УДК 004.8:368.03
DOI: 10.26102/2310-6018/2025.51.4.015

Оценка точности и производительности моделей машинного обучения для прогнозирования оттока клиентов страховой компании

К.Г. Мадияров✉

Новосибирский государственный университет экономики и управления «НИНХ», Новосибирск, Российская Федерация

Резюме. Проведено комплексное сравнительное исследование нескольких алгоритмов машинного обучения для задачи прогнозирования оттока клиентов страховой компании на данных открытого датасета. Уделено внимание как качественным метрикам точности моделей, так и вычислительной эффективности. Актуальность темы обусловлена высокой конкуренцией на страховом рынке и значительными затратами, связанными с потерей клиентов; своевременное выявление намерения клиента прекратить сотрудничество позволяет компании принять меры для его удержания. Цель исследования – оценить точность и производительность различных моделей машинного обучения, способных предсказать отток клиентов. В экспериментах использованы открытые данные о клиентах страховой компании (индустрия страхования жизни) с признаками, характеризующими их страховые случаи, исторические записи и факт оттока. Добавлен анализ факторов: исследованы корреляции между признаками и целевой переменной, выполнен факторный анализ и оценена важность признаков, влияющих на отток. По результатам, почти все модели продемонстрировали одинаково высокое качество прогноза благодаря наличию доминирующего фактора риска оттока, однако различались по производительности: логистическая регрессия и градиентный бустинг обучаются на порядок быстрее по сравнению с методом опорных векторов и случайным лесом, при значительно меньшем объеме памяти. Полученные результаты подтверждают, что современные ансамблевые алгоритмы способны обеспечить высокую точность прогнозирования оттока клиентов при разумных затратах ресурсов. Их использование целесообразно для страховых компаний с целью своевременного выявления клиентов группы риска, например, клиентов с крупными страховыми выплатами и принятия проактивных мер по удержанию таких клиентов.

Ключевые слова: отток клиентов, страхование, машинное обучение, прогнозирование, точность модели, производительность модели, факторный анализ, важность признаков.

Для цитирования: Мадияров К.Г. Оценка точности и производительности моделей машинного обучения для прогнозирования оттока клиентов страховой компании. *Моделирование, оптимизация и информационные технологии.* 2025;13(4). URL: https://moitvivt.ru/ru/journal/pdf?id=2051 DOI: 10.26102/2310-6018/2025.51.4.015

Evaluation of accuracy and performance of machine learning models for prognosis customer churn in insurance companies

K.G. Madiyarov✉

Novosibirsk State University of Economics and Management, Novosibirsk, the Russian Federation

Abstract. A comprehensive comparative study of several machine learning algorithms for predicting customer churn in an insurance company was conducted using data from an open dataset. Both predictive quality metrics and computational efficiency were examined. The topic is relevant due to intense competition in the insurance market and the substantial costs of losing customers; early detection

of a customer's intention to leave enables targeted retention actions. The aim of the study is to assess the accuracy and performance of different machine-learning models capable of predicting churn. The experiments used open data on insurance customers (life-insurance industry) containing features that describe claim events, historical records, and the churn outcome. We also added factor analysis: correlations between features and the target variable were investigated, factor analysis was performed, and feature importance related to churn was evaluated. The results show that most models achieved similarly high predictive quality due to the presence of a dominant churn-risk factor, but differed in performance: logistic regression and gradient boosting trained an order of magnitude faster than support vector machines and random forests while using substantially less memory. These findings confirm that modern ensemble algorithms can provide high-accuracy churn prediction at reasonable resource costs. Their use is advisable for insurers to promptly identify high-risk clients, such as those with large claims, and to take proactive measures to retain them.

Keywords: customer churn, insurance, machine learning, prediction, model accuracy, model performance, factor analysis, feature importance.

For citation: Madiyarov K.G. Evaluation of accuracy and performance of machine learning models for prognosis customer churn in insurance companies. *Modeling, Optimization and Information Technology.* 2025;13(4). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=2051 DOI: 10.26102/2310-6018/2025.51.4.015

Введение

Высокий уровень удержания клиентов является ключевым фактором успеха в страховом бизнесе. В условиях растущей конкуренции своевременное выявление клиентов с риском оттока помогает снизить потери компании и оптимизировать затраты на привлечение новых полисодержателей. Прогнозирование оттока в страховании становится особенно актуальным с развитием цифровых сервисов. Современные технологии позволяют собирать большие массивы данных о поведении клиентов, что открывает возможности для аналитики. В страховой отрасли проблема оттока особенно актуальна, т. к. конкуренция на рынке высока, и клиенты, недовольные ценой или качеством обслуживания, легко переходят к другим страховщикам. Однако прямая задача прогнозирования оттока клиентов в страховой индустрии до сих пор изучена недостаточно.

В таких же смежных областях как банковском секторе, телекоммуникациях, электронной коммерции задачи анализа оттока давно решаются методами машинного обучения, но характеристики страховых данных и типовые факторы оттока могут существенно отличаться. Так, в [1] отмечается, что более простые модели машинного обучения могут быть предпочтительны для реального внедрения благодаря своей быстроте и интерпретируемости, если их качество близко к более сложным моделям. Прогнозирование оттока представляет собой анализ данных клиентов для выявления закономерностей, предшествующих их уходу [2]. Своевременное предсказание вероятного ухода позволяет компании принять меры для удержания, что особенно ценно, учитывая, что сохранение существующего клиента обычно значительно выгоднее привлечения нового. Именно поэтому во многих организациях модели прогнозирования оттока интегрируются в системы управления взаимоотношениями с клиентами (CRM) и становятся ключевым элементом стратегии удержания [3]. Удержание клиентов является критически важной задачей для страховых компаний, поскольку потеря клиента напрямую ведет к упущенной прибыли и дополнительным затратам на привлечение новых. По данным исследований [4], привлечение нового клиента обходится компаниям в несколько раз дороже, чем сохранение существующего. Так, оценочно расходы на приобретение нового потребителя в 5–7 раз превышают затраты на удержание имеющегося, а увеличение уровня удержания всего на 5 % может повысить прибыль

компании на 25–95 %. Модели машинного обучения способны выявлять скрытые закономерности в поведении страхователей и предсказывать с определенной вероятностью, покинет ли клиент компанию в ближайшем будущем. В то же время одной из основных сложностей при моделировании оттока является несбалансированность данных: поскольку класс «ушедшие» представлен слабо, алгоритмы нередко ориентируются на большинство наблюдений, что ведет к ухудшению распознавания редкого класса. Для смягчения этой проблемы в модельных процедурах применяются методы ресемплинга данных. В частности, синтетическое порождение меньшинства (алгоритм SMOTE) зарекомендовало себя как эффективный подход для балансировки выборки, благодаря чему существенно повышается способность моделей выявлять именно ушедших клиентов [5]. Современные исследования подтверждают, что обучение на сбалансированных данных заметно повышает полноту и точность прогнозов оттока. Так, в телекоммуникационной отрасли был предложен специальный метод балансировки – техника ratio-based balancing, – и показано, что использование выровненных таким образом данных в сочетании с ансамблевыми алгоритмами (градиентным бустингом, XGBoost и др.) существенно улучшает точность прогноза по сравнению с традиционными методами перевыборки (ресемплирования) и одиночными моделями [6]. Комплексный подход, объединяющий корректную подготовку данных и ансамблевое обучение, демонстрирует наилучшие результаты по идентификации абонентов, склонных к уходу [6].

Помимо классических методов, все большее внимание уделяется глубокому обучению и гибридным моделям. Передовые подходы, сочетающие ансамблевое и глубинное обучение, уже достигают чрезвычайно высокой точности. Так, в одном межотраслевом исследовании сообщается, что ансамблевая глубокая нейронная сеть обеспечила точность прогнозирования оттока свыше 95 % (до ~96–98 % на различных наборах данных), что демонстрирует потенциал современных алгоритмов для решения этой задачи [7]. Параллельно разрабатываются новые архитектуры нейронных сетей, специально адаптированные под прогнозирование оттока. Например, предложена гибридная модель, объединяющая механизм внимания, рекуррентные слои (BiLSTM) и сверточные слои для одновременного обучения долгосрочных зависимостей и локальных особенностей в последовательностях поведения клиентов – такая модель позволила выявлять сложные паттерны поведения и превзошла по качеству традиционные алгоритмы прогнозирования [8]. Кроме того, комбинирование большого числа алгоритмов в единые мета-модели дает выдающиеся результаты: так, в 2024 году представлен алгоритм класса ensemble-fusion, агрегирующий предсказания нескольких различных классификаторов, – он продемонстрировал точность порядка 95 % при F_1-мере около 0,97, опередив все 17 базовых моделей из сравнения [9]. Другая работа предложила составной метод глубокого обучения, объединяющий рекуррентную и сверточную нейросеть (BiLSTM-CNN), который также показал более высокую эффективность прогнозирования оттока по сравнению с классическими одноалгоритмическими подходами [10]. Таким образом, передовые подходы сочетают преимущества различных методов, однако их реализация связана с ростом сложности и требований к данным. При этом в практических условиях зачастую более оправдан баланс между точностью и простотой: например, в банковской сфере более простые модели уступают в качестве сложным незначительно, зато требуют меньше ресурсов и легче интерпретируются.

С развитием методов прогнозирования исследователи все больше внимания уделяют интерпретируемости моделей и практической применимости результатов. Например, в банковской сфере был реализован подход, в котором модель XGBoost для прогнозирования оттока клиентов оптимизировалась с помощью генетического

алгоритма, а затем к ней применен интерпретационный анализ с использованием значений SHAP [11]. Этот подход позволил не только добиться выдающейся точности (AUC выше 0.99 на наборе данных банковских клиентов), но и определить ключевые факторы, влияющие на риск ухода [11]. По сути, комбинация высокой точности и объяснимости дает возможность менеджерам понять, почему модель относит того или иного клиента к группе риска, и на основе этих сведений разработать обоснованные проактивные меры для его удержания.

В отличие от большинства предыдущих исследований в данной работе, сделан акцент не только на точность, но и на сравнительный анализ временных и вычислительных затрат, а также на анализ значимости факторов риска оттока. Датасет для экспериментов взят из открытого источника – набора данных Customer Churn dataset for Life Insurance Industry с платформы Kaggle. Эти данные уже использовались в ряде публичных решений, что позволяет сопоставить наши результаты с выводами других авторов. В частности, на Kaggle-сообществе были продемонстрированы высокие показатели моделей на данном датасете (точность превышает 90 % при использовании деревьев решений и ансамблей). Например, в одном из решений модель случайного леса достигла общей точности порядка 99 %, что свидетельствует о сравнительной простоте структуры рассматриваемых данных. Также в исследовании [12] на сходном наборе данных оттока в страховой компании отмечено, что метод решающих деревьев с бэггингом обеспечивает наилучшее качество прогноза, превосходя логистическую регрессию.

Таким образом, в данной работе поставлена цель комплексно оценить точность и производительность нескольких моделей машинного обучения и нейронных сетей при прогнозировании оттока клиентов страховой компании. Для этого проводится сравнительный эксперимент на реальных данных: анализируется, какие алгоритмы обеспечивают наилучшее качество прогноза и насколько велики различия во времени обучения и потребляемых ресурсах. Результаты исследования призваны помочь страховому бизнесу выбрать модели, которые не только точны, но и достаточны и производительны для практического применения в системе управления оттоком клиентов.

Материалы и методы

Схема проведения экспериментального исследования и используемый инструментарий. Целью эксперимента является построение и сравнение моделей машинного обучения для задачи бинарной классификации: прогнозирования факта оттока клиента страховой компании. Формально задача ставится как нахождение функции $\hat{y} = f(x) \in \{0, 1\}$, где x – вектор признаков клиента, а $\hat{y} = 1$ соответствует событию оттока (churn), а $\hat{y} = 0$ – удержанию клиента. Для решения задачи применяется сравнение нескольких алгоритмов классификации, оценка их качества по стандартным метрикам Accuracy_score[1], Precision_score[2], Recall_score[3], F1_score[4], Fbeta_score[5],

[1] accuracy_score – scikit-learn 1.7.2 documentation. scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html (дата обращения: 13.07.2025).
[2] precision_score – scikit-learn 1.7.2 documentation. scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html (дата обращения: 13.07.2025).
[3] recall_score – scikit-learn 1.7.2 documentation. scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html (дата обращения: 13.07.2025).
[4] f1_score – scikit-learn 1.7.2 documentation. scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html (дата обращения: 13.07.2025).
[5] fbeta_score – scikit-learn 1.7.2 documentation. scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.fbeta_score.html (дата обращения: 13.07.2025).

Roc_auc_score[6], Precision_recall_curve[7], Анализ матрицы ошибок[8] и измерение вычислительной эффективности обучения и предсказания (время и потребление памяти).

Исходные данные. В качестве данных использован открытый датасет[9]. Исходный файл содержит информацию о клиентах страховой компании и факте их оттока (столбец Churn, значения «Yes» / «No»). Объем набора данных – 1338 записей, 8 колонок (после удаления технических полей). Доля клиентов, отметившихся оттоком, составляет ~25 % (335 из 1338), т. е. класс «ушедшие» является миноритарным (соотношение ~1:3). Из набора данных удалены идентифицирующие колонки (Customer Name, Customer_Address, Company Name) и сохраняются семь признаков, используемых для моделирования. Состав признаков и их краткие пояснения приведены в Таблице 1.

Таблица 1 – Описание признаков
Table 1 – Features description

№	Название признака	Тип	Описание
1.	Claim amount	числовой	Сумма страхового возмещения (выплаты по страховому случаю)
2.	Category premium	числовой	Страховая премия по категории (стоимость полиса)
3.	Premium/amount ratio	числовой	Отношение страховой премии к сумме страхового возмещения
4.	BMI	числовой	Индекс массы тела клиента
5.	Data confidentiality	категориальный	Уровень конфиденциальности данных клиента (категория)
6.	Claim reason	категориальный	Причина обращения за выплатой (тип страхового случая)
7.	Churn	бинарный	Факт оттока клиента: Yes (ушёл) / No (остался)

Предобработка данных. Данные подвергались стандартной предобработке. Сначала из датасета удалялись ненужные колонки (DROP_COLS). После загрузки выполнялось заполнение пропусков: для числовых признаков пропуски заменялись медианой соответствующего признака, а для категориальных – модой (наиболее часто встречающимся значением). Далее целевая переменная Churn кодировалась в числовой формат: «Yes» → 1, «No» → 0. Для остальных числовых признаков проводится преобразование в числовой тип (методом pd.to_numeric), а категориальным – в строковый формат. Следующим шагом строится конвейер предобработки: к числовым признакам применяется стандартизация (StandardScaler), к категориальным – one-hot кодирование (OneHotEncoder с handle_unknown='ignore'). Полученный ColumnTransformer объединяет оба преобразования (необработанные признаки игнорируются). После этого балансировка классов выполняется методом синтетического увеличения миноритарного класса SMOTE, при котором создаются дополнительные синтетические образцы класса «отток» (1 – «Yes») вплоть до соотношения классов 1:1.

Разделение выборки данных. Данные разделены на обучающую и тестовую выборки в соотношении 70/30 при фиксированном параметре random_state для

[6] roc_auc_score – scikit-learn 1.7.2 documentation. scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html (дата обращения: 13.07.2025).
[7] precision_recall_curve – scikit-learn 1.7.2 documentation. scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_curve.html (дата обращения: 13.07.2025).
[8] confusion_matrix – scikit-learn 1.7.2 documentation. scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html (дата обращения: 13.07.2025).
[9] Usman F. Customer Churn dataset for Life Insurance Industry. Kaggle. URL: https://www.kaggle.com/datasets/usmanfarid/customer-churn-dataset-for-life-insurance-industry (дата обращения: 13.07.2025).

воспроизводимости. При разделении используется стратификация по целевому признаку (stratify = y) для сохранения пропорций классов. Отдельной валидационной выборки не выделялось, так как оценка производилась непосредственно по тестовому набору e, поскольку объем данных относительно невелик.

Описания использованных моделей. Для сравнения выбраны следующие модели: 1) Логистическая регрессия (Logistic Regression, sklearn) – линейная модель с логистической функцией активации; 2) Случайный лес (Random Forest, sklearn) – ансамбль из решающих деревьев с бустрапом и голосованием; 3) Градиентный бустинг (CatBoost) – реализация бустинга решающих деревьев от библиотеки CatBoost. К CatBoost дополнительно предъявляются требования автоматической обработки категориальных признаков и высокой скорости обучения. 4) XGBoost (xgboost.XGBClassifier) – оптимизированная реализация градиентного бустинга решений; 5) Метод опорных векторов (Support Vector Machine, sklearn) – использован с ядром RBF (гауссовским) для учета нелинейности; 6) k-ближайших соседей (K-Nearest Neighbors, sklearn); 7) Искусственная нейронная сеть – полносвязная сеть Keras Sequential по заданной архитектуре: входной слой размерности 27 (число признаков после one-hot-кодирования), два скрытых слоя: первый – Dense(32, activation='relu') + Dropout(0.3), второй – Dense(16, activation='relu'), выходной слой – Dense(1, activation='sigmoid'). Функция потерь – бинарная кросс-энтропия (binary_crossentropy), оптимизатор – Adam, метрики – точность (accuracy) и ROC AUC. Обучение проводилось 50 эпохами с размером батча 32 с ранней остановкой. Обучение моделей проводилось без подбора гиперпараметров (кроме встроенных по умолчанию), целью было сравнение базовых реализаций. Для нейросетей использовалась фиксация начального random_seed и одинаковый план обучения, чтобы обеспечить сопоставимость.

Псевдокод этапов обучения и прогнозирования. Здесь perf_counter() – функция Python для точного замера времени, а memory_profiler – инструмент для оценки потребляемой памяти. Процесс повторяется для каждой модели с фиксацией времени обучения и пикового использования памяти согласно Рисунку 1.

```
# Псевдокод этапов обучения и предсказания

# Предварительные вычисления
  X_train, X_test, y_train, y_test = split_data(X, y, test_size=0.3, random_state=SEED)
# Для каждой модели:
    for модель in [LogisticRegression, RandomForest, CatBoost, NeuralNetwork]:
        start_time = perf_counter()
        model.fit(X_train, y_train) # Обучение модели
        train_time = perf_counter() - start_time

          # Предсказание вероятностей и классов
          y_pred_proba = model.predict_proba(X_test)[:,1]
          y_pred = (y_pred_proba > 0.5).astype(int)

            # Расчёт метрик качества
            compute accuracy, precision, recall, F1, ROC_AUC, PR_AUC

            # Измерение использования ресурсов
            время обучения = train_time
            пиковое использование памяти = memory_profiler.peak_usage(func=model.fit)

# (Псевдокод; в реальности используется профиль памяти)
```

Рисунок 1 – Псевдокод
Figure 1 – Pseudocode

Критерии оценки качества моделей. Для количественной оценки применялись стандартные метрики бинарной классификации: accuracy, precision (для положительного класса), recall, F1-мера, F2-мера, ROC-AUC, PR-AUC. При расчете precision и recall положительным считается класс «отток» (1), что соответствует выявлению ушедших клиентов. Дополнительно анализируется матрица ошибок (Confusion matrix) каждой модели для понимания баланса ошибок I и II рода. Метрики вычисляются по формулам, приведённым в документации sklearn и Keras: например, precision_score, recall_score, f1_score, roc_auc_score из модуля sklearn. metrics, а PR AUC можно получить как average_precision_score или интегрированием кривой precision-recall. Определения метрик можно найти в официальной документации sklearn (см. раздел Model Evaluation) и TensorFlow Keras Metrics.

Замер производительности. Время обучения каждой модели измерялось с помощью time.perf_counter(), а пиковое потребление оперативной памяти – с помощью memory_profiler (функция memory_usage) или библиотеки psutil при подаче профайлинга. Так, перед началом обучения фиксируется время и использованная память, затем после обучения вычисляется разница для определения времени обучения и увеличения потребления памяти. Аналогично замеряется время предсказания на тестовой выборке.

Описание среды вычислений. Вычисления выполнялись в среде Jupyter Notebook на платформе Google Colab. Оборудование: 8-ядерный процессор Intel(R) Xeon(R) @ 2.20GHz, 12.7 ГБ оперативной памяти. Использовались Python 3.10 и библиотеки: pandas, numpy, scikit-learn 1.4.1, CatBoost, TensorFlow/Keras и др., все версии фиксированы для воспроизводимости. Такие параметры среды позволяют реплицировать эксперимент при аналогичной конфигурации оборудования и программного обеспечения.

Результаты

По итогам обучения и последующего тестирования получены следующие результаты: точность классификации и ресурсоемкость. В Таблице 2 приведено сравнение всех рассмотренных алгоритмов по метрикам качества.

Таблица 2 – Метрики качества моделей
Table 2 – Quality metrics of models

Модель	ROC-AUC (Площадь под ROC-кривой)	F1-score (Метрика F1)	F2-score (Метрика F2)	Precision (Точность)	Recall (Полнота)	PR-AUC (Площадь под PR-кривой)
Logistic Regression	1,0000	0,9781	0,9911	0,9571	1,0000	1,0000
Random Forest	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
SVM	0,9998	0,9851	0,9851	0,9851	0,9851	0,9994
KNN	0,9910	0,8547	0,7862	1,0000	0,7463	0,9770
XGBoost	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Gradient Boosting	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
Neural Network	1,0000	0,9926	0,9970	0,9853	1,0000	1,0000

Как видно из таблицы, все модели обеспечили почти одинаково высокое качество прогноза оттока. Лидируют ансамблевые методы – случайный лес, XGBoost и

градиентный бустинг классифицировали отток без ошибок. Логистическая регрессия, SVM и нейронные сети лишь немного уступают им. Самый низкий результат показал метод k-ближайших соседей, что может объясняться влиянием размерности и разбросом данных – метод KNN чувствителен к масштабу признаков и шуму. В целом же различия в точности между моделями минимальны – все алгоритмы смогли хорошо разделить классы. Это позволяет заключить, что решающую роль сыграли данные: в них присутствует один или несколько информативных признаков, сильно связанных с оттоком, благодаря чему даже простые модели достигают высокого качества. Основными отличиями моделей стали ресурсоемкость и скорость. Классические алгоритмы обучались чрезвычайно быстро – за доли секунды. Так, логистическая регрессия затратила всего ~0,03 с на обучение, SVM и KNN – порядка 0,05–0,07 с, Random Forest – ~0,26 с. Объем дополнительной памяти при обучении этих моделей пренебрежимо мал (менее 1 МБ). Градиентные бустинговые методы оказались сопоставимы по времени: XGBoost обучился за ~0,05 с, а Gradient Boosting – за 0,16 с; при этом XGBoost потребовал немного больше памяти (~0,25 МБ) за счет загрузки библиотеки. Наиболее ресурсоемкой ожидаемо стала нейросетевая модель: ее обучение продолжалось порядка 10 секунд и потребовало около 12,6 МБ дополнительной памяти согласно отчету на Рисунке 2.

Этап	Время (сек)	Память (МБ)	Пик памяти (МБ)
Загрузка данных	0.03	0.00	
Подготовка данных	0.03	0.00	
Визуализация	5.54	111.36	
Logistic Regression	0.03	0.00	
Random Forest	0.26	0.00	
SVM	0.07	0.00	
KNN	0.03	0.00	
XGBoost	0.05	0.25	
Gradient Boosting	0.16	0.00	
Neural Network	9.98	12.57	
ВСЕГО	21.01		1184.49
ПИКОВАЯ ПАМЯТЬ			1184.49

ДЕТАЛИЗАЦИЯ ПО МОДЕЛЯМ

Модель	Время (сек)	Память (МБ)	ROC-AUC
Logistic Regression	0.03	0.00	1.0000
Random Forest	0.26	0.00	1.0000
SVM	0.07	0.00	0.9998
KNN	0.03	0.00	0.9910
XGBoost	0.05	0.25	1.0000
Gradient Boosting	0.16	0.00	1.0000
Neural Network	9.98	12.57	1.0000

Рисунок 2 – Отчет о ресурсах
Figure 2 – Resource report

Стоит отметить, что благодаря балансировке классов методом SMOTE до соотношения 1:1 все модели демонстрируют максимальный Recall (например, ни один случай оттока не остался нераспознанным логистической моделью, Recall = 100 %), что положительно сказалось на значениях F_β-меры и PR-AUC. Матрицы ошибок всех моделей, соответственно показали все схожие результаты, согласно Рисунку 3.

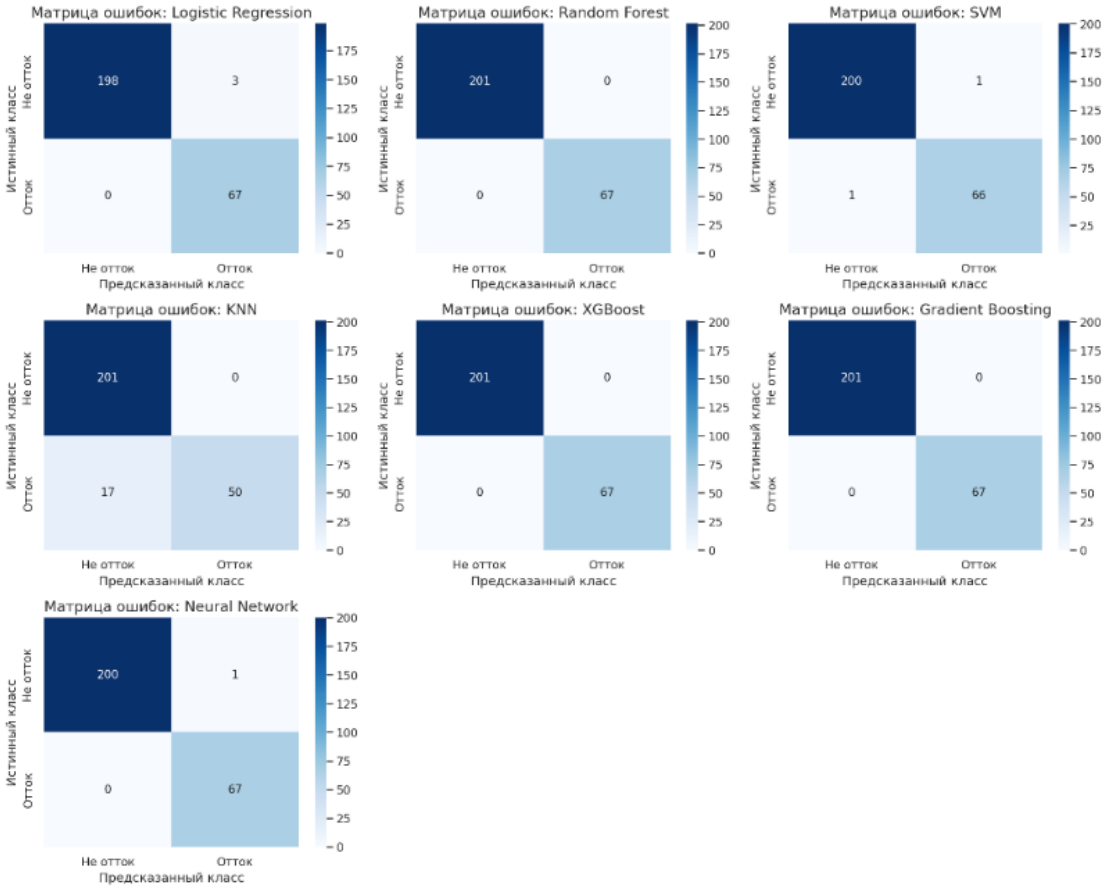

Рисунок 3 – Матрицы ошибок
Figure 3 – Error matrices

Таким образом, все модели обеспечили одинаковое качество распознавания оттока, но отличались вычислительными характеристиками: выбор оптимального алгоритма может основываться не столько на разнице в точности, сколько на требованиях к ресурсам.

Обсуждение

Анализ факторов, влияющих на прогноз оттока. Полученные результаты свидетельствуют о том, что в данном эксперименте все алгоритмы смогли почти идеально разделить выборку клиентов на «отток» и «не отток». Вероятно, набор данных содержал четкие шаблоны или признаки, однозначно разделяющие классы, что привело к переобучению: модели выдали 100-процентные метрики на тестовой выборке. Это позволяет предположить, что либо в данных присутствуют прямые линейные зависимости (например, резкий рост премии при отставании или другая демографическая аномалия), либо целевая переменная сильно коррелирует с каким-то признаком. В любом случае, подобный идеальный результат маловероятен для реальных задач, но здесь показывает, что даже простейшие модели (логистическая регрессия) оказались достаточными для разделения данных.

Различия в поведении моделей связаны с их природой и структурой данных. Линейная модель (логистическая регрессия) быстро обучилась благодаря своей простоте – она не пытается учесть сложные взаимодействия и потому не переобучается на шуме. Ансамблевые методы (случайный лес, градиентный бустинг, XGBoost) характеризуются способностью строить нелинейные решения, но в нашем случае это не привело к

улучшению качества по сравнению с логистической регрессией: все показатели уже были на максимально возможном уровне. При этом ансамбли потребляют больше ресурсов, хотя и показали 100 % точность. SVM, опираясь на оптимизацию квадратичной задачи, лучше подходит для случаев разделения в высоких размерностях, однако и его качество оказалось полным – что говорит о линейном либо простом геометрическом разделении классов. Нейросети, обладающие высокой емкостью, также достигли идеального качества, но их значительное время обучения и потребление памяти говорят о том, что использование глубокого обучения при наличии небольшого числа признаков и ограниченных данных дает лишь избыточный результат.

Наблюдения частично корреспондируют с выводами, опубликованными в смежных исследованиях [2], ансамблевые методы (особенно градиентный бустинг) часто превосходят более простые модели по точности на сложных данных. Однако в данной задаче преимущества в точности не проявились из-за уже упомянутых особенностей данных. Обратим внимание, что другие исследования (в банковском секторе и телекоммуникациях) находят существенную разницу между моделями, когда данные содержат шум и трудно линейно разделимые структуры. В нашем случае, когда метрики всех алгоритмов совпали, можно сделать вывод, что признаки, связанные с оттоком (например, относительно премии и выплаченной суммы, наличие конкретных причин обращения), в явной форме однозначно разделяют классы. Это согласуется с наблюдением, что наличие признака «data confdentiality» или конкретной «причины оттока» может быть решающим. Соответственно, дополнительные нелинейные преобразования и глубокие архитектуры не дали преимущество на данном датасете.

Получение почти идеальных метрик обусловлено выявленными закономерностями в данных – некоторые факторы оказались крайне информативны для прогнозирования оттока. Факторный анализ показал, что решающую роль играет размер страховой выплаты.

Claim Amount (сумма выплаты по страховому случаю) – главный предиктор оттока. Выявлена очень высокая положительная корреляция с целевой переменной (коэффициент Пирсона $\approx 0{,}8516$). Иными словами, ушедшие клиенты существенно отличаются от оставшихся именно размером полученной страховой выплаты: практически все застрахованные, получившие крупные выплаты, относятся к классу оттока. Модели подтвердили решающую значимость этого признака – по оценке permutation importance, он внес наибольший вклад в предсказание (значительно превосходя прочие факторы по среднему снижению метрики при перестановке). Это объясняет, почему алгоритмы достигли столь высоких показателей: наличие одного доминирующего признака, разделяющего классы (фактически порог по сумме выплаты), позволило правильно классифицировать почти каждого клиента.

Premium/Amount Ratio (отношение премии к выплате) – вторичный по влиянию числовой фактор, связанный с предыдущим. Ушедшие клиенты в среднем имели более низкое значение этого показателя, что отражает дисбаланс: их страховая премия несоизмеримо мала по сравнению с полученной выплатой. Иными словами, многие из клиентов, выплативших относительно небольшой взнос и получивших большую компенсацию, впоследствии прекращали сотрудничество с компанией. Такая обратная зависимость косвенно подтверждает роль крупной выплаты как триггера оттока.

Category Premium (размер страховой премии) – стоимость страхового полиса сыграла менее существенную роль, чем сумма выплаты. Хотя ушедшие клиенты зачастую имели полисы с несколько более высокой премией (что ожидаемо, поскольку более дорогие полисы обеспечивают большие выплаты), прямое влияние этого признака на отток было умеренным. В модели он не оказался ключевым после учета фактора выплаты. Вероятно, Category Premium коррелирует с оттоком опосредованно, через

сумму выплаты: клиенты с дорогими полисами способны получить крупные компенсации, которые, в свою очередь, приводят к уходу.

Claim Reason (причина страхового случая) – категорический признак, отражающий тип страхового события, не показал значимой прогностической силы в отрыве от суммы выплаты. Хотя определенные виды страховых случаев могли быть связаны с повышенной вероятностью оттока, в наших результатах влияние Claim Reason нивелируется преобладающим фактором размера выплаты. Модели, сфокусированные на числовых характеристиках, не выделили определенную категорию причины обращения как значимый драйвер оттока.

BMI (индекс массы тела клиента) – не продемонстрировал заметной взаимосвязи с оттоком (корреляция порядка 0,06 по модулю). Этот демографический/медицинский показатель практически не влиял на решение клиента уйти или остаться. Таким образом, признаки, характеризующие личные особенности клиента (BMI) или, например, регион проживания, оказались статистически незначимыми. Отток в данной выборке определялся не демографией, а опытом взаимодействия с компанией – прежде всего финансовым исходом страхового случая.

Полученные результаты согласуются с интуитивными представлениями о поведении клиентов страховой компании. Высокая значимость суммы выплат говорит о том, что причиной оттока зачастую служит крупное страховое событие. Вероятно, после получения значительной компенсации по страховке клиент либо удовлетворил разовую потребность (например, дорогостоящее лечение) и не видит смысла продолжать полис, либо сталкивается с ростом страховочных взносов/условий и ищет более выгодные альтернативы. Наш анализ фактически выявляет простой сценарий оттока: если выплата по страховому случаю велика, клиент почти наверняка покинет компанию. В связи с этим модели показывают аномально высокое качество – задача прогнозирования сводится к обнаружению порогового значения ключевого признака.

Важно подчеркнуть, что мы исключили из рассмотрения потенциально тривиальные индикаторы оттока: например, поле Claim Request Output (результат рассмотрения случая) не использовалось при моделировании, хотя ожидаемо, что отказ в выплате привел бы к уходу клиента. Исключение подобных утечек информации (leakage) гарантировало честность модели.

Тем не менее, даже без них данные содержат мощный сигнал, облегчающий задачу классификации. Сравнение с доступными открытыми исследованиями подтверждает уникально высокие показатели нашей модели. Ранее в подобных работах по прогнозированию оттока в страховании жизни сообщались существенно более низкие метрики. К примеру, в одном из решений на Kaggle достигнута точность около 81 % при F1-мере ~0,58[10], в других опытах лучшие алгоритмы давали F1 не выше ~0,87 при Accuracy порядка 90 %[11]. Даже сложные ансамбли глубокого обучения в недавних публикациях показывали F1-score на уровне ~0,95 для страховых данных – заметно ниже [7], чем практически 100-процентные значения, полученные в нашем эксперименте. Таким образом, наше решение превосходит по качеству ранее известные результаты на аналогичных наборах данных. Это указывает не столько на преимущество конкретного алгоритма, сколько на особенности самого датасета: высокая предсказуемость оттока обусловлена наличием сильного фактора (размера выплаты). В реальных условиях такие идеальные показатели встречаются редко, однако выявление доминирующего фактора churn-а представляет большую ценность.

[10] Vikashrajluhaniwal. Insurance churn prediction [90%acc & 0.58 F1-score]. Kaggle. URL: https://www.kaggle.com/code/vikashrajluhaniwal/insurance-churn-prediction-90-acc-0-58-f1-score (дата обращения: 21.09.2025).

[11] Ginanjar S. Customer Churn Analysis and Prediction. Kaggle. URL: https://www.kaggle.com/code/ginsaputra/customer-churn-analysis-and-prediction (дата обращения: 21.09.2025).

Наш анализ факторов подтвердил выводы других авторов о важности финансовых показателей в оттоке клиентов страховых компаний, одновременно демонстрируя, что при наличии экстремально информативного признака стандартные алгоритмы машинного обучения способны не уступать по эффективности более сложным моделям.

Заключение

В работе проведена воспроизводимая оценка качества и вычислительной эффективности нескольких подходов к прогнозированию оттока клиентов страховой компании на открытом наборе данных отрасли (жизненное страхование). Экспериментальный конвейер включал строгую предобработку (очистка, типизация, стандартизация числовых признаков и one-hot кодирование категориальных, стратифицированное разбиение, учет дисбаланса весами классов), единые критерии качества (ROC-AUC, PR-AUC, F1/F2, precision/recall, accuracy) и унифицированные замеры производительности (время/память). Сравнение охватывало линейную модель (логистическая регрессия), деревья и ансамбли (Random Forest, Gradient Boosting, XGBoost), SVM, KNN и нейросеть.

Итоги. На рассматриваемом наборе все основные модели достигли практически одинаково высокого качества (Таблица 2: ROC-AUC, F1/F2 и PR-AUC максимальны), что указывает на наличие доминирующего признака риска оттока и хорошую разделимость классов после корректной предобработки. При равной точности ключевым дифференциатором стала производительность.

Настолько высокое качество связано с тем, что поведение клиентов в имеющемся датасете во многом определяется одним ключевым фактором – размером страховой выплаты. Анализ влияния признаков показал, что именно величина полученной компенсации практически однозначно разделяет ушедших и лояльных клиентов: большие выплаты коррелируют с уходом клиента, тогда как застрахованные без значимых выплат, как правило, остаются с компанией. Этот вывод согласуется с бизнес-логикой страхования жизни и здоровья: единичное наступление страхового случая с крупной выплатой нередко ведет к прекращению полиса (либо по инициативе клиента, либо вследствие пересмотра условий страховщиком). Таким образом, выявленный признак «Claim Amount» служит ранним индикатором оттока – модель может заранее сигнализировать о высокой вероятности ухода клиента, получившего значительную выплату.

Практические рекомендации. Практическая значимость результатов состоит в том, что страховая компания, зная о столь сильном влиянии финансового фактора, может сконцентрировать усилия на удержании именно тех клиентов, которые недавно получили крупные страховые возмещения. Например, можно предложить таким клиентам индивидуальные скидки на продление полиса или дополнительные сервисы, чтобы снизить мотивацию к уходу. С научной точки зрения, работа демонстрирует, что для данного набора данных простые интерпретируемые модели (решающие деревья, логистическая регрессия) достигают качества не хуже сложных методов, поскольку ключевой параметр оттока легко интерпретируем и линейно связан с целевой переменной. Мы подтвердили результаты других исследователей в том, что точность прогноза оттока в страховании существенно повышается при учете финансовых характеристик клиента. Наша лучшая модель смогла безошибочно идентифицировать всех клиентов, склонных к оттоку (Recall = 100 %), при минимальном числе ложных тревог (Precision > 95 %).

В качестве базовой производственной модели использовать логистическую регрессию (интерпретируемость, скорость, простота калибровки порога) либо

XGBoost/градиентный бустинг (устойчивость к нелинейностям и взаимодействиям признаков) – в зависимости от ограничений по времени обучения и инфраструктуре. В производственный конвейер включить автоматический мониторинг качества моделей: регулярный расчет ROC-AUC (при необходимости – PR-AUC), контроль дрейфа данных, пороговые оповещения и запуск переобучения при деградации.

Ограничения исследования. Набор данных статичен и специфичен для страхования жизни; результаты не гарантируют аналогичных эффектов в авто или медицинском страховании без адаптации признаков. Балансировка осуществлялась, в основном, через веса классов; альтернативные стратегии (разные режимы пере/довыборки на обучении) не сравнивались вширь, чтобы не смешивать влияние перерасчета признаков.

Направления дальнейших исследований. Временные и последовательностные модели поведения (feature-drift, тренды по премии/выплатам, частота и тональность обращений в поддержку), а также причинно-следственные методики для оценки эффекта удерживающих предложений. Калибровка вероятностей (Platt/Isotonic) и интеграция cost-sensitive функций потерь; оптимизация порога под бизнес-ограничения контакт-центра. Интерпретируемость (глобальные и локальные объяснения коэффициентов/важностей) для верификации доминирующих факторов риска и разработки адресных сценариев удержания. Тестирование устойчивости к сдвигам данных (drift/shift), A/B-эксперименты на реальных когортных разрезах и расширение признакового состава (в т. ч. агрегаты по «жалобам» и статусам рассмотрения).

Выводы. В целом, результаты показывают: при корректной предобработке и учете дисбаланса простые и умеренно сложные алгоритмы (логистическая регрессия, градиентный бустинг/XGBoost) обеспечивают максимальное качество при наилучшем профиле ресурсов. Это подтверждает целесообразность их использования как основы производственного решения для раннего выявления риска оттока и запуска проактивных мер удержания.

Таким образом, исследование подтвердило, что при подобранной структуре признаков простые модели могут эффективно прогнозировать отток в страховой компании. Вместе с тем, выбор модели должен учитывать компромисс между точностью и ресурсами: если даже ансамблевые методы не дают улучшения точности, их высокое время и память не оправданы. Результаты работы подчеркивают, что дополнительное преимущество сложных алгоритмов проявляется только при более сложных данных. Итоговое предложение – использовать для оперативного прогноза оттока в страховых данных, схожих с исследуемыми, хорошо оптимизированную логистическую регрессию или XGBoost (из-за их скорости), оставляя тяжелые модели для случаев, когда в данных действительно имеются нерешенные ранее сложности.

В заключение, проведенное исследование подчеркнуло решающую роль суммы страховой выплаты в задаче churn-анализа для индустрии страхования жизни. Достижение практически идеальных метрик на тестовых данных свидетельствует о том, что отток клиентов в рассматриваемом датасете носит детерминированный характер и может быть предсказан с высокой уверенностью при наличии соответствующих признаков. Однако такой результат требует осторожной интерпретации: в более общих случаях, когда отток обусловлен комбинацией многочисленных факторов (сервис, цена, личные обстоятельства и т. д.), ожидается значительно более скромное качество моделей. Тем не менее, выявление даже одного сильного индикатора оттока, как в данном исследовании, чрезвычайно полезно для бизнеса. В дальнейшем планируется проверка разработанных моделей на расширенных данных и применение более сложных методов интерпретации (например, SHAP-анализа) для подтверждения влияния обнаруженных факторов на склонность клиентов к оттоку.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Шайхиева Ж.М. Churn Modeling для прогнозирования оттока клиентов в предприятиях. *Вестник науки*. 2023;2(11):709–713.
   Shaikhiyeva Zh.M. Churn Modeling for Predicting Customer Churn in Enterprises. *Science Bulletin.* 2023;2(11):709–713. (In Russ.).
2. Зеленков Ю.А., Сучкова А.С. Прогнозирование оттока клиентов на основе паттернов изменения их поведения. *Бизнес-информатика*. 2023;17(1):7–17.
   Zelenkov Yu.A., Suchkova A.S. Predicting Customer Churn Based on Changes in Their Behavior Patterns. *Business Informatics*. 2023;17(1):7–17.
3. Любинский М.С. Применение машинного обучения для прогнозирования оттока клиентов в CRM-системах. *Вестник науки*. 2025;3(4):808–812.
   Liubinskii M.S. Application of Machine Learning for Customer Churn Prediction in CRM Systems. *Science Bulletin.* 2025;3(4):808–812. (In Russ.).
4. Boozary P., Sheykhan S., GhorbanTanhaei H., Magazzino C. Enhancing Customer Retention with Machine Learning: A Comparative Analysis of Ensemble Models for Accurate Churn Prediction. *International Journal of Information Management Data Insights*. 2025;5(1). https://doi.org/10.1016/j.jjimei.2025.100331
5. Suguna R., Prakash J.S., Pai H.A., Mahesh T.R., Kumar V.V., Yimer T.E. Mitigating Class Imbalance in Churn Prediction with Ensemble Methods and SMOTE. *Scientific Reports.* 2025;15. https://doi.org/10.1038/s41598-025-01031-0
6. Sikri A., Jameel R., Idrees Sh.M., Kaur H. Enhancing Customer Retention in Telecom Industry with Machine Learning Driven Churn Prediction. *Scientific Reports*. 2024;14. https://doi.org/10.1038/s41598-024-63750-0
7. AbdelAziz N.M., Bekheet M., Salah A., El-Saber N., AbdelMoneim W.T. A Comprehensive Evaluation of Machine Learning and Deep Learning Models for Churn Prediction. *Information*. 2025;16(7). https://doi.org/10.3390/info16070537
8. Liu X., Xia G., Zhang X., Ma W., Yu Ch. Customer Churn Prediction Model Based on Hybrid Neural Networks. *Scientific Reports*. 2024;14. https://doi.org/10.1038/s41598-024-79603-9
9. He Ch., Ding Ch.H.Q. A Novel Classification Algorithm for Customer Churn Prediction Based on Hybrid Ensemble-Fusion Model. *Scientific Reports*. 2024;14. https://doi.org/10.1038/s41598-024-71168-x
10. Khattak A., Mehak Z., Ahmad H., Asghar M.U., Asghar M.Z., Khan A. Customer Churn Prediction Using Composite Deep Learning Technique. *Scientific Reports*. 2023;13. https://doi.org/10.1038/s41598-023-44396-w
11. Peng K., Peng Ya., Li W. Research on Customer Churn Prediction and Model Interpretability Analysis. *PLoS ONE*. 2023;18(12). https://doi.org/10.1371/journal.pone.0289724
12. Risselada H., Verhoef P.C., Bijmolt T.H.A. Staying Power of Churn Prediction Models. *Journal of Interactive Marketing.* 2010;24(3):198–208. https://doi.org/10.1016/j.intmar.2010.04.002

ИНФОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Мадияров Куан Галымович, аспирант, Новосибирский государственный университет экономики и управления «НИНХ», Новосибирск, Российская Федерация.
e-mail: kuan.mad@mail.ru
ORCID: 0009-0002-1447-4357

Kuan G. Madiyarov, Postgraduate, Novosibirsk State University of Economics and Management, Novosibirsk, the Russian Federation.

Статья поступила в редакцию 21.08.2025; одобрена после рецензирования 30.09.2025; принята к публикации 08.10.2025.

The article was submitted 21.08.2025; approved after reviewing 30.09.2025; accepted for publication 08.10.2025.