АНАЛИЗ И ПРОГНОЗИРОВАНИЕ ДИНАМИКИ ПРОДАЖ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Голубева Л. Л., Мурашко А. С.

Белорусский государственный университет, Минск, Беларусь,
e-mail: goloubeva@bsu.by, sashamurashko@yahoo.com

В работе рассматриваются вопросы выявления внутренних закономерностей больших объемов данных методами машинного обучения для выполнения анализа и прогнозирования динамики продаж на примере исследования статистических данных о продажах парфюмерно-косметических товаров в магазинах торговой сети Rossmann.

Прогноз продаж является основой для планирования закупочной деятельности, страховых запасов, работы складской службы, привлечения или увольнения персонала, загрузки производственных мощностей и маркетинговых акций по стимулированию спроса [3, с. 135]. Ошибки прогноза приводят к дополнительным издержкам, в частности, увеличение ошибки прогноза на 1 % приводит к увеличению оборотных средств предприятия, расходуемых на страховые запасы, примерно на 2–3 % [3, с. 135–136].

Dirk Rossmann GmbH – одна из крупнейших торговых сетей в Европе, насчитывающая около 56 300 сотрудников и более 4000 магазинов косметики, товаров для дома, продуктов питания с ассортиментом более 12000 наименований. В 2020 году оборот компании в восьми европейских странах составил более 10 миллиардов евро. На продажи в магазинах влияют многие факторы, в том числе локация, промоакции, конкуренция, сезонность, государственные праздники и школьные каникулы. Менеджерам магазинов Россманн поручено прогнозировать ежедневные продажи на срок до шести недель.

Объектом исследования является динамика продаж данной торговой сети. Цель работы заключается в выявлении внутренних закономерностей, присущих исследуемому временному ряду, и прогнозирование его будущих значений с помощью различных методов машинного обучения. Исходные данные содержат 1 017 209 записей в тренировочном наборе и 41 088 записей в тестовом наборе. Ежедневные данные предоставлены за период с января 2013 по июль 2015 (2 года 7 месяцев). Дополнительно предложен набор данных о каждом из 1115 магазинов, в которых осуществлялись продажи. Задача – спрогнозировать величину ежедневных продаж в каждом из этих магазинов на указанный период [5].

Табл. 1 содержит описание данных, предоставленных в тестовом и тренировочном наборах.

Табл. 1. Структура тренировочного и тестового набора данных

Название столбца	Описание	Тип данных
Store	Уникальный идентификатор магазина	int
DayOfWeek	День недели совершения продаж (1 – понедельник, …, 7 – воскресенье)	int
Date	Дата совершения продаж (год-месяц-день)	object

Sales	Величина продаж для указанной даты (указана только для тренировочного набора, так как эту величину и требуется спрогнозировать)	int
Customers	Число покупателей в указанную дату	int
Open	Индикатор работы магазина в указанную дату (0 – закрыт, 1 – открыт)	int
Promo	Индикатор проведения промоакции в указанную дату (0 – нет, 1 – да)	int
StateHoliday	Параметр, указывающий является ли указанная дата праздником и тип праздника: a – государственный праздник, b – пасхальные выходные, c – рождественские каникулы, 0 – нет	object
SchoolHoliday	Индикатор школьных каникул в указанную дату (0 – нет, 1 – да)	int

Таблица 2 содержит описание данных о магазинах.

Табл. 2. Структура набора данных о магазинах

Название столбца	Описание	Тип данных
Store	Уникальный идентификатор магазина	int
StoreType	Тип магазина: a, b, c, d (пояснения типам не даны)	object
Assortment	Тип ассортимента для указанного магазина: a – базовый ассортимент, b – доступен дополнительный ассортимент, c – расширенный ассортимент	object
Competition-Distance	Расстояние в метрах до ближайшего магазина-конкурента	float
Competition-Open-Since	Приблизительная дата открытия ближайшего магазина-конкурента	float
Promo2	Идентификатор участия магазина в продолжительной промоакции (0 – нет, 1 – да)	int
Promo2SinceYear Promo2SinceWeek	Год и календарная неделя начала участия магазина в продолжительной промоакции	float
PromoInterval	Месяцы начала очередных раундов промоакции Promo2. Например, объект “Feb,May,Aug,Nov” означает, что новые раунды промоакции для указанного магазина начинаются в феврале, мае, августе и ноябре каждого года	object

Первоначально в работе были проведены предварительный анализ и предобработка данных (создание новых признаков, логарифмирование величин, обработка пропущенных значений, приведение данных к удобному формату). **В результате был получен единый набор данных с восемнадцатью признаками, включающими как признаки из исходного тренировочного набора, так и данные о магазинах, а также новые, созданные вручную для повышения точности прогнозов, признаки.**

Далее был выполнен анализ внутренних зависимостей между параметрами исследуемого временного ряда (компонентный анализ [1], тест на стационарность),

установлена его стационарность и получены графики компонент временного ряда, выявлено наличие сезонности и его автокорреляционной функции. Эти действия позволили определить параметры модели ARIMA, которая является классической моделью, применяемой в анализе временных рядов. Были реализованы модели SARIMA и обобщенные аддитивные модели (библиотека Prophet [2]), применяемые для анализа временных рядов, а также общеприменимые методы машинного обучения, такие как: линейная регрессия, случайные леса, градиентный бустинг и их комбинация с помощью стекинга. Работа включала исследование информативности признаков и подбор параметров для моделей случайного леса и градиентного бустинга. Для оценки качества моделей использовалась RMSPE метрика – значение относительной среднеквадратичной ошибки.

Ниже приведена сводная таблица (см. табл. 3), содержащая названия реализованных моделей и соответствующие им значения относительной среднеквадратичной ошибки (RMSPE) [4].

Табл. 3. Сравнение результатов прогнозирования по метрике RMSPE

Модель	RMSPE
SARIMA (p=0, d=1, q=1)□(P=0, D=1, Q=1, m=12)	0.28451
Prophet	0.37456
Линейная регрессия	0.25933
Случайный лес	0.11821
Градиентный бустинг	0.09941
Стекинг (Случайный лес + Линейная регрессия + Градиентный бустинг (500) -> Линейная регрессия)	0.10007

Литература

1. Лернер, Э.Ю. Экономическое моделирование и прогнозирование на компьютере. Учебное пособие / Э.Ю. Лернер, О.А. Кашина. – Казанский Государственный Университет, 2001. [Электронный ресурс] Режим доступа: http://kek.ksu.ru/eos/Model/Content.htm . Дата доступа: 19.05.2018.

2. Taylor, S.J. Forecasting at scale / S.J. Taylor, B. Letham. – PeerJ Preprints, 2017 [Electronic resource] Mode of access: https://doi.org/10.7287/peerj.preprints.3190v2 Date of access: 02.02.2018.

3. Егоров, А.М. Алгоритм правильного прогнозирования продаж / Управление продажами. – ООО «ИД Гребенников», 2012, №03 (64). – С.134-144.

4. Мурашко, А.С. Анализ временных рядов методами машинного обучения: магистерская диссертация: специальность 1-31 81 08 «Компьютерная математика и системный анализ». – Минск, БГУ, Механико-математический факультет, Кафедра дифференциальных уравнений и системного анализа; науч. рук. Голубева Л.Л, 2018. – 77 с. [Электронный ресурс] Режим доступа: http://elib.bsu.by/handle/123456789/202712 . Дата доступа: 13.06.2018.

5. Kaggle- Rossmann Store Sales [Electronic resource] / Mode of access: https://www.kaggle.com/c/rossmann-store-sales/. – Date of access: 21.09.2017.