http://dx.doi.org/10.35596/1729-7648-2025-31-1-31-41

УДК 331.108.2

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ТЕКУЧЕСТИ КАДРОВ НА ОСНОВЕ ОТКРЫТЫХ ДАННЫХ

А. Н. КОЗИНЕЦ

Белорусский государственный университет информатики и радиоэлектроники (Минск, Республика Беларусь)

Аннотация. Исследовано применение методов машинного обучения для прогнозирования текучести кадров в организациях с использованием открытых данных. Проведен анализ существующих подходов к прогнозированию текучести персонала, обоснована необходимость использования современных алгоритмов машинного обучения. На базе открытого набора данных разработана модель, позволяющая с высокой точностью определять вероятность увольнения сотрудников. Результаты исследования демонстрируют практическую значимость предлагаемого подхода и могут быть использованы для повышения эффективности управления человеческими ресурсами в организациях. Представлены формальные описания и архитектура применяемых моделей машинного обучения, что обеспечивает прозрачность и воспроизводимость рассматриваемого подхода.

Ключевые слова: текучесть кадров, машинное обучение, прогнозирование, управление человеческими ресурсами, открытые данные, HR-аналитика.

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Для цитирования. Козинец, А. Н. Применение методов машинного обучения для прогнозирования текучести кадров на основе открытых данных / А. Н. Козинец // Цифровая трансформация. 2025. Т. 31, № 1. С. 31–41. http://dx.doi.org/10.35596/1729-7648-2025-31-1-31-41.

APPLICATION OF MACHINE LEARNING METHODS FOR EMPLOYEE TURNOVER PREDICTION BASED ON OPEN DATA

ALIAKSANDR N. KAZINETS

Belarusian State University of Informatics and Radioelectronics (Minsk, Republic of Belarus)

Abstract. The application of machine learning methods for predicting staff turnover in organizations using open data is studied. An analysis of existing approaches to predicting staff turnover is conducted, the need to use modern machine learning algorithms is substantiated. Based on an open data set, a model is developed that allows for a high-precision determination of the probability of employee dismissal. The results of the study demonstrate the practical significance of the proposed approach and can be used to improve the efficiency of human resource management in organizations. Formal descriptions and architecture of the applied machine learning models are presented, which ensures the transparency and reproducibility of the approach under consideration.

Keywords: employee turnover, machine learning, prediction, human resource management, open data, HR analytics.

Conflict of interests. The author declares no conflict of interests.

For citation. Kazinets A. N. (2025) Application of Machine Learning Methods for Employee Turnover Prediction Based on Open Data. *Digital Transformation.* 31 (1), 31–41. http://dx.doi.org/10.35596/1729-7648-2025-31-1-31-41 (in Russian).

Введение

В условиях активной цифровой трансформации экономики и стремительного роста объемов доступных данных эффективное управление человеческими ресурсами становится стратегически важным фактором обеспечения конкурентоспособности и устойчивого развития организаций. Человеческий капитал рассматривается как ключевой ресурс, определяющий инновационный потенциал, гибкость и адаптивность предприятия к динамично изменяющейся рыночной среде. Однако текучесть кадров по-прежнему остается одной из наиболее острых и затратных проблем, вызывая существенные финансовые потери, снижение производительности, утрату накопленного опыта и ослабление корпоративной культуры.

Традиционные подходы к прогнозированию текучести, основанные главным образом на статистических методах и экспертных оценках, не учитывают многомерную природу и нелинейные зависимости факторов, влияющих на решение сотрудника об увольнении. Это создает потребность в более продвинутых методах анализа и прогнозирования, способных обрабатывать крупномасштабные и разнородные данные, а также выявлять скрытые закономерности в поведении персонала [1].

Современные методы машинного обучения, опирающиеся на алгоритмы искусственного интеллекта и анализ больших данных, открывают новые возможности для повышения точности прогнозов, выявления наиболее значимых факторов текучести и проактивного управления кадровыми ресурсами. Применение таких методов в HR-аналитике позволяет организациям своевременно определять сотрудников с высоким риском увольнения, разрабатывать адресные стратегии удержания и повышать эффективность HR-процессов.

Цель исследования – разработка и эмпирическая проверка модели прогнозирования текучести кадров с использованием методов машинного обучения на основе открытых данных. Для ее достижения необходимо:

- проанализировать существующие подходы к прогнозированию текучести и определить их ограничения;
- обосновать выбор наиболее результативных алгоритмов машинного обучения;
- обучить и протестировать разработанные модели на открытом наборе данных;
- интерпретировать результаты с позиций практического использования в управлении человеческими ресурсами.

Научная новизна исследования заключается в интеграции передовых алгоритмов машинного обучения для решения задачи прогнозирования текучести, позволяя учесть сложные многомерные и нелинейные зависимости. Применение открытых данных обеспечивает воспроизводимость и прозрачность результатов, соответствующих принципам современной научной методологии, и дает возможность другим исследователям верифицировать и развивать предложенный подход. Практическая значимость состоит во внедрении разработанной модели в практику управления персоналом, что позволит снизить уровень текучести, оптимизировать затраты, повысить устойчивость и конкурентоспособность организации. В статье представлено описание архитектуры и принципов функционирования используемых моделей машинного обучения (логистической регрессии, случайного леса, XGBoost), обеспечивающее концептуальную ясность и воспроизводимость методологии, а также формирующее основу для ее дальнейшего совершенствования и расширения.

Анализ существующих подходов к прогнозированию текучести кадров

Текучесть кадров представляет собой сложный и многогранный феномен, оказывающий существенное влияние на эффективность и устойчивость организаций. Она приводит к значительным прямым и косвенным издержкам, способным подорвать конкурентоспособность предприятия и его позиции на рынке. Прямые издержки включают расходы на рекрутинг, отбор и адаптацию новых сотрудников, такие как размещение вакансий, проведение собеседований и обучение персонала. Косвенные издержки проявляются в снижении производительности труда вследствие утраты накопленного опыта и знаний, нарушении рабочих процессов, ухудшении морально-психологического климата и снижении удовлетворенности оставшихся сотрудников.

Традиционные методы прогнозирования текучести кадров, основанные на статистическом анализе исторических данных и экспертных оценках, часто опираются на субъективные мнения

и ограниченный набор параметров. Эти методы обычно применяют линейные модели и предполагают независимость факторов, что не позволяет учитывать сложные и нелинейные взаимосвязи между переменными, влияющими на решение сотрудника об увольнении. Кроме того, они неэффективны при обработке больших объемов данных и недостаточно адаптивны к быстро меняющимся условиям внешней и внутренней среды организации.

В условиях усложнения управленческих задач и увеличения доступности данных становится необходимым внедрение современных методов анализа, способных выявлять скрытые закономерности в многомерных данных. Машинное обучение, являясь важным направлением искусственного интеллекта, предлагает широкий спектр алгоритмов и инструментов для обработки больших данных, обнаружения нелинейных зависимостей и прогнозирования сложных явлений. В сфере HR-аналитики методы машинного обучения позволяют разрабатывать модели, которые точно прогнозируют вероятность увольнения сотрудников, оценивают их потенциал и оптимизируют процессы подбора и развития персонала.

Применение машинного обучения в прогнозировании текучести кадров открывает новые возможности для проактивного управления человеческими ресурсами. Например, алгоритмы классификации могут сегментировать сотрудников по уровню риска увольнения, а методы регрессии и кластеризации – выявлять ключевые факторы текучести и разрабатывать индивидуализированные стратегии удержания. Кроме того, модели машинного обучения способны постоянно улучшаться с накоплением новых данных, повышая свою адаптивность и актуальность в условиях динамичной среды.

Важным аспектом современного исследования является использование открытых данных, которые позволяют проводить масштабные и репрезентативные исследования без затратных процессов сбора собственных данных. Открытые данные способствуют развитию открытой науки, повышению прозрачности и воспроизводимости исследований. Однако работа с такими данными требует внимания к их качеству, релевантности и соблюдения этических норм, включая конфиденциальность и защиту персональных данных. Использование открытых данных в прогнозировании текучести кадров позволяет исследователям применять и сравнивать различные модели и подходы, что способствует обобщению знаний и разработке более универсальных решений. Тем не менее необходимо учитывать ограничения, связанные со спецификой отраслей, культурными особенностями и актуальностью данных, что требует тщательного анализа и корректной интерпретации результатов.

Таким образом, анализ существующих подходов к прогнозированию текучести кадров демонстрирует, что традиционные методы имеют существенные ограничения и не соответствуют современным требованиям эффективного управления человеческими ресурсами. Применение методов машинного обучения в сочетании с использованием открытых данных представляет собой перспективное направление, способное значительно повысить точность прогнозирования и предоставить организациям инструменты для принятия обоснованных решений в сфере HR-менеджмента.

Методология исследования

В исследовании для разработки модели прогнозирования текучести кадров с использованием методов машинного обучения применялся открытый набор данных HR Analytics Employee Attrition & Performance, доступный на платформе Kaggle [2]. Этот набор включает информацию о 1470 сотрудниках, охватывая такие параметры, как демографические характеристики, профессиональные показатели, уровни удовлетворенности и производительности, а также факт увольнения. В совокупности предусмотрено 35 признаков, среди которых целевой переменной выступает Attrition (Yes/No) (табл. 1).

Таблица 1. Основные характеристики набора данных
Table 1. Main characteristics of the dataset

Параметр	Описание
Количество записей	1470
Количество признаков	35
Целевая переменная	Attrition (факт увольнения: Yes/No)

Набор данных характеризуется разнообразием признаков, включающих:

- демографические данные: возраст, пол, семейное положение, уровень образования;
- профессиональные данные: должность, отдел, стаж работы;
- показатели удовлетворенности: удовлетворенность работой, взаимоотношения в коллективе;
- показатели производительности: количество проектов, количество часов переработки.

Качественная предварительная обработка данных является критически важным этапом построения модели машинного обучения, поскольку качество входных данных напрямую влияет на точность и надежность модели. Этапы обработки данных включали следующие шаги:

1) анализ пропущенных значений: в этом наборе данных отсутствуют пропущенные значения, что исключает необходимость проведения иммутации или удаления записей;

2) кодирование категориальных признаков: Label Encoding применялся ко всем категориальным признакам (например, Gender, Department, JobRole, EducationField), преобразуя категории в числовые значения (0 и 1 или более);

3) нормализацию числовых признаков: проводилась стандартизация (StandardScaler) для обеспечения сопоставимости масштабов признаков и ускорения сходимости алгоритмов.

Для решения задачи бинарной классификации и повышения точности прогнозирования были выбраны следующие алгоритмы машинного обучения, зарекомендовавшие себя в области HR-аналитики [3]:

– логистическая регрессия: простая и интерпретируемая модель с низкими вычислительными затратами, однако ограниченная линейностью;

– случайный лес (Random Forest): ансамблевый метод, устойчивый к переобучению, способный моделировать сложные нелинейные зависимости и оценивать важность признаков;

– градиентный бустинг (XGBoost): высокопроизводительный алгоритм, оптимизированный для скорости и точности, с гибкими настройками гиперпараметров.

Процесс машинного обучения был реализован на языке Python с использованием библиотек Scikit-learn и XGBoost и включал следующие этапы.

1. Предварительная обработка данных:

– кодирование категориальных переменных с помощью LabelEncoder;

– масштабирование числовых признаков с использованием StandardScaler;

– разделение выборки на обучающую (80 %) и тестовую (20 %) с фиксированным параметром random_state=42 для воспроизводимости результатов.

2. Обучение моделей:

– логистическая регрессия с параметром max_iter=1000;

– Random Forest с базовыми параметрами;

– XGBoost с отключенным LabelEncoder и метрикой logloss.

3. Оценка качества моделей:

– расчет метрик: accuracy, precision, recall, f1_score;

– анализ важности признаков на основе модели Random Forest.

Набор данных был разделен на обучающую и тестовую выборки в соотношении 80:20 с использованием функции train_test_split из библиотеки Scikit-learn, что позволило оценить обобщающую способность моделей на новых данных. Для оценки эффективности моделей использовались следующие метрики:

– accuracy (точность): доля правильно классифицированных наблюдений;

– precision (точность позитивного прогноза): доля правильно предсказанных положительных случаев;

– recall (полнота): доля правильно предсказанных положительных случаев из всех фактических;

– f1 score: гармоническое среднее между precision и recall, особенно полезное при несбалансированных данных [3].

Для учета многомерности и нелинейности данных применялись:

– Random Forest, способный автоматически учитывать нелинейные взаимодействия между признаками через построение множества деревьев решений;

– XGBoost, использующий градиентный бустинг для моделирования сложных зависимостей;

– стандартизация признаков (StandardScaler) для обеспечения корректной работы с многомерными данными.

Для каждой модели были настроены ключевые гиперпараметры [4, 5]:

● логистическая регрессия: регуляризация (C), функция потерь;

● случайный лес: количество деревьев (n_estimators), максимальная глубина (max_depth), количество признаков для разбиения (max_features);

● XGBoost: скорость обучения (learning_rate), максимальная глубина (max_depth), количество деревьев (n_estimators), параметр регуляризации (gamma).

Используемые инструменты и библиотеки на языке Python включают:

● Pandas и NumPy: для обработки и анализа данных;

● Scikit-learn: для реализации алгоритмов машинного обучения, предобработки данных, настройки гиперпараметров и оценки моделей;

● XGBoost: для реализации градиентного бустинга;

● Matplotlib и Seaborn: для визуализации данных и результатов анализа.

Особое внимание уделялось конфиденциальности и соблюдению этических норм при работе с данными. Использование открытого набора данных, предназначенного для исследовательских целей, обеспечивает соблюдение требований конфиденциальности и защиту персональных данных сотрудников.

Примененная методология сочетает передовые практики обработки данных и машинного обучения, что обеспечивает получение надежных и интерпретируемых результатов. Такой подход гарантирует высокую точность прогнозирования и практическую применимость разработанных моделей в управлении человеческими ресурсами. В целях обеспечения прозрачности и воспроизводимости в статье приводится формальное описание применяемых моделей машинного обучения, включая их математические основы и архитектуру. Это позволит более полно понять принципы функционирования рассмотренных алгоритмов до перехода к анализу результатов их применения.

Описание моделей

Под разработанными моделями подразумеваются конкретные, обученные на реальных данных алгоритмы машинного обучения, сконфигурированные для решения задачи прогнозирования текучести кадров. Каждая модель представляет собой формализованную математическую конструкцию (классификатор), которая по входным характеристикам сотрудника оценивает вероятность его увольнения.

Прогнозирование текучести кадров сводится к задаче бинарной классификации. Пусть $X \in R^n$ – вектор признаков, описывающих отдельного сотрудника. Эти признаки могут включать демографические данные (возраст, пол), характеристики занятости (должность, стаж), показатели удовлетворенности (оценка рабочего климата, отношение к переработкам), а также параметры компенсационной политики (месячный доход, почасовые ставки). Целевая переменная $y \in \{0,1\}$ указывает факт увольнения: $y = 1$ – если сотрудник ушел из организации, $y = 0$ – если он остался. Цель модели – оценить вероятность $P(y = 1|X)$.

На вход модели подается вектор признаков X, прошедший этапы предобработки (масштабирование, кодирование категориальных переменных). Результатом работы модели является численное значение вероятности увольнения. Если эта вероятность превышает некоторый порог (обычно 0,5), модель классифицирует сотрудника как имеющего высокий риск увольнения. Данный подход позволяет на основе открытых данных о сотрудниках выявлять «уязвимые» группы персонала и предпринимать превентивные меры по удержанию ключевых специалистов. Логистическая регрессия — классический линейный классификатор, предполагающий, что логарифм отношения шансов увольнения к удержанию является линейной комбинацией признаков. Формально

$$\log it\left(P(y=1 \mid X)\right) = w_0 + \sum_{i=1}^{n} w_i x_i,$$

где w_i – параметр модели.

Вероятность увольнения определяется через логистическую функцию

$$P(y=1|X)=\frac{1}{1+e^{-\left(w_0+\sum_{i=1}^{n}w_i x_i\right)}}.$$

Логистическая регрессия легко интерпретируется: знаки и значения коэффициентов w_i указывают направление и степень влияния соответствующего признака на вероятность увольнения. Это делает модель удобной для первичной аналитики, позволяя HR-специалистам понять, какие факторы сильнее всего воздействуют на риск текучести.

Случайный лес – ансамблевый метод, представляющий собой совокупность M деревьев решений. Каждое дерево обучается на случайной подвыборке исходных данных и случайном подмножестве признаков, что снижает коррелированность деревьев и уменьшает риск переобучения. Предсказание ансамбля получается путем голосования

$$\hat{y}=\underset{c\in\{0,1\}}{\operatorname{argmax}}\sum_{k=1}^{M}I\left(\hat{y}_k=c\right),$$

где $\hat{y}_k$ – предсказание модели; $I(.)$ – индикаторная функция.

Случайный лес способен улавливать сложные нелинейные закономерности и, что особенно важно для анализа текучести, предоставляет механизм оценки важности признаков. Это позволяет определить, какие факторы (например, уровень дохода или переработки) оказывают наибольшее влияние на риск увольнения.

XGBoost – эффективный алгоритм градиентного бустинга, который итеративно строит последовательность деревьев решений, каждое из которых направлено на уменьшение ошибки предыдущего ансамбля. Итоговая модель может быть представлена в следующем виде:

$$f(x)=\sum_{k=1}^{K}\gamma_k h_k(x),$$

где $h_k(x)$ – дерево-регрессор на k-й итерации; γ_k – соответствующий вес.

Процесс обучения сводится к минимизации функции потерь, учитывающей как точность предсказания, так и регуляризацию, что обеспечивает высокую устойчивость к переобучению. XGBoost славится высокой производительностью и возможностью тонкой настройки параметров, позволяя достичь оптимального баланса между скоростью обучения, точностью и обобщающей способностью модели.

На практике для каждого сотрудника X-алгоритм (логистическая регрессия, случайный лес или XGBoost) вычисляет $P(y = 1|X)$. Если вероятность высока, руководство может своевременно принять меры, направленные на удержание ценного специалиста: пересмотреть систему компенсаций, предложить карьерный рост или оптимизировать рабочую нагрузку. Таким образом, формальные математические описания моделей непосредственно связаны с реальной управленческой задачей – со снижением текучести кадров за счет вовремя предпринятых действий. Для большей наглядности на рис. 1 представлена общая схема применения моделей.

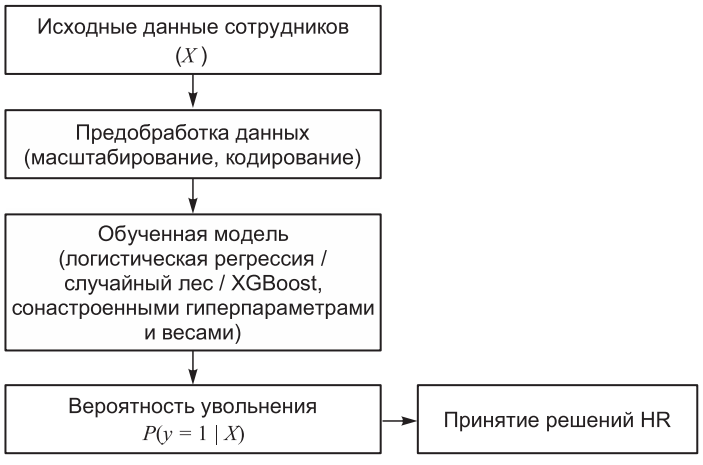

Рис. 1. Применение разработанных моделей для прогнозирования текучести кадров
Fig. 1. Application of the developed models for forecasting employee turnover

Для достижения наилучших результатов в рамках поставленной задачи были эмпирически выбраны оптимальные значения гиперпараметров. В случае логистической регрессии использовалось число итераций max_iter=1000, что обеспечило сходимость алгоритма. Для случайного леса подбирались число деревьев (n_estimators), максимальная глубина (max_depth) и критерий разбиения, позволяющие учесть сложные взаимосвязи и избежать переобучения. Для XGBoost варьировались скорость обучения (learning_rate), максимальная глубина деревьев (max_depth) и число итераций (n_estimators), а также параметры регуляризации, что позволило достичь высокой точности и устойчивости результатов.

Таким образом, сформированы и настроены три модели машинного обучения: логистическая регрессия, случайный лес и XGBoost. Каждая из них обладает формальной математической основой, четко определяемой архитектурой и набором подогнанных под данные гиперпараметров. Такой подход обеспечивает как высокую точность прогнозирования текучести кадров, так и возможность интерпретации полученных результатов, что непосредственно способствует принятию обоснованных управленческих решений в сфере HR.

Результаты исследований и их обсуждение

Разработанная система прогнозирования текучести кадров основана на применении трех моделей машинного обучения: логистической регрессии, случайного леса (Random Forest) и градиентного бустинга (XGBoost). Для воспроизводимости и объективности экспериментов был использован единый процесс предобработки данных, включающий кодирование категориальных признаков (LabelEncoder) и стандартизацию числовых переменных (StandardScaler), а также фиксированное разбиение выборки на обучающую (80 %) и тестовую (20 %) с параметром random_state=42. Такой подход гарантировал сопоставимость полученных результатов. На тестовой выборке были вычислены ключевые метрики: accuracy, precision, recall и f1 score, значения которых представлены в табл. 2.

Таблица 2. Результаты оценки моделей
Table 2. Results of model evaluation

Модель	Метрика			
	accuracy	precision	recall	f1 score
Логистическая регрессия	0,891156	0,684211	0,333333	0,448276
Случайный лес	0,867347	0,500000	0,102564	0,170213
XGBoost	0,877551	0,588235	0,256410	0,357143

Анализ показал, что все три рассмотренные модели обеспечивают высокую точность классификации (более 86 %). Логистическая регрессия продемонстрировала наилучшие результаты по всем основным метрикам: accuracy = 0,891156, precision = 0,684211, recall = 0,333333 и f1 score = 0,448276. Это свидетельствует о сбалансированном соотношении между точностью и полнотой при идентификации сотрудников с высоким риском увольнения. Модель XGBoost, хотя и уступила логистической регрессии, превзошла случайный лес по precision и recall. Случайный лес продемонстрировал высокую accuracy (0,867347), но оказался менее эффективным по recall (0,102564), что затрудняет своевременное выявление сотрудников, действительно склонных к увольнению. Таким образом, логистическая регрессия была признана моделью с наилучшим качеством прогнозирования.

После определения наилучшей модели по качественным показателям проводился интегрированный анализ важности признаков для всех трех алгоритмов. Для логистической регрессии в качестве меры важности использовались абсолютные значения коэффициентов, отражающие вклад каждого признака в логарифм отношения шансов. Для случайного леса и XGBoost применялись встроенные механизмы оценки значимости признаков. Сопоставление результатов после нормализации позволило сформировать сводную сравнительную таблицу (табл. 3).

Таблица 3. Сравнение важности признаков для всех моделей
Table 3. Comparison of feature importance for all models

Наименование признака	Логистическая регрессия	Случайный лес	XGBoost
OverTime	0,099822	0,064823	0,105581
YearsAtCompany	0,080857	0,041914	0,043086
YearsInCurrentRole	0,077865	0,027299	0,031993
YearsSinceLastPromotion	0,055244	0,026549	0,034487
NumCompaniesWorked	0,054052	0,036035	0,033700
MaritalStatus	0,052193	0,025151	0,052048
YearsWithCurrManager	0,051192	0,028512	0,021118
JobLevel	0,050121	0,026893	0,057950
JobSatisfaction	0,047524	0,025368	0,032001
Department	0,044301	0,011564	0,051286
TotalWorkingYears	0,043679	0,048104	0,047878
EnvironmentSatisfaction	0,042066	0,024602	0,026775
JobInvolvement	0,036494	0,022657	0,037186
DistanceFromHome	0,032076	0,042012	0,026622
Age	0,026867	0,056865	0,026889
WorkLifeBalance	0,023424	0,019841	0,025055
StockOptionLevel	0,022516	0,03184	0,063194
Gender	0,022028	0,007987	0,013753
MonthlyIncome	0,02161	0,074968	0,037411
RelationshipSatisfaction	0,018697	0,020104	0,017982
TrainingTimesLastYear	0,018343	0,025413	0,018744
JobRole	0,015388	0,032731	0,025498
PercentSalaryHike	0,013914	0,03253	0,020214
EducationField	0,011636	0,023972	0,020013
DailyRate	0,010048	0,050505	0,021936
MonthlyRate	0,008955	0,04709	0,019752
Education	0,008867	0,018488	0,024337
PerformanceRating	0,00463	0,004958	0
EmployeeNumber	0,002301	0,045296	0,024224
HourlyRate	0,001684	0,043352	0,021568
BusinessTravel	0,001606	0,012577	0,017706
EmployeeCount	0	0	0
StandardHours	0	0	0
Over18	0	0	0

Табл. 3 наглядно демонстрирует как совпадения, так и расхождения в оценке важности признаков тремя моделями. Так, факторы, связанные с оплатой труда и переработками (например, MonthlyIncome, OverTime), подтверждают свою ключевую роль во всех алгоритмах, указывая на их универсальное влияние на риск увольнения. Напротив, ряд признаков (таких как YearsInCurrentRole или JobLevel) проявляет себя неодинаково в разных моделях, что отражает специфику выявления закономерностей в линейных (логистическая регрессия) и более сложных нелинейных (случайный лес, XGBoost) методах. Отдельные признаки получили нулевое значение важности в одной или нескольких моделях (например, EmployeeCount, StandardHours, Over18), что говорит о том, что данные алгоритмы не обнаружили статистически значимого вклада этих факторов в оценку риска увольнения. Таким образом, параллельное использование нескольких моделей позволяет повысить интерпретируемость результатов, учитывать широкий спектр детерминант текучести и дает основания для более точной настройки управленческих решений в HR-сфере.

Для повышения наглядности был проведен сравнительный анализ десяти наиболее значимых признаков, нормированных относительно суммы их важностей в каждой модели (рис. 2).

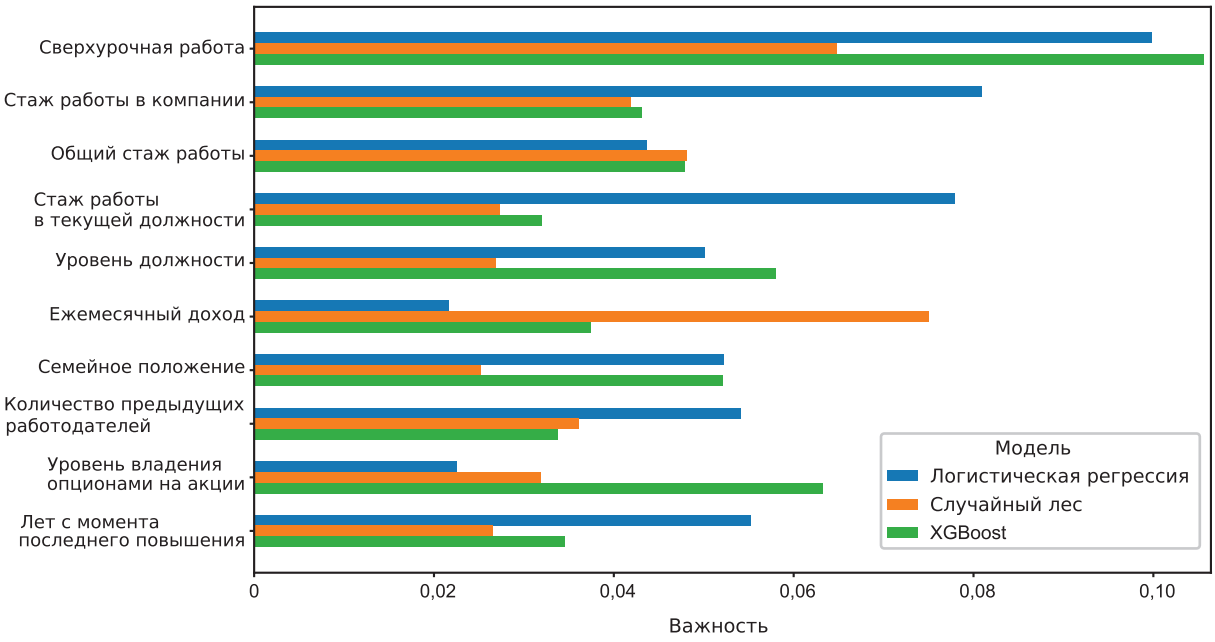

Рис. 2. Сравнение нормализованной важности топ-10 признаков между моделями
Fig. 2. Comparison of normalized importance of top 10 features between models

Нормализация позволяет привести значения важности к единому масштабу, упрощая сравнение вклада факторов между моделями. Визуализация подтверждает выявленные ранее закономерности: признаки, устойчиво проявляющие себя как значимые в разных моделях, следует рассматривать в качестве приоритетных целей для управленческого воздействия. Различия в оценке важности отдельных факторов между моделями могут указывать на потенциальные резервы для дальнейшего уточнения модели, расширения набора признаков или применения более продвинутых методов анализа.

Интегрированный анализ выявил, что основными детерминантами риска увольнения выступают показатели переработок (OverTime), трудового стажа (YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion), а также такие параметры, как уровень должности (JobLevel), размер заработной платы (MonthlyIncome), количество предыдущих мест работы (NumCompaniesWorked), семейное положение (MaritalStatus) и наличие опционов на акции (StockOptionLevel). В то же время различия в оценке важности отдельных факторов свидетельствуют о том, что разные алгоритмы по-своему улавливают и интерпретируют связи между признаками и вероятностью увольнения. Такой интегрированный анализ способствует более точной настройке методов прогнозирования и подталкивает к расширению набора признаков или применению более продвинутых алгоритмов, если возникают несовпадения в выявленных закономерностях. Значение практической ценности выводов иллюстрирует пример высокой интенсивности переработок при относительно низком уровне дохода, которая существенно повышает риск увольнения. Для снижения текучести кадров организации могут оптимизировать нагрузку, пересмотреть компенсационные пакеты и более целенаправленно контролировать карьерные перспективы сотрудников. Этот подход обеспечивает системное понимание сложных взаимодействий факторов и облегчает принятие управленческих решений в направлении повышения удовлетворенности персонала и устойчивости HR-политики.

Таким образом, использованный подход, основанный на сравнительном анализе производительности и важности признаков для нескольких моделей, не только повышает точность прогнозирования, но и укрепляет обоснованность принимаемых управленческих решений в сфере управления персоналом. Стратегии, основанные на данных, позволяют более эффективно использовать человеческий капитал, снижать текучесть и повышать конкурентоспособность организации.

Практическая значимость и рекомендации

Результаты проведенного исследования обладают высокой практической значимостью для организаций, стремящихся оптимизировать управление человеческими ресурсами и снизить

уровень текучести кадров. Разработанная модель прогнозирования текучести на основе методов машинного обучения представляет собой эффективный инструмент для принятия обоснованных управленческих решений и реализации проактивных стратегий в области HR-менеджмента. Модель может быть эффективно использована организациями для:

– идентификации сотрудников с высоким риском увольнения: модель обеспечивает точное прогнозирование вероятности увольнения каждого сотрудника, что позволяет своевременно выявлять потенциально рисковых сотрудников. Это создает основу для разработки индивидуализированных мер по удержанию ключевых кадров, включая программы профессионального развития, персонализированные мотивационные схемы и улучшение условий труда;

– оптимизации программ мотивации и компенсации: анализ значимых факторов, влияющих на текучесть, таких как сверхурочная работа, уровень дохода и удовлетворенность работой, предоставляет возможность пересматривать и совершенствовать существующие программы мотивации и компенсации. Это способствует повышению удовлетворенности сотрудников и снижению уровня текучести. Например, внедрение гибких графиков работы, пересмотр политики оплаты труда в соответствии с рыночными тенденциями и обеспечение возможностей карьерного роста;

– планирования кадровой политики: прогнозирование текучести с использованием модели позволяет более точно планировать кадровую политику, включая прогнозирование потребностей в персонале, планирование найма и обучение новых сотрудников. Это способствует снижению издержек, связанных с незапланированным уходом сотрудников, и повышению эффективности HR-процессов.

Для успешного внедрения разработанной модели прогнозирования текучести кадров рекомендуются:

– интеграция модели в информационные системы управления персоналом (HRIS): эффективное использование модели предполагает ее интеграцию в существующие HRIS организации. Это позволит автоматизировать процессы сбора данных, обновления модели и предоставления результатов прогнозирования в режиме реального времени, что повысит оперативность и точность принятия управленческих решений;

– обучение HR-специалистов: успешное внедрение модели требует повышения квалификации сотрудников HR-отдела в области анализа данных и интерпретации результатов моделей машинного обучения. Рекомендуется проведение специализированных тренингов и обучающих программ, чтобы HR-специалисты могли эффективно использовать результаты модели и принимать обоснованные решения на основе данных;

– учет этических и правовых аспектов: при использовании модели необходимо обеспечить конфиденциальность данных сотрудников и соблюдение нормативно-правовых требований, включая законодательство о защите персональных данных. Организациям следует разработать и внедрить политику защиты данных, предотвращающую несанкционированный доступ и использование персональной информации. Также важно обеспечить прозрачность алгоритмов модели и исключить возможную дискриминацию при принятии решений на основе прогнозов;

– постоянное обновление и совершенствование модели: для поддержания высокой точности и актуальности модели рекомендуется регулярно обновлять ее на основе новых данных и проводить мониторинг производительности. Это позволит учитывать изменения во внешней и внутренней среде организации, адаптироваться к новым условиям и поддерживать эффективность прогнозирования;

– оценка эффективности внедрения: организациям следует регулярно оценивать эффективность внедрения модели, анализируя показатели текучести кадров до и после ее внедрения, а также оценивать экономическую эффективность и возврат инвестиций. Это позволит определить степень достижения поставленных целей и принять решения о дальнейшем развитии и совершенствовании модели.

Разработка и внедрение модели прогнозирования текучести кадров на основе методов машинного обучения представляет собой перспективное направление для организаций, стремящихся повысить эффективность управления персоналом и снизить издержки, связанные с текучестью кадров. Реализация предложенных рекомендаций позволит организациям получить конкурентные преимущества за счет более эффективного использования человеческого капитала и принятия обоснованных управленческих решений.

Заключение

1. Разработана и эмпирически проверена модель прогнозирования текучести кадров на основе методов машинного обучения, использующих открытые данные о сотрудниках. Применение логистической регрессии, случайного леса и градиентного бустинга (XGBoost) позволило не только достичь высокой точности прогнозирования, но и выявить ключевые факторы, определяющие вероятность увольнения. Наиболее эффективной оказалась логистическая регрессия, продемонстрировавшая оптимальный баланс между полнотой и точностью позитивных прогнозов.

2. Выделенные значимые признаки, такие как месячный доход, сверхурочная работа, возраст и показатели стажа, представляют ценную информацию для управленцев при разработке стратегий удержания талантов и повышения удовлетворенности сотрудников. Интеграция разработанной модели в информационные системы управления персоналом (HRIS) позволит перейти к проактивному, основанному на данных, кадровому менеджменту, снижая риск невосполнимых кадровых потерь и повышая общую эффективность работы организации.

3. Методологическая прозрачность алгоритмов машинного обучения повышает воспроизводимость полученных результатов, создает условия для их критической оценки и дальнейшего совершенствования подхода. Перспективными направлениями развития представляются расширение набора признаков, анализ временных рядов, применение глубоких нейронных сетей и изучение влияния организационной культуры и внешних факторов. Также целесообразны исследование экономического эффекта от внедрения модели и разработка рекомендационных систем, предлагающих индивидуализированные меры по удержанию персонала.

4. Проведенное исследование свидетельствует о целесообразности и эффективности интеграции методов машинного обучения в HR-аналитику. Полученные результаты закладывают основу для дальнейшего развития научной и практической базы по управлению человеческими ресурсами, способствуя формированию более устойчивой, инновационной и адаптивной кадровой политики в условиях цифровой экономики.

Список литературы / References

1. Becker B. E., Huselid M. A. (1998) High Performance Work Systems and Firm Performance: A Synthesis of Research and Managerial Implications. *Research in Personnel and Human Resources Management*. 16, 53–101.
2. Subhash P. (2017) IBM HR Analytics Employee Attrition & Performance. *Kaggle*. Available: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset (Accessed 1 October 2024).
3. Kotsiantis S. B., Zaharakis I., Pintelas P. (2007) Supervised Machine Learning: A Review of Classification Techniques. *In Emerging Artificial Intelligence Applications in Computer Engineering*. 160–175.
4. Russell S. J., Norvig P. (2020) Artificial Intelligence: A Modern Approach. *Pearson*.
5. Kazinets A. (2024) *Application of Machine Learning Methods for Employee Turnover Prediction Based on Open Data*. Available: https://colab.research.google.com/drive/1p6yQv8rZm8jIdtFc2DxYmwROXIGbDH-F?usp=sharing (Accessed 1 October 2024).

Поступила 16.11.2024 Принята в печать 16.01.2025 Доступна на сайте 10.04.2025
Received: 16 November 2024 Accepted: 16 January 2025 Available on the website: 10 April 2025

Сведения об авторе

Козинец А. Н., асп. каф. экономики, Белорусский государственный университет информатики и радиоэлектроники

Адрес для корреспонденции

220013, Республика Беларусь,
Минск, ул. П. Бровки, 6
Белорусский государственный университет
информатики и радиоэлектроники
Тел.: +375 17 293-80-46
E-mail: kozinets.science@gmail.com
Козинец Александр Николаевич

Information about the author

Kazinets A. N., Postgraduate at the Department of Economics, Belarusian State University of Informatics and Radioelectronics

Address for correspondence

220013, Republic of Belarus,
Minsk, P. Brovki St., 6
Belarusian State University
of Informatics and Radioelectronics
Tel.: +375 17 293-80-46
E-mail: kozinets.science@gmail.com
Kazinets Aliaksandr Nikolaevich