Дипломная работа по программе «Аналитик данных»

Прогнозирование спроса в онлайн ритейле

Александр Сипко
Группа: AML-42
Руководитель: Даниил Корбут

1. Содержание

1. Введение и постановка задачи

Прогнозирование спроса — это ключевая задача для онлайн-ритейла, оказывающая значительное влияние на управление запасами, логистику и планирование продаж.

> *«Точные прогнозы позволяют снизить издержки, дефицита товаров и улучшить обслуживание клиентов, что даёт компаниям конкурентное преимущество, повышая эффективность операций и прибыль.»*

Стейкхолдеры:

Коммерческий директор – повышение прибыли.

Менеджеры по закупкам – минимизация издержек, планирование поставок.

Логистический отдел – сокращение затрат и управление запасами.

Маркетологи – планирование активностей и промо-компаний.

IT-отдел – поддержка инфраструктуры для сбора и обработки данных.

Финансовый отдел – бюджетирование и оценка рентабельности.

2. Описание данных и их особенностей их подготовка для обучения моделей

Источник данных:

Выгруза Excel содержащая данные о продажах за 3 года

- InvoiceNo — номер заказа
- StockCode — номер товара
- Description — название товара
- Quantity — колличество
- InvoiceDate — дата и вермя транзакции
- UnitPrice — стоимость
- CustomerID — id клиента
- Country — страна (нет данных о значении поля)

Пред обработка данных:

- проверены пропуски в данных
- убраны пустые значения (замена на общее значение)
- для целей кластеризации данных, были выбраны только уникальные наименования товаров
- для целей предсказания временных рядов, данные были сгруппированы по категориям (кластерам) по недельным интервалам и суммой по количеству продаж.

3. Описание возможных решений и планируемая архитектура

Можно выдвинуть несколько гипотез, которые описывают факторы влияющие на продажи:

1. Гипотеза о предсказуемости трендов на оснвоании истории продаж

2. Гипотеза о сезонности продаж

Для проверки данных гипотез необходимо катетеризировать данные, это можно сделать с помощью Кластеризация данных.

Это позволит выявить логически сгруппированные категории товаров и клиентов, что, в свою очередь, поможет более точно анализировать поведение потребителей и выявлять скрытые зависимости.

4. Описание обучения

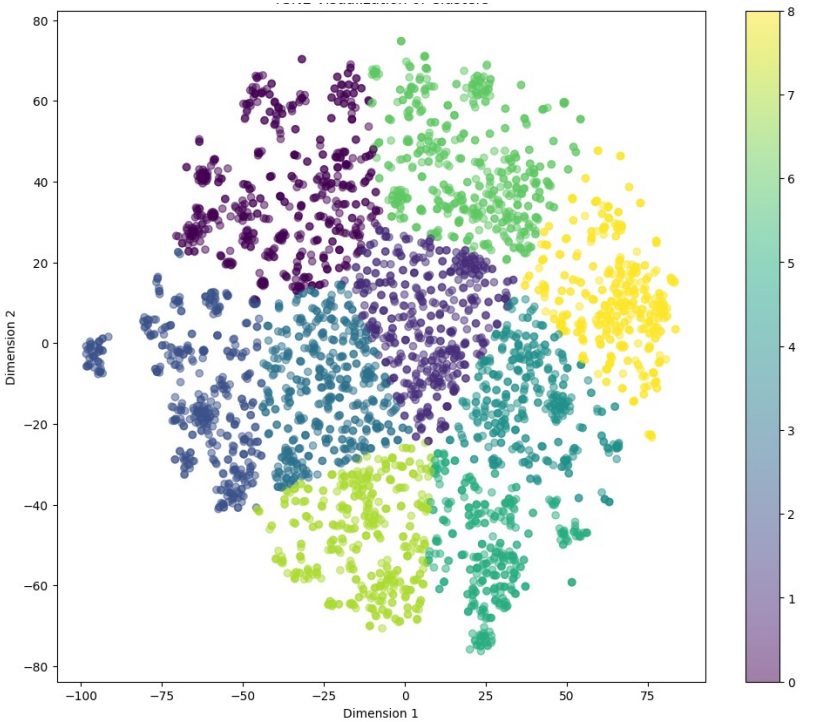
*Использование текстовых моделей позволило преобразовать наименование товаров в эмбеддинги, а затем, используя **K-Means**, разделить товары на кластеры. Для оценки качества кластеризации использовалась оценка **Silhouette Score** (значение от -1 до 1, где 1 указывает на хорошую кластеризацию).*

*В работе применялись предобученные модели, такие как **paraphrase-MiniLM-L6-v2**, **all-MiniLM-L6-v2** и другие, для представления товаров в векторном пространстве.*

Best model: *paraphrase-MiniLM-L6-v2*

Best number of clusters (K): *9*

Best Silhouette Score: *0.08163392543792725*



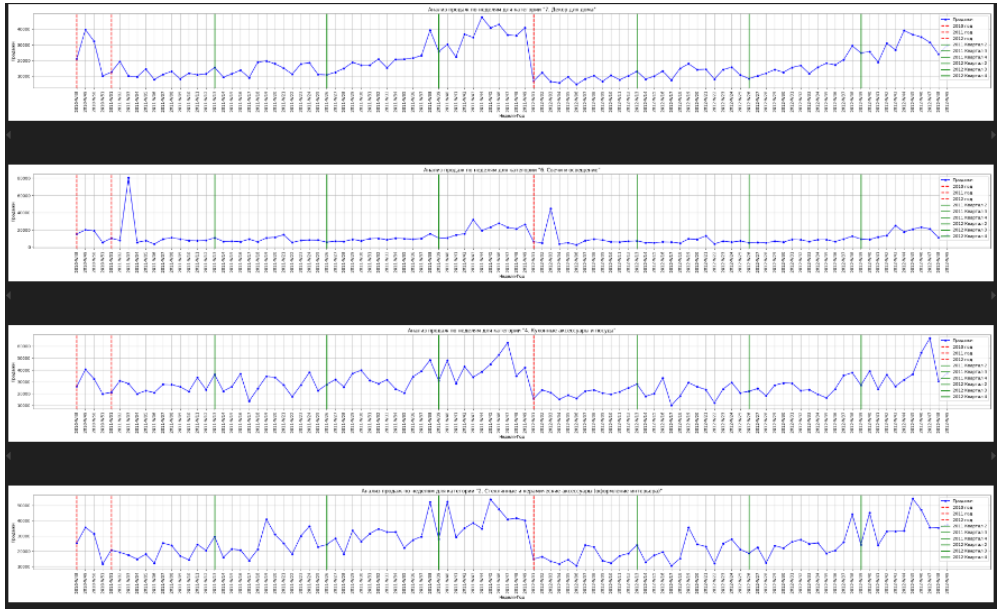
4. Описание обучения

«Прогнозирование спроса в онлайн ритейле»

** нумерация фото слева на право, начиная с верхнего ряда*

1. Стилизованные вещи интерьера с уникальным дизайном
2. Стеклянные и керамические аксессуары (оформление интерьера)
3. Часы и рождественские украшения
4. Кухонные аксессуары и посуда
5. Игрушки или детские товары
6. Свечи и освещение
7. Декор для дома
8. Праздничные украшения
9. Браслеты бижутерия

Временные ряды кластеров товаров были разбиты с шагом в 1 неделю для визуальной оценки и определения наличия сезонности. Ниже представлены примеры построения графиков для визуального анализа.



Временные ряды отражают сезонность для различных категорий.

Также отмечается короткий диапазон временного ряда, необходимый для формирования устойчивого сезонного паттерна.

** полный перечень графиков представлен в приложенном к роекту *.ipynb файле*

Для прогнозирования временного ряда использовались модели Prophet, ARIMA. Была сформирована функция для определения наилучшей модели и оптимальных гиперпараметров. Для этого применялись следующие метрики:

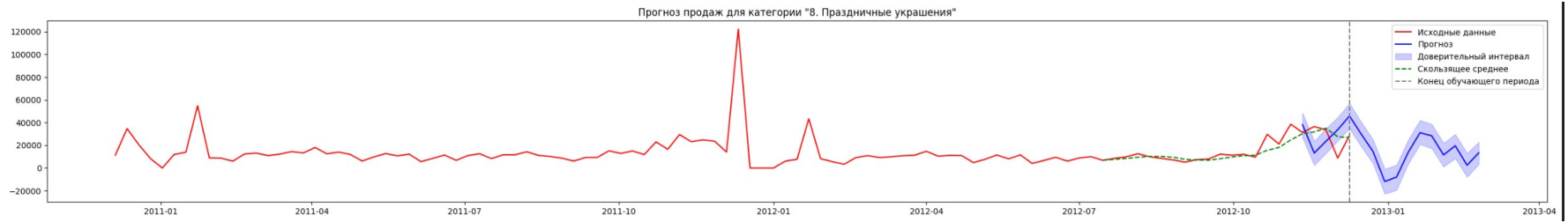
Для оценки моделей временных рядов использовались метрики:

- *Среднеквадратическая ошибка (MSE)*
- *Средняя абсолютная ошибка (MAE)*
- *Коэффициент детерминации (R^2)*
- *Средняя абсолютная процентная ошибка (MAPE)*
- *Сравнение с базовой линией, построенной на основании скользящего среднего.*

Для прогнозирования продаж была применена модель Prophet, которая показала высокую точность на тестовых данных. Построенные прогнозы охватывали период в 12 недель, с использованием данных последних 4 недель в качестве исторического контекста.

Примеры категории товаров демонстрирующие улучшение метрик, по сравнению с бейзлайн

8. Праздничные украшения



3. Часы и рождественские украшения

** полный перечень графиков представлен в приложенном к роекту *.ipynb файле*

Оценка качества прогнозирования продаж для каждой категории

	Category	Baseline MSE	Baseline MAE	Baseline R^2	Prophet MSE	Prophet MAE	Prophet R^2
0	7. Декор для дома	1.890789e+07	3342.723684	0.684877	1.848973e+07	3315.071250	0.729523
1	6. Свечи и освещение	1.127422e+07	2185.328947	0.643953	2.253267e+07	3962.301665	0.320746
2	4. Кухонные аксессуары и посуда	8.003183e+07	7222.868421	0.426662	6.589700e+07	6478.019519	0.495356
3	2. Стеклянные и керамические аксессуары (оформ...	5.283409e+07	5454.565789	0.422399	5.362271e+07	6075.268235	0.477214
4	8. Праздничные украшения	4.353351e+07	4145.197368	0.654384	1.095873e+08	7927.023640	0.076631
5	1. Стилизованные вещи интерьера с уникальным д...	1.364643e+07	3223.578947	0.448022	2.097828e+07	3840.286861	0.121624
6	5. Игрушки или детские товары	5.240364e+06	1786.486842	0.547259	9.362651e+06	2493.913050	0.185415
7	9. Браслеты бижутерия	4.651652e+05	509.197368	0.586350	1.076343e+06	893.679541	-0.051781
8	3. Часы и рождественские украшения	7.531862e+07	5792.407895	0.555005	3.908192e+07	4488.381398	0.761454
9	11. Прочие	2.732996e+05	426.539474	0.147828	1.358731e+06	1014.865336	-3.644213

Показатель с высоким R2 и низким MSE и MAE, а так же показавшим улучшение показателей по сравнению с байзлайн, могут быть использованы для принятия управленческих решений в бизнесе, другие категорию требуют более аккуратного использования или увеличения диапазона данных для прогнозирования и повышения качества модели

5. Заключение с выводами и планами на дальнейшее развитие (практическое применение)

В результате работы по прогнозированию спроса были подтверждены гипотезы о предсказуемости трендов и сезонности продаж. Анализ данных с использованием моделей Prophet и ARIMA выявил четкую сезонность в категориях, таких как праздничные украшения и кухонные аксессуары.

Проведенная кластеризация товаров позволила выделить 10 категорий, что улучшило качество прогнозов за счет работы с более однородными группами данных. Полученные результаты уже применяются для оптимизации запасов и планирования закупок, что снижает издержки и повышает уровень обслуживания клиентов.

В будущем планируется улучшение моделей путем увеличения объема данных и использования более сложных алгоритмов машинного обучения, что повысит точность прогнозов и адаптирует бизнес-процессы к изменениям на рынке.

6. Список литературы

1. Hyndman, R.J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.
2. Box, G.E.P., Jenkins, G. M., & Reinsel, G. C. (2015). Time Series Analysis: Forecasting and Control. Wiley.
3. Taylor, S.J., & Letham, B. (2018). Forecasting at Scale. The American Statistician, 72(1), 37-45.
4. Shumway, R.H., & Stoffer, D. S. (2017). Time Series Analysis and Its Applications: With R Examples. Springer.
5. Prophet Documentation. (n.d.). Forecasting at Scale.
6. Chen, J., & Liu, L. (2020). A Review of Time Series Forecasting Methods. Journal of Statistics and Management Systems, 23(2), 245-260.
7. Kourentzes, N., & Petropoulos, F. (2020). Forecasting with Exponential Smoothing: A Review. International Journal of Forecasting, 36(1), 1-17.
8. Datta, A., & Khamis, M. (2021). Machine Learning for Time Series Forecasting. Journal of Business Research, 124, 1-12.

Спасибо за внимание!