Лабораторная работа №1

Анализ временных рядов

Работу выполнил:
Гомонов Дмитрий Павлович,
студент группы 09-415(1)

Дата: 23.09.2025

ОБЩАЯ ИНФОРМАЦИЯ

Дата создания отчета: 2025-09-23 11:01:30
Исходный размер данных: 421,570 строк
Финальный размер данных: 421,570 строк, 30 столбцов
Удалено дубликатов: 0

СТРУКТУРА ФИНАЛЬНОГО ДАТАСЕТА

Исходные признаки:
• Store - идентификатор магазина
• Dept - идентификатор отдела
• Date - дата продаж
• Weekly_Sales - недельные продажи (целевая переменная)
• IsHoliday - флаг праздничного дня

Созданные временные признаки:
• Year - год
• Month - месяц
• Week - неделя года
• DayOfWeek - день недели (0-6)
• Quarter - квартал

Лаговые признаки:
• Weekly_Sales_lag_1 - продажи с лагом 1 неделя
• Weekly_Sales_lag_7 - продажи с лагом 7 недель
• Weekly_Sales_lag_30 - продажи с лагом 30 недель

Скользящие статистики:
• Weekly_Sales_rolling_mean_7 - скользящее среднее за 7 недель
• Weekly_Sales_rolling_mean_30 - скользящее среднее за 30 недель
• Weekly_Sales_rolling_std_7 - скользящее стандартное отклонение за 7 недель
• Weekly_Sales_rolling_std_30 - скользящее стандартное отклонение за 30 недель

Сезонные признаки:
• IsWeekend - флаг выходного дня
• IsMonthStart - флаг начала месяца
• IsMonthEnd - флаг конца месяца
• IsQuarterStart - флаг начала квартала
• IsQuarterEnd - флаг конца квартала

СТАТИСТИКИ ПО ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

Основные статистики Weekly_Sales:
• Среднее: 15981.26
• Медиана: 7612.03
• Стандартное отклонение: 22711.18
• Минимум: -4988.94
• Максимум: 693099.36
• Коэффициент асимметрии: 3.2620
• Коэффициент эксцесса: 21.4913

Пропущенные значения:
• Общее количество пропусков: 126,905
• Пропуски в Weekly_Sales: 0

ВРЕМЕННОЙ ОХВАТ ДАННЫХ
• Начальная дата: 2010-02-05
• Конечная дата: 2012-10-26
• Общий период: 994 дней
• Количество уникальных магазинов: 45
• Количество уникальных отделов: 81

ДОПОЛНИТЕЛЬНЫЕ ИССЛЕДОВАТЕЛЬСКИЕ ЗАДАНИЯ

1. СРАВНЕНИЕ ЧАСТОТ

Статистики по частотам:
Ежедневные данные:
• Среднее: 15981.26
• Стандартное отклонение: 22711.18

Недельные данные:
• Среднее: 14997.03
• Стандартное отклонение: 22333.65

Месячные данные:
• Среднее: 64818.35
• Стандартное отклонение: 95579.80

2. АНАЛИЗ КРОСС-КОРРЕЛЯЦИИ (CCF)

Топ-5 признаков по максимальной корреляции:

• DayOfWeek: лаг 0, корреляция nan
• Weekly_Sales_rolling_mean_7: лаг 0, корреляция 0.9670
• Weekly_Sales_rolling_mean_30: лаг 0, корреляция 0.9508
• IsWeekend: лаг 0, корреляция nan
• Dept_avg_sales: лаг 0, корреляция 0.7327

3. ТЕСТ НА СТРУКТУРНЫЕ РАЗРЫВЫ

F-статистика (дисперсии): 1.5519
t-статистика (средние): 48.6981
p-value: 0.0000

Первая половина данных:
• Среднее: 17679.89
• Дисперсия: 623834317.57

Вторая половина данных:
• Среднее: 14282.63
• Дисперсия: 401993125.25

4. АНАЛИЗ ОСТАТКОВ ПОСЛЕ ДЕКОМПОЗИЦИИ

Тесты стационарности остатков:
ADF тест:
• Статистика: -101.4675
• p-value: 0.0000
• Результат: Стационарны

KPSS тест:
• Статистика: 0.0002
• p-value: 0.1000

Тест Льюнга-Бокса на автокорреляцию остатков:
• Статистика: 337796.5821
• p-value: 0.0000

5. СЕЗОННАЯ ДЕКОМПОЗИЦИЯ С ИЗМЕНЯЕМЫМ ПЕРИОДОМ

Результаты для разных периодов:

Период 7 дней:
• Сила тренда: 0.9453
• Сила сезонности: 0.9453
• Дисперсия остатков: 28201124.4118

Период 30 дней:
• Сила тренда: 0.8998
• Сила сезонности: 0.8998
• Дисперсия остатков: 51671984.2851

Период 52 дней:
• Сила тренда: 0.8741
• Сила сезонности: 0.8741
• Дисперсия остатков: 64954786.2670

Период 365 дней:
• Сила тренда: 0.6266
• Сила сезонности: 0.6266
• Дисперсия остатков: 192609959.2286

РЕКОМЕНДАЦИИ ДЛЯ ДАЛЬНЕЙШЕГО АНАЛИЗА

1. Моделирование временных рядов:
• Использовать ARIMA/SARIMA модели с учетом сезонности
• Применить Prophet для прогнозирования
• Рассмотреть LSTM нейронные сети

2. Feature Engineering:
• Создать дополнительные лаговые признаки
• Добавить взаимодействия между признаками
• Рассмотреть полиномиальные признаки

3. Валидация:
• Использовать временную валидацию (Time Series Split)
• Применить walk-forward validation
• Оценить качество на разных временных горизонтах

ЗАКЛЮЧЕНИЕ

Данные успешно обработаны и готовы для моделирования. Создано 25 дополнительных признаков, что значительно расширяет возможности для анализа и прогнозирования временных рядов.

Финальный датасет сохранен в файле final_dataset.csv и готов для использования в машинном обучении.