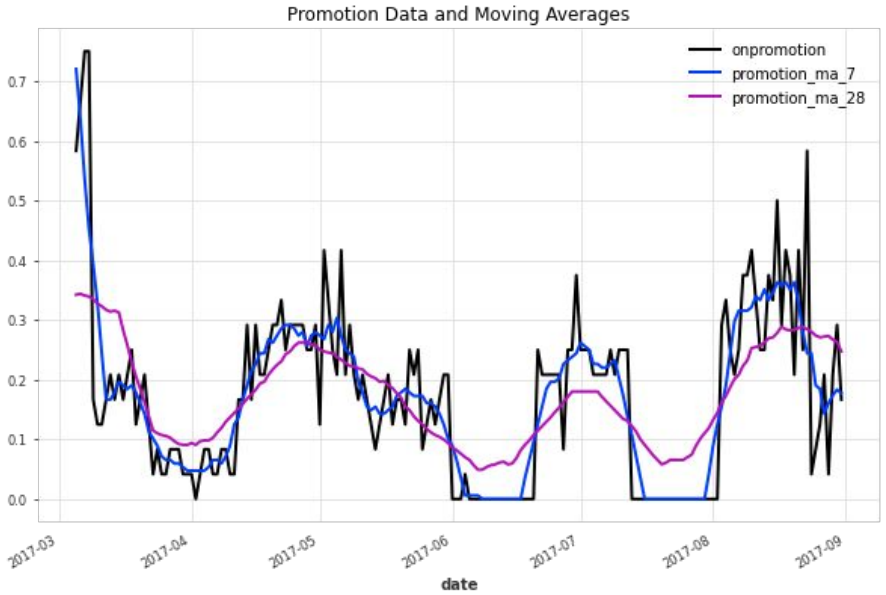

Only The Last 365 Days
Sales for BREAD/BAKERY
date

EDA анализ


Sales 7- and 28-day Moving Averages
sales
sales_ma_7
sales_ma_28
0.95
0.90
0.85
0.80
0.75
0.70
Mar
Apr
May
Jun
Jul
Aug
2017
date

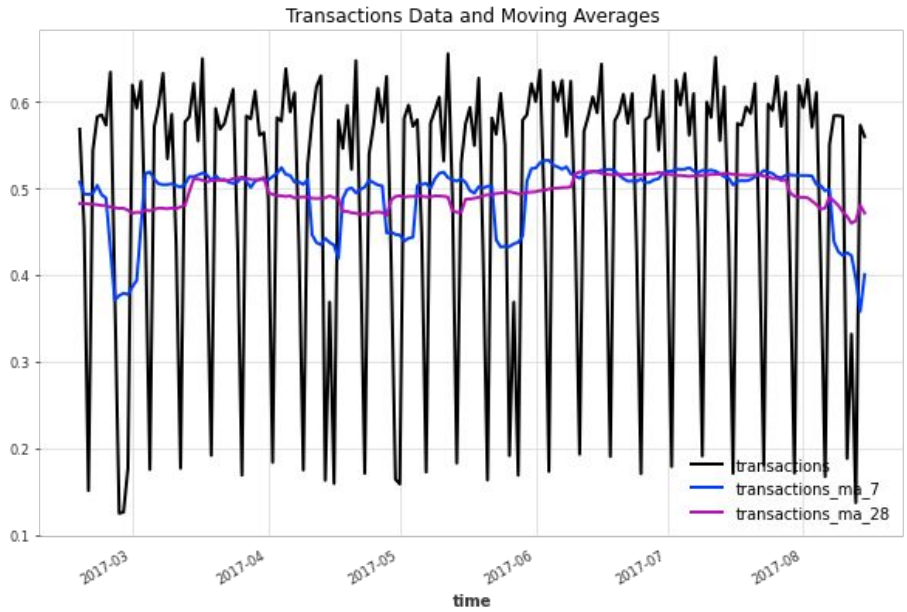
EDA анализ

Promotion Data



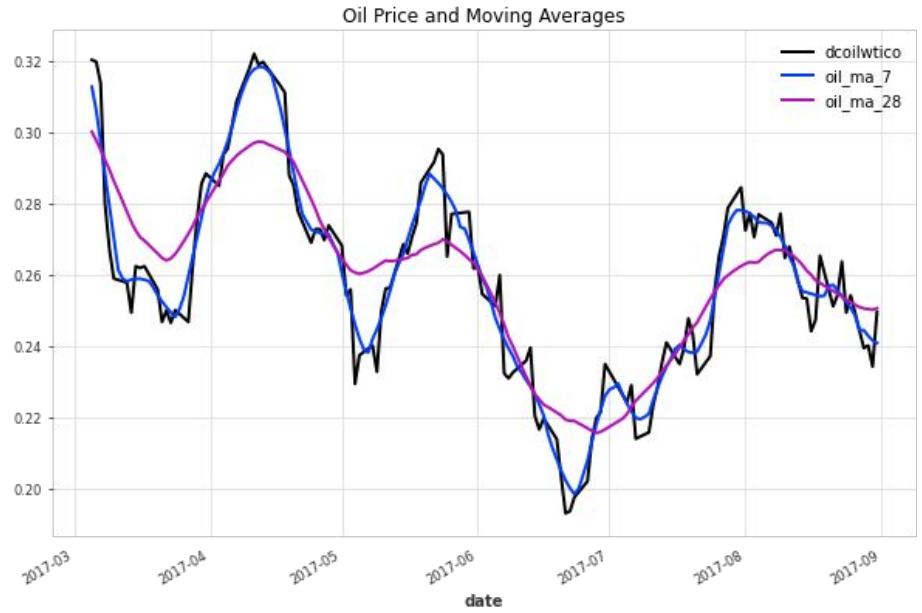
EDA анализ

Transactions



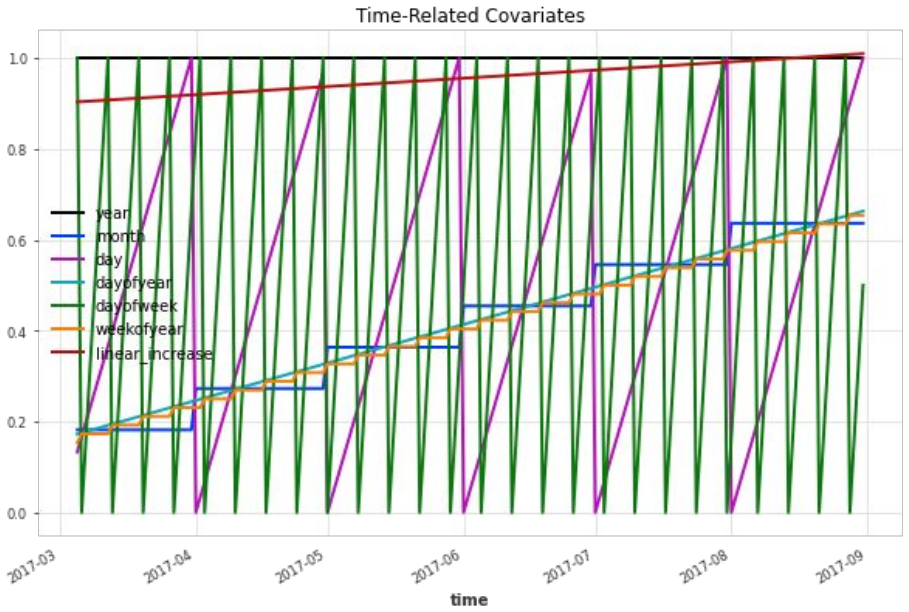
EDA анализ

Oil Price

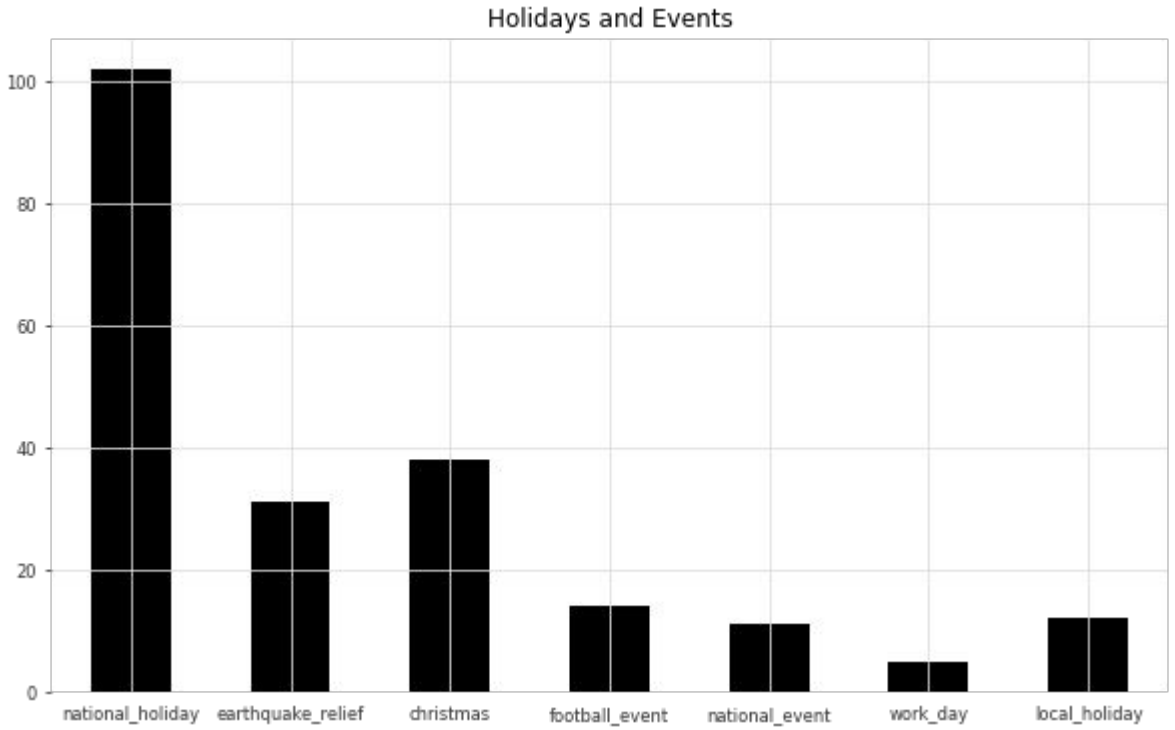


EDA анализ

Time Dummies and Covariates



EDA Анализ


Holidays and Events
100
80
60
40
20
0
national_holiday
earthquake_relief
christmas
football_event
national_event
work_day
local_holiday

Базовые модели (Бейзлайн)

Реализованные модели:

- Наивная сезонная модель (K=7 дней)
  - Простое повторение последних 7 дней
  - Используется как базовый уровень сравнения
  - Реализация через NaiveSeasonal из Darts
- Экспоненциальное сглаживание
  - Показала лучшие результаты среди базовых моделей
  - Реализация через ExponentialSmoothing из Darts
  - Более устойчива к колебаниям в данных
- Facebook Prophet
  - Автоматическое определение сезонности
  - Учет праздничных дней
  - Реализация через Prophet из Darts

Базовые модели (Бейзлайн)

Методология тестирования:

- Бэктестинг на исторических данных
- Начальная точка: 1 ноября 2016
- Горизонт прогнозирования: 16 дней
- Метрика оценки: RMSLE

Особенности реализации:

- Использование трансформации данных
- Обратное преобразование для оценки
- Визуализация результатов для последних 365 дней

Результаты:

- Тестирование на категории BREAD/BAKERY (стабильный ряд)
- Сравнение прогнозов с реальными данными
- Экспоненциальное сглаживание показало наилучшие результаты

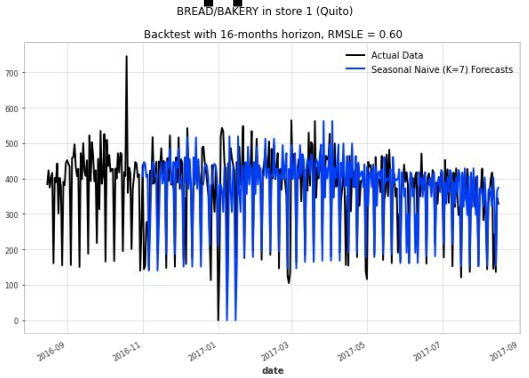
Сравнение базовых моделей на разных типах данных

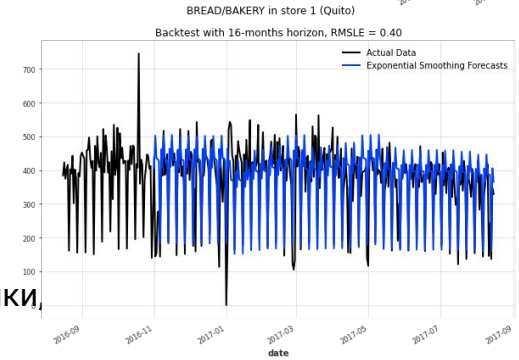
Анализ BREAD/BAKERY (стабильный паттерн):

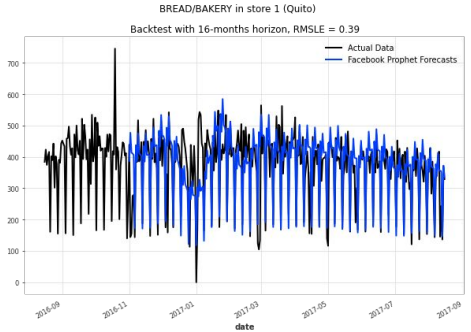
- Стабильный недельный паттерн продаж
- Единственная аномалия: нулевые продажи в период Рождества/Нового года 2017
- Все три модели хорошо прогнозируют регулярный паттерн
- Наивная сезонная модель плохо восстанавливается после резкого спада

Анализ CELEBRATION (сложный паттерн):

- Нерегулярные всплески продаж, связанные с особыми событиями
- Результаты моделей:
  - Наивная сезонная: генерирует ложные пики, пропускает реальные
  - Экспоненциальное сглаживание и Prophet: улавливают базовые сезонные паттерны, но пропускают пики







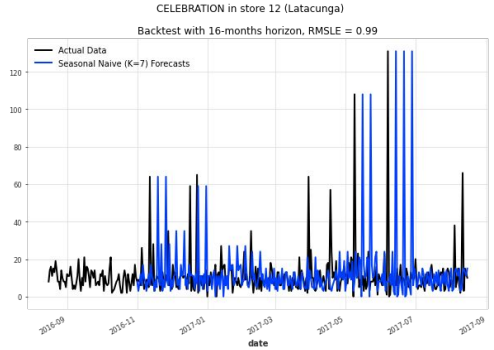
Сравнение базовых моделей на разных типах данных

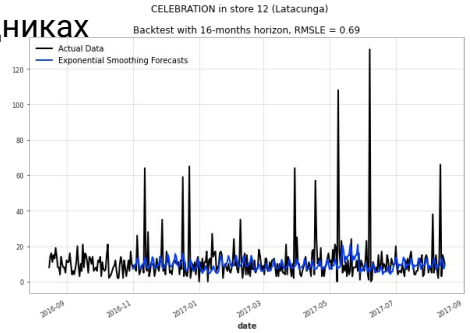
Бизнес-последствия:

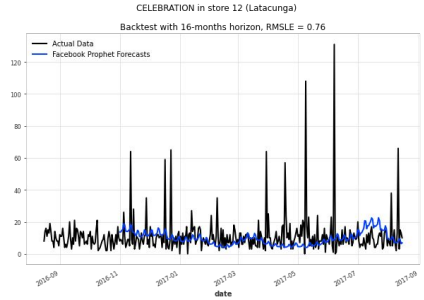
- Риск недостаточных запасов праздничных товаров в периоды пикового спроса
- Потенциальная потеря прибыли из-за неточных прогнозов

Выводы для улучшения моделей:

- Необходимость учета общей сезонности и трендов
- Важность прогнозирования предсказуемых пиков
- Возможность использования данных о праздниках для улучшения прогнозов







Оценка производительности моделей

Методология разделения данных:

- Простое разделение на обучающую и тестовую выборки
- Горизонт прогнозирования: 16 дней
- Причина выбора: оптимизация вычислительных ресурсов для нейронных сетей

Реализация экспоненциального сглаживания:

- Обучение 1782 отдельных моделей (для каждой комбинации магазин-семейство)
- Особенность: нулевые прогнозы для серий без продаж в последние 2 недели
- Время обучения и прогнозирования: ~653 секунд

Результаты:

- Средний RMSLE по всем сериям: 0.37411

От базовых к глобальным моделям

Ограничения простых моделей:

- Эффективны для небольших наборов данных
- Хорошо работают с базовыми паттернами
- Имеют фиксированную структуру

Особенности датасета:

- 1782 временных ряда
- Значительная длина рядов
- Наличие ковариат (праздники, промоакции)
- Схожие паттерны продаж между рядами

Глобальные модели глубокого обучения

LSTM (1995):

- Рекуррентная нейронная сеть
- Требует future_covariates
- Обучается на последних 60 образцах

N-HiTS (2022):

- Использует только past_covariates
- Обучается на последних 180 образцах
- Наиболее быстрая в обучении

TFT (2019):

- Поддерживает все типы ковариат
- Обучается на последних 7 образцах
- Наиболее ресурсоемкая модель

Технические особенности реализации

Ограничения вычислений:

- Использование подмножеств временных рядов
- Разное количество образцов для разных моделей

Оптимизация:

- Настройка гиперпараметров через Optuna
- Балансировка времени обучения моделей

Сравнение моделей:

- CatBoost: 365 образцов
- N-HiTS: 180 образцов
- LSTM: 60 образцов
- TFT: 7 образцов

Модель N-HiTS (Neural Hierarchical Interpolation for Time Series)

Особенности работы с ковариатами:

- Поддерживает только past_covariates
- Сдвиг future-known информации на 16 дней назад
- Объединение информации о промоакциях, праздниках и временных признаках

Архитектура и гиперпараметры:

- Лучшие параметры после оптимизации:
  - input_chunk_length: 266
  - num_stacks: 3
  - num_blocks: 3
  - num_layers: 2
  - layer_width: 2^8
  - dropout: 0.01
  - learning_rate: ~0.003

Технические детали реализации:

- Использование PyTorch Lightning
- Ранняя остановка обучения
- Максимум 50 эпох
- Размер батча: 128
- Максимум 180 сэмплов на временной ряд

Результаты:

- RMSLE: 0.43265
- Время обучения и прогнозирования: ~1103 секунды
- Хуже базовой модели экспоненциального сглаживания (0.37411)

Сравнение моделей прогнозирования

Базовая модель:

- Экспоненциальное сглаживание
  - RMSLE: 0.37411
  - Время обучения: 653 секунд
  - Лучший результат среди всех моделей

Глобальные модели:

- N-HiTS:
  - RMSLE: 0.43265
  - Время обучения: 1103 секунды
  - Работает только с past_covariates
- LSTM:
  - RMSLE: 0.55443
  - Время обучения: 1438 секунд
  - Требует future_covariates
  - Худший результат среди всех моделей
- TFT:
  - RMSLE: 0.43226
  - Время обучения: 1091 секунда
  - Поддерживает все типы ковариат
  - Единственная модель с информацией о конкретных сериях

Особенности реализации

Ограничения:

- Неполное использование данных
- Ограниченная настройка гиперпараметров
- Проблемы с памятью для бустинговых моделей

Важные наблюдения:

- Только TFT различает серии между собой
- Остальные модели считают все серии однородными
- Возможный компромисс между объемом данных и схожестью серий

Рекомендации по улучшению:

- Обучение отдельных моделей для каждого семейства продуктов
- Использование более мощного оборудования
- Полная оптимизация гиперпараметров

Спасибо за внимание!