Мидуков Данила Сергеевич
магистрант
Научный руководитель
Аркадьева Ольга Геннадьевна
канд. экон. наук, доцент
ФГБОУ ВО «Чувашский государственный университет им. И.Н. Ульянова»
г. Чебоксары, Чувашская Республика

ПРОГНОЗИРОВАНИЕ ОТТОКА КЛИЕНТОВ БАНКА С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ МАШИННОГО ОБУЧЕНИЯ

Аннотация: *в статье исследуется задача прогнозирования банковского клиентского оттока с привлечением методов машинного обучения. Эмпирическую основу составили несколько банковских датасетов, предоставленных автором; при этом в качестве центрального массива для построения модели выбран churn.csv, включающий 10 000 наблюдений и 14 признаков. В работе проведены структурный анализ данных, сопоставление вспомогательных массивов, разведочное изучение факторов, связанных с уходом клиентов, а также сравнительное тестирование моделей Logistic Regression, Random Forest и Gradient Boosting. Показано, что на вероятность оттока в наибольшей степени воздействуют возраст клиента, число используемых продуктов, уровень активности и географическая принадлежность. Наиболее высокие показатели качества продемонстрировал алгоритм Gradient Boosting. Результаты могут быть использованы при проектировании банковской CRM-системы, ориентированной на раннее выявление клиентов из риск-сегмента.*

Ключевые слова: *отток клиентов, банк, машинное обучение, клиентская аналитика, прогнозирование.*

При высокой плотности конкуренции на финансовом рынке удержание существующей клиентской базы приобретает для банков не просто операционный, а экономически чувствительный характер. Уход клиента означает сокращение

процентных и комиссионных поступлений, увеличение расходов на привлечение новых потребителей банковских услуг и одновременное ослабление возможностей кросс-продаж. По этой причине прогнозирование оттока в банковской аналитике трактуется не исключительно как статистическая процедура, а как управленческий инструмент, непосредственно связанный с обеспечением устойчивости бизнеса [4].

Инструментарий современного машинного обучения позволяет одновременно учитывать демографические, финансовые и поведенческие параметры клиента, обнаруживая между ними и вероятностью ухода такие зависимости, которые не сводятся к линейным соотношениям. Исследования, посвященные банковскому churn, показывают, что наиболее убедительные результаты нередко демонстрируют ансамблевые алгоритмы, способные работать со сложной конфигурацией признакового пространства и с умеренно несбалансированными классами [1; 2; 8; 9].

Цель исследования состоит в построении и сопоставлении моделей машинного обучения, предназначенных для прогнозирования оттока клиентов банка на основе анализа нескольких банковских датасетов, а также в выявлении факторов, представляющих наибольшую практическую значимость для системы клиентского удержания. Исследовательская гипотеза исходит из того, что поведенческие и продуктовые характеристики обладают более высокой прогностической силой по сравнению с отдельными формальными параметрами, тогда как ансамблевые методы оказываются результативнее линейного подхода при решении указанной задачи.

В рамках исследования были рассмотрены несколько банковских датасетов, загруженных автором и отражающих различные стороны взаимодействия клиента с кредитной организацией. Их сопоставительный анализ показал, что лишь датасет churn.csv непосредственно содержит целевую переменную оттока Exited, вследствие чего именно он может рассматриваться как базовая эмпирическая основа для задачи бинарной классификации. Прочие массивы не решают задачу прогнозирования оттока напрямую, однако представляют интерес как источники

контекстных признаков и позволяют оценить, какие дополнительные поведенческие характеристики могут быть впоследствии включены банком в расширенную модель удержания.

Основной массив churn.csv насчитывает 10 000 наблюдений и 14 признаков. В нем не обнаружено пропусков, полных дубликатов и повторяющихся значений по CustomerId. На этапе предобработки из набора были исключены технические признаки RowNumber, CustomerId и Surname; категориальные переменные Geography и Gender были переведены в числовое представление. После этого выборка была разделена на обучающую и тестовую части в соотношении 80:20 с сохранением структуры целевой переменной.

Доля наблюдений с Exited = 1 составила 20,37%, что свидетельствует о наличии умеренного дисбаланса классов. По этой причине качество моделей оценивалось не только при помощи accuracy, но и на основе ROC-AUC, PR-AUC, precision, recall и F1. Подобная схема согласуется с современными подходами к анализу бинарных классификаторов на несбалансированных выборках [5; 7]. В качестве интерпретируемой базовой модели использовалась Logistic Regression, а среди ансамблевых методов были выбраны Random Forest и Gradient Boosting [3; 6].

Таблица 1

Сопоставление проанализированных датасетов и их функций в исследовании

Датасет	*Размер*	*Содержание*	*Роль в исследовании*
churn.csv	10 000 × 14	Клиентские, финансовые и поведенческие признаки; целевая переменная Exited	Основной массив для построения churn-модели
sale_tasks_dataset.csv	18 691 × 28	Продажные коммуникации, длительность звонка, причины отказа, исход контакта	Показал ценность коммуникационных признаков
DANO.zvonki.robotov.csv	78 434 × 23	Результаты роботизированных звонков, доступность клиента, сценарии отклика	Показал значимость факта дозвона и контактируемости
application_dataset.csv	1 807 426 × 12	Операционная воронка по заявкам, встречам и успеш-	Показал ценность продуктовых и воронковых

		ности продукта	признаков

Изучение дополнительных массивов позволило выявить, что в продажных и коммуникационных датасетах на исход взаимодействия заметно влияют длительность разговора, сам факт успешного контакта и форма клиентского отклика. В sale_tasks_dataset.csv доля успешных исходов составила 12,93%, причем средняя продолжительность успешного контакта оказалась значительно выше по сравнению с неуспешным сценарием. В DANO_zvonki_robotov.csv доминирующим исходом оказался недозвон – 64,39%, что делает канал коммуникации и достижимость клиента потенциально важными переменными для последующего расширения churn-модели. В application_dataset.csv высокий уровень успешности внутри продуктовой воронки дополнительно указывает на практическую значимость продуктовых и воронковых индикаторов в рамках клиентской аналитики.

Разведочное изучение основного массива данных выявило выраженную неоднородность распределения оттока между различными клиентскими сегментами. Наиболее высокая доля клиентов, прекративших взаимодействие с банком, зафиксирована в Германии – 32,44%; для Франции и Испании аналогичные показатели заметно ниже и составляют 16,15% и 16,67% соответственно. Сегментация по признаку активности обнаруживает сопоставимо отчетливый разрыв: в группе неактивных клиентов уровень оттока достигает 26,85%, тогда как среди активных он снижается до 14,27%. Не менее показателен возрастной срез. Если в категории 18–30 лет доля ушедших клиентов равна 7,52%, а в интервале 31–40 лет – 12,09%, то в группе 41–50 лет она возрастает до 33,97%, а среди клиентов 51–60 лет достигает 56,21%.

Особую аналитическую значимость демонстрирует показатель, отражающий количество используемых банковских продуктов. Для клиентов, располагающих одним продуктом, уровень оттока составил 27,71%; при наличии двух продуктов он сокращается до 7,58%; в случае трех продуктов, напротив, возрастает до 82,71%. Здесь просматривается обстоятельство, имеющее прямое прикладное

значение: устойчивость клиентских отношений определяется не только фактом присутствия клиента в банковской базе как таковым, но и глубиной его включенности в продуктовую конфигурацию банка. Подобная трактовка согласуется с исследованиями, в которых именно продуктовые и поведенческие характеристики рассматриваются как одни из наиболее содержательных факторов в задачах churn prediction [2; 8; 9].

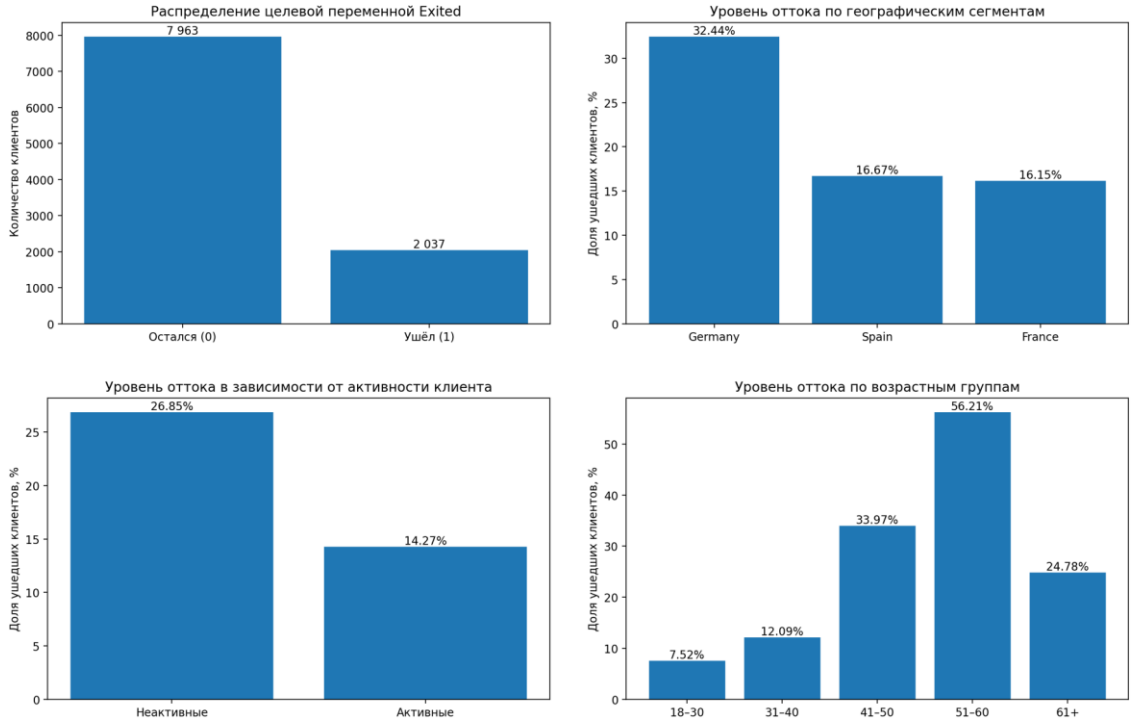

Рис. 1. Распределение целевой переменной и зависимость уровня оттока от географии, активности и возраста клиента

С точки зрения прикладного использования выявленные соотношения означают, что банк получает возможность строить дифференцированную retention-политику. Возраст, клиентская активность и глубина использования продуктовой линейки должны трактоваться как ключевые маркеры риска. При такой постановке прогнозная модель выступает средством раннего обнаружения клиентов, в отношении которых еще до фактического ухода можно запускать индивидуализированные сценарии удержания [4].

Таблица 2

Сравнение качества моделей машинного обучения

Модель	*ROC-AUC*	*PR-AUC*	*Precision*	*Recall*	*F1*
Logistic Regression	0,775	0,479	0,589	0,187	0,284
Random Forest	0,853	0,678	0,756	0,450	0,564
Gradient Boosting	0,870	0,719	0,782	0,494	0,605

Сравнительная оценка моделей выявила преимущество ансамблевых алгоритмов по отношению к линейному подходу. Logistic Regression обеспечила ROC-AUC = 0,775, Random Forest – 0,853, тогда как Gradient Boosting достиг значения 0,870. По метрике PR-AUC, наиболее чувствительной к качеству работы на положительном классе в условиях дисбаланса, наилучший результат также показал Gradient Boosting – 0,719. В прикладном измерении это означает, что ансамблевая модель точнее выделяет клиентов, находящихся в риск-сегменте, и по этой причине лучше адаптирована для использования в банковском CRM-контуре [2; 8; 9].

В то же время Random Forest продемонстрировал более уравновешенное сочетание precision, recall и F1, поэтому его можно рассматривать как устойчивую альтернативу в тех ситуациях, когда для банка особенно важен баланс между полнотой выявления и числом ложных срабатываний. Полученные результаты поддерживают исходную гипотезу: решающий вклад в прогноз вносят возраст, активность, географический сегмент и число банковских продуктов, а ансамблевые методы адекватнее воспроизводят сложную конфигурацию связей между признаками и фактом оттока.

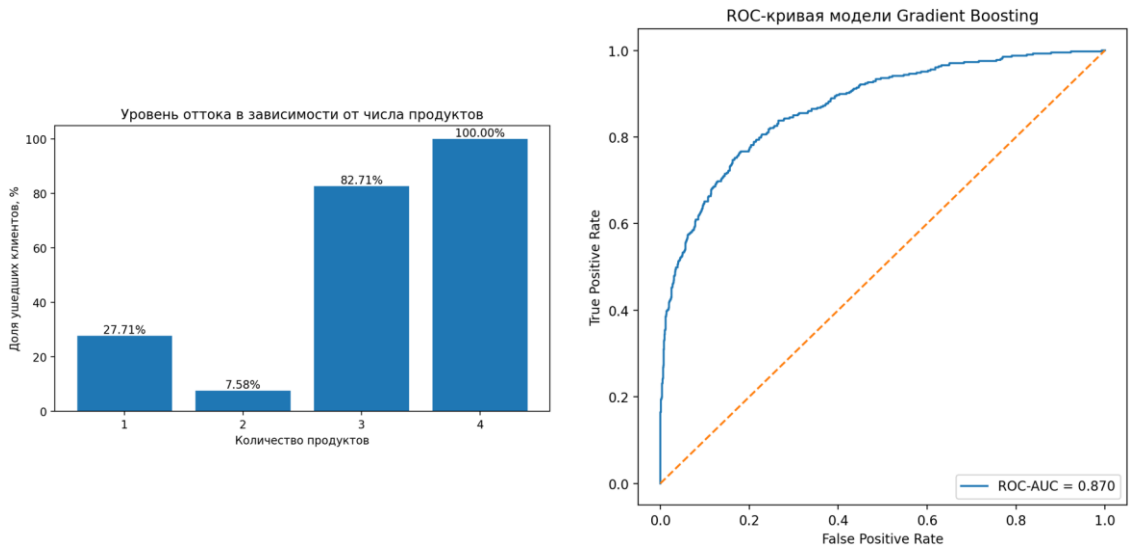

Рис. 2. Уровень оттока по числу продуктов
и ROC-кривая модели Gradient Boosting

С практической стороны на основании полученных результатов может быть предложен следующий сценарий применения модели: банк на регулярной основе рассчитывает вероятность ухода для каждого клиента, выделяет верхний риск-сегмент и передает его retention-команде для запуска адресных мероприятий. При precision, близкой к 0,78, модель Gradient Boosting способна выполнять функцию инструмента приоритизации клиентской базы, повышая результативность удерживающих кампаний и переводя анализ оттока в плоскость конкретных управленческих действий.

Проведенное исследование показало, что прогнозирование банковского клиентского оттока с применением методов машинного обучения обладает одновременно аналитической и прикладной значимостью. Сопоставление нескольких загруженных датасетов позволило, с одной стороны, выделить churn.csv как основной массив для моделирования, с другой – обозначить перспективные группы дополнительных признаков, связанных с коммуникациями, продуктовой воронкой и доступностью клиента. В качестве ключевых факторов риска были выявлены возраст, географический сегмент, уровень активности клиента и число используемых банковских продуктов. Наилучшие показатели качества прогноза продемонстрировал алгоритм Gradient Boosting. Применение ансамблевых моде-

лей в банковской клиентской аналитике может рассматриваться как действенный механизм раннего выявления клиентов из риск-группы и повышения результативности CRM-политики банка.

References

1. Al-Najjar D., Al-Rousan N., Al-Najjar H. Machine Learning to Develop Credit Card Customer Churn Prediction // Journal of Theoretical and Applied Electronic Commerce Research. 2022. Vol. 17. No. 4. Pp. 1529–1542. DOI: 10.3390/jtaer17040077. EDN: MIBVVI

2. Bharathi S.V., Pramod D., Raman R. An Ensemble Model for Predicting Retail Banking Churn in the Youth Segment of Customers // Data. 2022. Vol. 7. No. 5. Art. 61. DOI: 10.3390/data7050061. EDN: RIDLBA

3. Breiman L. Random Forests // Machine Learning. 2001. Vol. 45. No. 1. Pp. 5–32. DOI: 10.1023/A:1010933404324. EDN: ARROTH

4. Farquhar J.D., Panther T. Acquiring and retaining customers in UK banks: an exploratory study // Journal of Retailing and Consumer Services. 2008. Vol. 15. No. 1. Pp. 9–21. DOI: 10.1016/j.jretconser.2007.02.001.

5. Fawcett T. An introduction to ROC analysis // Pattern Recognition Letters. 2006. Vol. 27. No. 8. Pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.

6. Friedman J.H. Greedy Function Approximation: A Gradient Boosting Machine // The Annals of Statistics. 2001. Vol. 29. No. 5. Pp. 1189–1232. DOI: 10.1214/aos/1013203451.

7. Saito T., Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets // PLOS ONE. 2015. Vol. 10. No. 3. Art. e0118432. DOI: 10.1371/journal.pone.0118432. EDN: YBEDRE

8. Tékouabou S.C.K., Gherghina Ș.C., Toulni H., Mata P.N., Martins J.M. Towards Explainable Machine Learning for Bank Churn Prediction Using Data Balancing and Ensemble-Based Methods // Mathematics. 2022. Vol. 10. No. 14. Art. 2379. DOI: 10.3390/math10142379. EDN: HVEEDE

9. Tran H.D., Le N., Nguyen V.-H. Customer Churn Prediction in the Banking Sector Using Machine Learning-Based Classification Models // Interdisciplinary Journal of Information, Knowledge, and Management. 2024. Vol. 19. DOI: 10.28945/ijikm.v19i1.51.