

ПОСТРОЕНИЕ МОДЕЛИ ПРОГНОЗИРОВАНИЯ ОТМЕНЫ ЗАКАЗА В ИНТЕРНЕТ-МАГАЗИНЕ (НА ПЛАТФОРМЕ LOGINOM)

Аннотация. В статье представлено построение модели прогнозирования отмены заказа в интернет-магазине, основанной на градиентном бустинге. Предварительно проведен первичный анализ данных, использованы методы предварительной обработки, сконструированы новые признаки на основе исходных. В работе использован метод оптимального квантования и оценена значимость признаков для прогнозирования.

Ключевые слова: конструирование признаков, градиентный бустинг, прогнозирование отмены заказа, оптимальное квантование.

Введение. Крупные онлайн-ритейлеры ежедневно обрабатывают огромное количество заказов, которое варьируется до нескольких миллионов транзакций в год. Некоторая часть из них отменяется, вследствие чего компания несет значительные финансовые убытки. Согласно исследованиям, доля отмен в онлайн торговле может достигать 20 — 40% в зависимости от категории товаров.

Проблема отмен заказов имеет серьезные экономические последствия для бизнеса: возникают ситуации нехватки запасов, проблемы с логистикой, что приводит к потере дохода [1]. В некоторых случаях компании вынуждены нести дополнительные расходы на возврат уже отправленных товаров после отмены заказа. Например, американский бизнес-журнал Fast Company провел исследование, которое показало, что обработка возврата товаров заказа может обходиться в 59% от первоначальной стоимости заказа [2].

С каждым годом онлайн-торговля стремительно развивается. Так, в России в 2023 году сумма транзакций достигла 6,2 трлн. российских рублей, увеличившись на 28% по сравнению с предыдущим годом [3]. Вместе с этим влияние отказов от заказов становится все более значительным: снижается прибыль, появляется необходимость в дополнительных ресурсах. В долгосрочной перспективе это может снизить конкурентоспособность компании. Например, уже в 2023 году общая сумма убытков из-за изменения заказов (включая отмены) составила 1,77 трлн. долларов США по всему миру [4].

Современные исследования показывают, что применение методов машинного обучения позволяет минимизировать убытки за счет прогнозирования вероятности отмен. Это, в свою очередь, помогает оптимизировать бизнес-процессы и снижать затраты на логистику и обработку заказов. В результате чего компании смогут сохранить часть своих ресурсов и повысить эффективность работы [5]. Более того, ансамблевые методы машинного обучения, в том числе градиентный бустинг, обеспечивают высокую точность прогнозирования в условиях непредсказуемости [6]. В исследовании крупной американской компании PVH использование ML-моделей позволило более точно прогнозировать возвраты и отмены заказов [7].

На фоне растущего интереса к практическому применению этих технологий особенно актуальным становится анализ конкретных случаев из российской розничной онлайн-торговли. В данном исследовании изучался набор из более 300 тыс. заказов российского онлайн-

гипермаркета продуктов питания «Утконос», представлявший на конкурсе SAS Data Hack Platypus в 2019 г.

Проблема исследования. Интернет–ритейлер «Утконос» также, как и любой другой онлайн-магазин, сталкивается с существенными финансовыми потерями из-за отмен заказов. Затраты на упаковку, транспортировку и хранение, понесенные до момента отмены, не компенсируются, что напрямую влияет на прибыльность. Инструмент, способный на ранних этапах жизненного цикла заказа спрогнозировать его потенциальную отмену позволит компании заранее принимать управленческие решения: например, приоритизировать обработку устойчивых заказов, изменить условия доставки, провести дополнительную верификацию или же отправить клиенту уведомление, влияющее на вероятность завершения покупки.

Прогнозирование отмены заказов — это междисциплинарная задача, включающая в себя анализ пользовательского поведения, динамики товарооборота, а также логистических параметров. Исследования показывают, что применение методов машинного обучения в данной области может существенно повысить точность прогнозов по сравнению с традиционными статистическими подходами.

Ключевым требованием к разрабатываемому решению является его способность учитывать широкий спектр факторов: к примеру, характеристики товара, историю покупок, особенности клиента, а также временные параметры заказа [2]. Таким образом, задача заключается не только в построении предсказательной модели, но и в выявлении важных для этой модели признаков.

Материалы и методы. Основание и исходные данные для разработки темы — транзакционные данные компании «Утконос» за 2018 год, изначально имевшие около 9 млн строк. Исходный набор данных содержал следующие признаки, которые в процессе исследования были разделены на 5 категорий:

1. **Идентификаторы** — включают уникальные коды, позволяющие различать объекты в системе: ID клиента, ID заказа, ID товара, ID группы товаров. Эти признаки в последствии не использовались как входные для модели прогнозирования, и были использованы для формирования истории заказов по каждому клиенту.

2. **Временные характеристики** — отражают временные параметры: дата заказа, дата доставки, интервал доставки.

3. **Параметры доставки** — описывают логистические особенности выполнения заказа: тип доставки, кластер доставки.

4. **Характеристики заказа** — содержат информацию, связанную с составом и обработкой заказа: канал сбыта, количество позиции, число редактирований заказа, признак предоплаты.

5. **Целевая переменная** — представляет собой бинарную метку, указывающую на отмену заказа.

Реализация аналитического процесса осуществлялась в программной среде Loginom, представляющей собой low-code платформу, предназначенную для визуального проектирования моделей без необходимости написания программного кода. На рис. 1 представлен сценарий Loginom для задачи прогнозирования отказов в онлайн-магазине «Утконос».

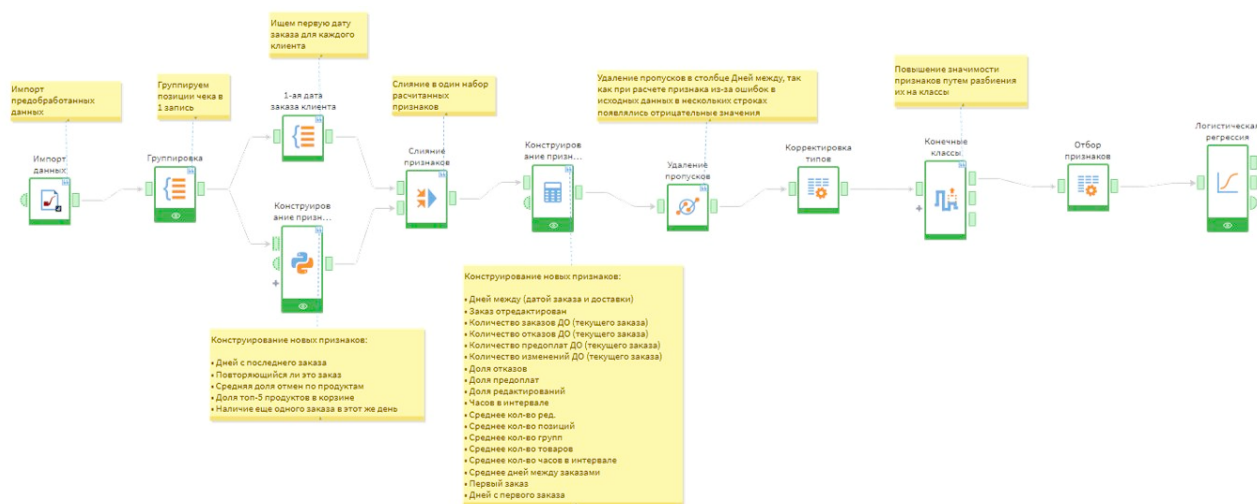


Рис. 1. Сценарий Loginom

Работа с данными в Loginom началась с первичного изучения и преобработки данных. На первом этапе подготовки данных была выполнена обработка пропусков: был исключен из дальнейшего рассмотрения признак Кластер доставки, содержащий 63% пропущенных значений. Такое количество пропусков делает этот признак непригодным для обучения модели [8]. Для ряда признаков (ID товара, ID группы товаров, количество товаров) были удалены строки с пропущенными значениями. В процессе изучения данных было выявлено, что признаки Дата заказа и Дата доставки имеют некорректные значения в некоторых строках: дата доставки раньше, чем дата заказа. Небольшое количество таких записей позволило также решить эту проблему путем удаления некорректных строк — во избежание лишнего шума при обучении модели прогнозирования. На рис. 2 представлен сценарий преобработки данных, который включает в себя следующие компоненты: импорт исходных файлов, исключение «лишнего» поля, удаление пропусков, и сохранение обработанных данных в промежуточном файле.

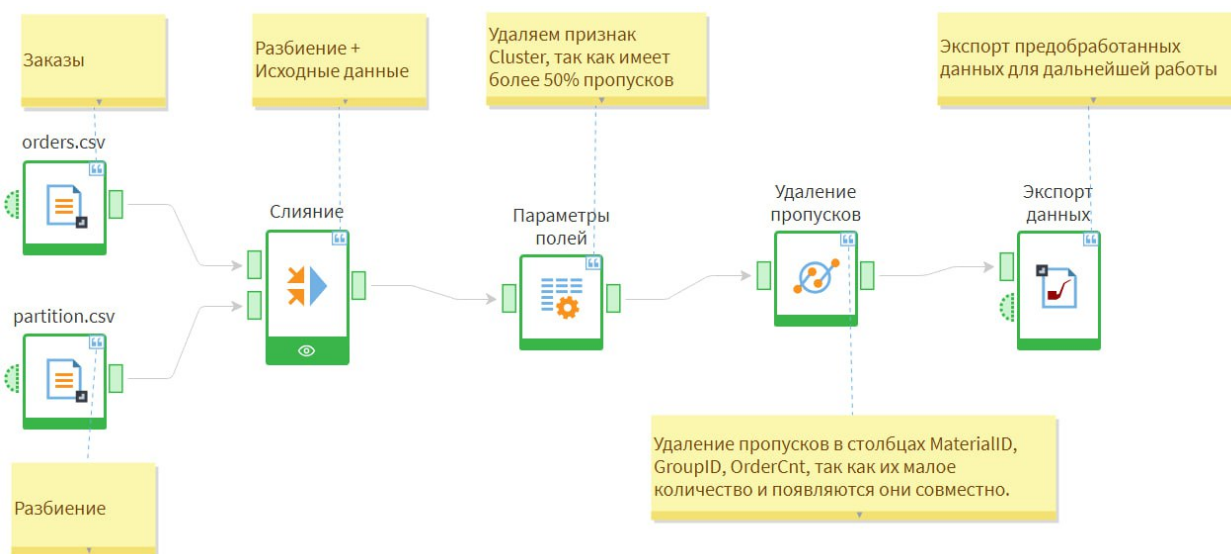


Рис. 2. Сценарий преобработки

Далее были сгруппированы записи, относящиеся к одному заказу.

После изучения признаков исходного набора данных становится понятно, что сами по себе первичные признаки совершенно непригодны для целей прогнозирования: как даты, так и характеристики конкретного заказа или способы доставки могут лишь в очень незначительной степени влиять на вероятность отмены. Поэтому важнейшей задачей исследования было конструирование из первичных признаков новых, более значимых признаков.

Для построения модели прогнозирования применялись алгоритмы: логистическая регрессия и градиентный бустинг. Несмотря на возрастающий интерес к вычислительно сложным подходам, классические методы сохраняют высокую практическую значимость благодаря прозрачности интерпретации результатов и невысоким требованиям к объему обучающей выборки [9].

Для улучшения качества прогнозирования дополнительно применялся оптимальный биннинг непрерывных переменных (использовался встроенный компонент Logitom «Конечные классы»).

Результаты. На этапе конструирования признаков был сформирован комплексный набор характеристик, направленных на описание истории поведения клиента (на основе данных о предыдущих заказах этого клиента) и структуры заказа. Основной сложностью являлось то, что при их создании важно было учитывать хронологическую последовательность каждого заказа, чтобы избежать «утечки данных». Поэтому при расчете признаков данные были упорядочены по дате заказа, номеру заказа и клиенту, и рассчитывались динамически: для каждого заказа использовались данные только предыдущих периодов.

Для повышения точности модели было важно сформировать целостное представление о поведении клиента. Поэтому все полученные признаки разделены на две ключевые группы: информационное досье клиента и признаки, не связанные с его историей заказов.

1. Информационное досье клиента.

Данная группа признаков формируется на основе истории заказов и отражает как общее поведение клиента, так и контекстные особенности его текущей активности.

1. Доля отмен — отношение количества отмененных заказов к общему числу заказов клиента.

2. Среднее количество редактирований заказа — отражает склонность клиента вносить изменения в заказы.

3. Количество заказов — общая активность клиента в сервисе.

4. Среднее количество позиций в заказе — характер закупок: точечные или комплексные.

5. Среднее количество дней между заказами — регулярность покупок.

6. Доля предоплат — доля заказов, оплаченных заранее; этот признак скорее отражает доверие к сервису.

7. Доля редактирований — доля отредактированных заказов среди всех; признак нестабильности потребностей.

8. Средняя доля отмен по товарам — отражает склонность к отменам определенных типов товаров.

9. Признак повторяющегося заказа клиента — наличие схожих по составу заказов в прошлом.

10. Признак первого заказа — индикатор нового клиента.
11. Наличие более одного заказа на один и тот же день — сигнал возможной отмены заказа, поскольку это может указывать на дублирование или изменение клиентом своего заказа.
12. Количество дней с первого заказа — длина жизненного цикла клиента, некий «возраст» клиента.
13. Количество дней с последнего заказа — временной разрыв, потенциально указывающий на снижение интереса клиента.
14. Доля топ-5 продуктов в корзине клиента — доля самых популярных товаров клиента, добавленных в корзину.
15. Среднее количество товаров в заказе — средний размер заказов по числу позиций.

2. Признаки, не зависящие от истории заказов клиента.

Эта группа признаков особо информативна для новых клиентов, которые еще не сформировали историю своих заказов. Более того, эта группа была выделена специально для подобных случаев, так как модель часто ошибается на первых заказах клиентов.

1. Дней между датой заказа и доставки — отражает интервал времени между оформлением заказа и его фактической доставкой.
2. Признак редактирования заказа — бинарная переменная, указывающая, вносились ли изменения в заказ после его оформления.
3. Часовой интервал доставки — предпочтения по времени получения заказа.

Было выявлено, что некоторые из вышеупомянутых признаков продемонстрировали более высокую информационную ценность после применения оптимального биннинга, выполненного с помощью компонента Logitom «Конечные классы». Этот компонент автоматически подбирает наилучшее разбиение непрерывной переменной на интервалы и рассчитывает информационную значимость признака после разбиения (Information Value, IV) по следующей формуле [10]:

$$IV = \sum_{i=1}^k \left\{ \left(\frac{N_i}{N} - \frac{P_i}{P} \right) \times \ln \frac{F_i}{F} \right\}, \quad (1)$$

где N_i — количество «отрицательных» объектов (в данном случае, неотмененных заказов) в i -м интервале; P_i — количество «положительных» объектов (в данном случае, отмененных заказов) в i -м интервале; N , P — общее количество отрицательных и положительных объектов соответственно; $F_i = P_i + N_i$; $F = P + N$.

В состав признаков, значимость которых улучшилась после конечных классов, вошли:

1. Дней между датой заказа и доставки — *средняя значимость*,
2. Количество заказов — *средняя*,
3. Количество отмен — *высокая*,
4. Доля отмен — *высокая*,
5. Среднее количество редактирований заказа — *средняя*,
6. Среднее количество дней между заказами — *средняя*,
7. Количество дней с последнего заказа — *средняя*.

Далее на основе сформированного набора признаков были обучены две модели бинарной классификации: логистическая регрессия и градиентный бустинг. Для оценки качества моделей использовалась метрика AUC-ROC (Area Under the Receiver Operating Characteristic Curve), которая является стандартным показателем для задач классификации с несбалансированными классами.

Результаты продемонстрировали хорошее качество обеих моделей по экспертной шкале от Loginom: значение AUC-ROC на тестовых данных составило 0,77 для логистической регрессии и 0,78 для градиентного бустинга, что соответствует хорошему уровню разделяющей способности [11].

Графическое представление AUC-ROC кривых для логистической регрессии и градиентного бустинга приведено на рис. 3 и 4 соответственно. Оба графика демонстрируют гладкие кривые, что свидетельствует о стабильности моделей и отсутствии переобучения на обучающих данных. Гладкость кривой также указывает на уверенное ранжирование объектов по вероятности принадлежности к положительному классу. Целевой признак имел сильный дисбаланс классов, примерно 6,7% событий и 92,3% не-событий, поэтому пороговое значение для модели логистической регрессии было опущено до 0,23, чтобы повысить ее чувствительность к редкому классу.

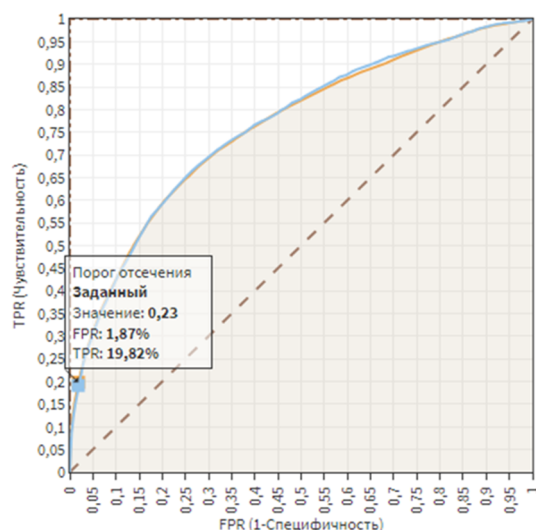


Рис. 3. Графическое представление AUC-ROC кривой для логистической регрессии

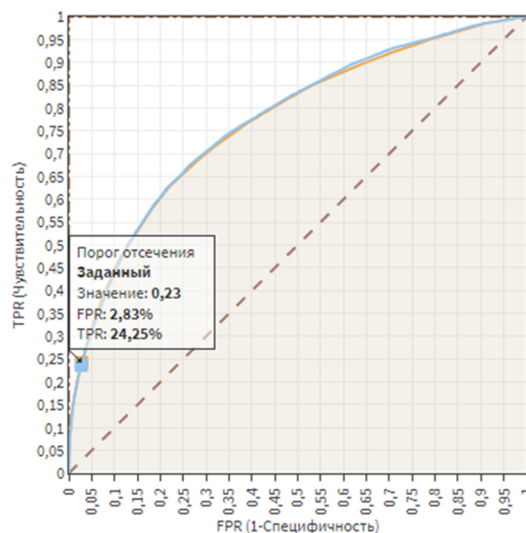


Рис. 4. Графическое представление AUC-ROC кривой для градиентного бустинга

Заключение. В данной работе разработана модель прогнозирования отмены заказов в интернет-магазине на базе платформы Loginom. Исследование охватывало полный цикл построения модели: от анализа исходных данных и конструирования признаков, до обучения алгоритмов и оценки их качества. Особое внимание было уделено созданию комплексного набора характеристик, отражающих поведение клиента и структуру заказа, а также использованию методов оптимального биннинга для повышения интерпретируемости и качества модели.

Результаты экспериментов показали, что как логистическая регрессия, так и градиентный бустинг способны эффективно решать поставленную задачу: полученные значения AUC-ROC (0,77 и 0,78 соответственно) подтверждают высокое качество прогнозирования. Гладкие

ROC-кривые также свидетельствуют о стабильности моделей и их способности надежно ранжировать заказы по вероятности отмены.

Практическая значимость работы заключается в возможности интеграции построенной модели в бизнес-процессы интернет-ритейлера для реализации предиктивных мер управления рисками отмен. Это может включать приоритизацию заказов или изменение условий доставки и оплаты. В условиях высокой конкуренции и необходимости оптимизации логистических операций внедрение подобных решений способствует снижению издержек и повышению общей операционной эффективности.

Дальнейшие исследования могут быть направлены на расширение набора признаков, включая поведенческие и контекстные характеристики клиентов (например, частота визитов, время активности, способ оформления заказа). Кроме того, перспективным является использование внешних данных, таких как праздничные дни или маркетинговые кампании, которые могут влиять на поведение пользователей.

СПИСОК ЛИТЕРАТУРЫ

1. Son J., Kang J. H., Jang S. The effects of out-of-stock, return, and cancellation amounts on the order amounts of an online retailer // *Journal of Retailing and Consumer Services*. — 2019. — Vol. 51. — С. 421–427. — ISSN 0969–6989. — DOI: 10.1016/j.jretconser.2019.02.008.
2. Sarı Z. S., Taşkapı B., Dağdaş G., Akay M. F. Development of machine learning-based sales cancellation/return forecasting models for the e-commerce industry // *Proceedings of the International Conference on Advances in Electrical–Electronics Engineering and Computer Science*, 9–10 November 2024, Ankara, Türkiye. — 2024. — С. 17–21. — ISSN 2687–5527. — DOI: 10.36287/sets.19.4.022.
3. Ситюков А. Как поменялся рынок интернет-торговли в 2023 году. Инфографика // РБК. — 2024. — 12 февр. — URL: https://www.rbc.ru/technology_and_media/12/02/2024/65c64b479a794726261d49f7 (дата обращения: 01.05.2025).
4. IHL Group. Retail Inventory Distortion Study — The Good, The Bad, The Ugly — 2023. — Franklin, TN: IHL Group, 2023. — URL: <https://www.ihlservices.com/product/retail-inventory-distortion-study-the-good-the-bad-the-ugly-2023/> (дата обращения: 02.05.2025).
5. Chidroop I., Moharir M. Predicting the propensity of order cancellation in the ecommerce domain // *International Journal of Research in Engineering, Science and Management*. — 2020. — Vol. 3, № 6. — ISSN 2581–5792. — URL: <https://www.ijresm.com> (дата обращения: 09.04.2025).
6. Adıguzel Tuylu, Ayse & Eroglu, Ergun. The prediction of product return rates with ensemble machine learning algorithms // *Journal of Engineering Research*. — 2022. — DOI: <https://doi.org/10.36909/jer.13725>
7. Bhulai S., Gorissen B. L., Sluis N. Predicting e-commerce orders returns and cancellations using machine learning [магистерская диссертация] — Amsterdam : Vrije Universiteit Amsterdam, Faculty of Sciences, Business Analytics, 2017. — 36 л.
8. Kong W., Hui H. W. H., Peng H., Goh W. W. B. Dealing with missing values in proteomics data // *Proteomics*. 2022. Vol. 22, N 23–24. e2200092. DOI: 10.1002/pmic.202200092.
9. Schulz M. A., Yeo B. T. T., Vogelstein J. T. et al. Different scaling of linear models and deep learning in UK Biobank brain images versus machine-learning datasets // *Nature Communications*. — 2020. — Vol. 11. — Article № 4238. — DOI: 10.1038/s41467-020-18037-z.
10. Информационный индекс (Information value) // Loginom Wiki [Элек-тронный ресурс]. — URL: <https://wiki.loginom.ru/articles/information-value.html> (дата обращения: 02.05.2025).
11. Площадь под ROC-кривой (Area Under The ROC Curve) // Loginom Wiki. — URL: <https://wiki.loginom.ru/articles/auc.html> (дата обращения: 02.05.2025).