**Министерство Образования и Исследований Молдовы**

**Технический Университет Молдовы**

**Факультет Вычислительной техники, Информатики и Микроэлектроники**

**Департамент Информатики и Системной Инженерии**

**Предмет:** Analiza Datelor

**Научная статья**

**Тема: «Анализ сотрудников компании и их возможный уход»**

**Выполнил:** Ясыбаш Степан, гр. MI-213

**Кишинёв 2023**

**Анализ сотрудников компании и их возможный уход**

**Iasîbaş Stepan**

*Universitatea Tehnică a Moldovei, Facultatea Calculatoare, Informatică și Microelectronică*  *Departamentul Informatică și Inginerie Sistemelor grupa MI-213, Chișinău, Moldova*

\* Autor corespondent: Iasîbaş Stepan, stepan.iasibas@iis.utm.md

*Абстракт*

*Данная научная статья посвящена исследованию факторов, влияющих на уход сотрудников из компании с использованием анализа данных. Проект представляет собой комплексный анализ информации о сотрудниках, включая их образование, опыт работы, место проживания, возраст, пол, а также другие характеристики. В ходе исследования была построена и настроена модель машинного обучения для предсказания вероятности ухода сотрудников.*

*Исследование включает в себя использование различных методов машинного обучения, включая логистическую регрессию, случайный лес и градиентный бустинг, с последующей настройкой гиперпараметров для повышения производительности моделей. Дополнительно, применяется балансировка классов и метод Synthetic Minority Over-sampling Technique (SMOTE) для учета дисбаланса классов в данных.*

*Результаты экспериментов представлены в виде метрик точности (accuracy), матрицы ошибок и отчета классификации для каждой модели. Полученные результаты предоставляют полезные инсайты для компании, позволяя выявить факторы, способствующие удержанию персонала, а также эффективные стратегии для предотвращения ухода сотрудников. Этот проект представляет собой важный шаг в области управления человеческими ресурсами и обеспечивает основу для принятия информированных решений по управлению персоналом.*

*Abstract*

*This scientific article explores the factors influencing employee attrition within a company using data analysis techniques. The project involves a comprehensive analysis of employee information, including education, work experience, location, age, gender, and other relevant characteristics. A machine learning model has been constructed and fine-tuned to predict the probability of employee turnover.*

*The research encompasses various machine learning methods, such as logistic regression, random forest, and gradient boosting, with subsequent hyperparameter tuning to enhance model performance. Additionally, class balancing techniques and the Synthetic Minority Over-sampling Technique (SMOTE) are applied to address class imbalance in the data.*

*Results of the experiments are presented through accuracy metrics, confusion matrices, and classification reports for each model. The obtained insights provide valuable information for the company, allowing the identification of factors contributing to employee retention and effective strategies to prevent turnover. This project represents a significant advancement in human resources management and serves as a foundation for making informed personnel management decisions.*

*Rezumat*

*Acest articol științific explorează factorii care influențează atritionarea angajaților în cadrul unei companii folosind tehnici de analiză a datelor. Proiectul implică o analiză cuprinzătoare a informațiilor despre angajați, inclusiv educație, experiență de muncă, locație, vârstă, gen și alte caracteristici relevante. A fost construit și ajustat un model de învățare automată pentru a prezice probabilitatea de rotire a angajaților.*

*Cercetarea cuprinde diverse metode de învățare automată, cum ar fi regresia logistică, pădurea aleatoare și creșterea gradientului, cu ajustarea ulterioară a hiperparametrilor pentru îmbunătățirea performanței modelului. În plus, sunt aplicate tehnici de echilibrare a claselor și Tehnica de Suprapunere Minoritară Sintetică (SMOTE) pentru a aborda dezechilibrul de clasă în date.*

*Rezultatele experimentelor sunt prezentate prin metrici de acuratețe, matrici de confuzie și rapoarte de clasificare pentru fiecare model. Insight-urile obținute furnizează informații valoroase companiei, permițând identificarea factorilor care contribuie la retenția angajaților și strategii eficiente pentru prevenirea fluctuației personalului. Acest proiect reprezintă un progres semnificativ în domeniul managementului resurselor umane și servește ca bază pentru luarea deciziilor informate privind gestionarea personalului.*

***Ключевые слова****: анализ, метод, прогнозирование, визуализация данных, предсказывающая модель*

**Введение**

В условиях современного бизнеса и устойчивой конкуренции, управление человеческими ресурсами становится важным стратегическим аспектом для обеспечения устойчивости и успешности компаний. Одним из ключевых аспектов в этом контексте является удержание квалифицированных и талантливых сотрудников. Несмотря на это, вопрос ухода сотрудников (оттока) остается актуальным вызовом для многих организаций.

**Целью** данного исследования является проведение анализа факторов, влияющих на возможный уход сотрудников из компании, с использованием методов машинного обучения и анализа данных. Понимание этих факторов позволит компаниям принимать более информированные решения в области управления человеческими ресурсами и разрабатывать эффективные стратегии удержания персонала.

В данном исследовании мы ставим перед собой задачу анализа факторов, влияющих на уход сотрудников из компании. Основываясь на предварительном анализе, мы выдвигаем два ключевых предположения, которые направлены на выявление возможных связей между характеристиками сотрудников и вероятностью их ухода.

Мы предполагаем, что уровень образования сотрудника может оказывать влияние на решение об уходе из компании. Наши ожидания основаны на предположении, что сотрудники с различными уровнями образования могут иметь различные ожидания от карьеры, и эти различия могут отразиться на вероятности ухода.

Наше второе предположение связано с возрастом сотрудников и его влиянием на решение об уходе. Мы предполагаем, что разные возрастные группы могут иметь различные потребности и мотивации, что влияет на стабильность и продолжительность их трудовой деятельности в компании.

Эти предположения станут основой для нашего дальнейшего исследования, включая сбор данных, построение моделей и анализ результатов. Проведение анализа по этим предположениям поможет нам лучше понять динамику ухода сотрудников и разработать целенаправленные стратегии управления персоналом для повышения устойчивости и продуктивности компании.

Проект включает в себя сбор и обработку данных о сотрудниках, создание прогностических моделей для предсказания вероятности ухода, и дальнейший анализ результатов. Мы также рассмотрим различные методы улучшения производительности моделей, такие как балансировка классов и использование методов увеличения данных.

В результате этого исследования ожидается получение глубокого понимания факторов, влияющих на уход сотрудников, что позволит компаниям принимать про активные меры для удержания своего персонала и, таким образом, повысить устойчивость и конкурентоспособность на рынке.

**Описание Данных**

Для достижения целей нашего исследования, мы использовали набор данных, содержащий информацию о сотрудниках компании. Всего в датасете 4653 строки, каждая из которых предоставляет информацию о различных характеристиках сотрудника. Вот краткое описание ключевых переменных, которые мы использовали в нашем анализе:

1. Education (Образование): Уровень образования сотрудника, представленный в виде категорий (например, Bachelors, Masters).
2. JoiningYear (Год присоединения): Год, в который сотрудник присоединился к компании.
3. City (Город): Местоположение сотрудника, где он проживает или работает.
4. PaymentTier (Уровень оплаты труда): Категория, отражающая уровень заработной платы сотрудника.
5. Age (Возраст): Возраст сотрудника.
6. Gender (Пол): Пол сотрудника.
7. EverBenched (Был ли в запасе): Индикатор, указывающий, был ли сотрудник когда-либо в запасе.
8. ExperienceInCurrentDomain (Опыт в текущей области): Количество лет опыта работы в текущей области.
9. LeaveOrNot (Уход или нет): Целевая переменная, указывающая, ушел ли сотрудник (1) или остался (0).

Эти переменные предоставляют широкий спектр информации о каждом сотруднике, позволяя нам провести комплексный анализ и выделить ключевые факторы, влияющие на уход сотрудников. В следующих этапах исследования мы приступим к предобработке данных, построению моделей машинного обучения и дальнейшему анализу результатов.

**Методы**

Для достижения конкретных результатов в моем проекте были использованы следующие методы:

* Логистическая регрессия — это разновидность [множественной регрессии](https://wiki.loginom.ru/articles/multiple-linear-regression.html), общее назначение которой состоит в анализе связи между несколькими [независимыми переменными](https://wiki.loginom.ru/articles/input-variable.html) (называемыми также регрессорами или предикторами) и [зависимой переменной](https://wiki.loginom.ru/articles/output-variable.html).[1]
* Алгоритм случайного леса (**Random Forest**) — универсальный алгоритм машинного обучения, суть которого состоит в использовании ансамбля решающих деревьев. Само по себе решающее дерево предоставляет крайне невысокое качество классификации, но из-за большого их количества результат значительно улучшается. Также это один из немногих алгоритмов, который можно использовать в абсолютном большинстве задач. [2]
* Градиентный бустинг — это техника машинного обучения для задач классификации и регрессии, которая строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений. [3]
* Площадь под ROC-кривой (AUC) –способ оценить эффективность модели прогнозирования. Вычисляется площадь под ROC-кривой (Receiver Operating Characteristic – операционная характеристика приемника). Площадь AUC связана с качеством прогноза (PP) в соответствии со следующей формулой: PP = 2 \* AUC - 1. Для модели прогнозирования простой рейтинговой оценки с бинарной целью эта площадь представляет собой вероятность, что случайно выбранное наблюдение сигнала будет иметь более высокую оценку, чем случайно выбранное отрицательное наблюдение (несигнал). [4]
* Критерий хи-квадрат – метод в математической статистике. Он показывает различия между фактическими данными в выборке и теоретическими результатами, которые предположил исследователь. С помощью метода оценивают, соответствует ли выборка законам распределения. Частный случай – критерий согласия Пирсона, который употребляется чаще всего. [5]
* Scatter Plot - График рассеяния представляет собой точки на плоскости, где одна ось отображает возраст, а другая - опыт работы в текущей области. Каждая точка на графике представляет одного сотрудника, и их распределение дает представление о том, как связаны эти два признака. Такой график позволяет выявить возможные тренды, кластеры или выбросы в данных. [6]

**Визуализация данных**

[Качественная визуализация данных](https://www.oracle.com/cis/business-analytics/data-visualization.html) имеет критическое значение для анализа данных и принятия решений на их основе. Визуализация позволяет быстро и легко замечать и интерпретировать связи и взаимоотношения, а также выявлять развивающиеся тенденции, которые не привлекли бы внимания в виде необработанных данных. В большинстве случаев для интерпретации графических представлений не требуется специальное обучение, что сокращает вероятность недопонимания.

Продуманное графическое представление не только содержит информацию, но и повышает эффективность ее восприятия за счет наглядности, привлечения внимания и удержания интереса в отличие от таблиц и документов. [7]

Для моих данных я использовал различные визуализации. Например, для визуализации связи между возрастом и опытом работы в текущей области я использовал Scatter plot (см. Приложение 1, Рисунок 1). График рассеяния представляет собой точки на плоскости, где одна ось отображает возраст, а другая - опыт работы в текущей области. Каждая точка на графике представляет одного сотрудника, и их распределение дает представление о том, как связаны эти два признака.

Также я использовал столбчатую диаграмму[8] для визуализации связи между Образованием и Возможным уходом сотрудника из компании (см. Приложение 1, Рисунок 2). Также мы вычислили Chi2 Value (Значение хи-квадрат). Это статистика, которая измеряет разницу между фактическим и ожидаемым распределением данных. В данном случае, значение хи-квадрат равно 101.8301. Чем больше это значение, тем сильнее отклонение от ожидаемого, что может свидетельствовать о наличии статистически значимой связи между переменными (образованием и решением об уходе). P-value (Уровень значимости): Это вероятность получить такое или более экстремальное значение статистики хи-квадрат при условии, что нулевая гипотеза верна. В данном случае, p-value равно 7.7244e-23, что практически равно нулю. Малое значение p-value говорит о том, что мы имеем достаточно доказательства для отвержения нулевой гипотезы (гипотезы о независимости переменных). Таким образом, существует статистически значимая связь между уровнем образования и решением сотрудников об уходе. В данном контексте, результаты статистического теста указывают на то, что уровень образования сотрудников имеет статистически значимое влияние на их решение об уходе из компании. Вероятность получить такие результаты случайно крайне мала, что подтверждает статистическую значимость связи.

Далее я применил гистограмму для выявления связи между Возрастом и Возможным уходом сотрудника (см. Приложение 1, Рисунок 3). Также вычислили Chi2 Value (Значение хи-квадрат). В данном случае, значение хи-квадрат равно 5.172, что является относительно невысоким значением.

P-value (Уровень значимости) в данном случае равно 0.0753, что является относительно высоким значением. Учитывая высокое значение p-value, мы не отвергаем нулевую гипотезу. Это означает, что нет статистически значимой связи между возрастом сотрудников и решением об уходе. Таким образом, уровень статистической значимости не достигнут, и мы не можем сделать вывод о том, что возраст сотрудников влияет на решение об уходе из компании.

**Важность признаков**

Оценка важности признаков в моделях машинного обучения играет ключевую роль в понимании влияния каждой переменной на результаты прогнозирования. Важность признаков отражает степень влияния каждого фактора на целевую переменную и может быть использована для принятия важных бизнес - решений. Это полезное понимание может помочь в определении ключевых факторов, влияющих на исход задачи, а также в выявлении наиболее значимых трендов или закономерностей.

Важность признаков может быть использована для:

* Управления ресурсами: При наличии информации о важности различных факторов бизнес может выделить ресурсы наиболее эффективным образом, сфокусировав усилия на ключевых областях.
* Оптимизации стратегий: Понимание влияния различных переменных позволяет оптимизировать стратегии и тактики в зависимости от бизнес-целей.
* Выявления факторов риска: Идентификация наиболее важных признаков помогает выявить потенциальные риски и угрозы, что полезно для принятия проактивных мер.
* Повышения интерпретируемости: Важность признаков обеспечивает лучшее понимание внутреннего механизма модели, что может быть важно для интерпретации результатов и принятия доверительных решений.

Таким образом, оценка важности признаков не только улучшает качество прогнозов моделей, но также предоставляет ценную информацию для более эффективного управления и принятия стратегических решений в различных областях бизнеса.

Важность признаков, рассчитанная моделями случайного леса и градиентного бустинга, предоставляет ценную информацию о влиянии различных факторов на прогнозирование ухода сотрудников.

Feature Importance:

* JoiningYear 0.362420
* City\_Pune 0.201405
* PaymentTier 0.179759
* Education\_Masters 0.099129
* Gender\_Male 0.052566
* Gender\_Female 0.027650
* Education\_Bachelors 0.020525
* Age 0.020383
* City\_Bangalore 0.009906
* ExperienceInCurrentDomain 0.009571
* City\_New Delhi 0.006551
* EverBenched\_Yes 0.005396
* Education\_PHD 0.002779
* EverBenched\_No 0.001959

В нашем исследовании признаки, такие как "JoiningYear", "City\_Pune" и "PaymentTier", оказались наиболее значимыми. Также я вывел результаты в виде диаграммы (см. Приложение 1, Рисунок 4). Эти результаты могут служить основой для разработки стратегий удержания персонала и улучшения управления кадрами в организации, учитывая особую значимость указанных факторов.

**Визуализация модели**

**Logistic Regression**

Далее мы переходим, непосредственно, к визуализации нашей модели. В первую очередь, я использую метод логистической регрессии для моей модели. Мы получаем следующие результаты:

1. Accuracy (Точность): Это доля правильных предсказаний относительно общего числа предсказаний. Здесь значение accuracy равно 0.7143, что означает, что примерно 71.43% предсказаний модели были верны.
2. Confusion Matrix (Матрица ошибок): Эта матрица показывает количество верных и неверных предсказаний для каждого класса. В данном случае:

* Верные отрицательные (True Negatives, TN): 551
* Ложные положительные (False Positives, FP): 59
* Ложные отрицательные (False Negatives, FN): 207
* Верные положительные (True Positives, TP): 114

1. Classification Report (Отчет о классификации): Этот отчет предоставляет различные метрики для каждого класса (0 и 1) и их взвешенные средние. В данном случае:

* Precision (Точность): Показывает, как много из предсказанных положительных примеров действительно являются положительными. Для класса 0 это 0.73, а для класса 1 - 0.66.
* Recall (Полнота): Показывает, как много из фактически положительных примеров было успешно предсказано. Для класса 0 это 0.90, а для класса 1 - 0.36.
* F1-Score (F1-мера): Среднее гармоническое между точностью и полнотой. Для класса 0 это 0.81, а для класса 1 - 0.46.

**Random Forest**

Далее для определения лучшей модели, я применяю метод случайного леса

(Random Forest) и получаю следующие результаты:

1. Random Forest Accuracy (Точность модели): Доля правильных предсказаний относительно общего числа предсказаний. В данном случае, точность модели Random Forest составляет примерно 84.96%, что означает, что примерно 85% предсказаний были верны.
2. Random Forest Confusion Matrix (Матрица ошибок Random Forest): Эта матрица показывает количество верных и неверных предсказаний для каждого класса. В данном случае:

* Верные отрицательные (True Negatives, TN): 559
* Ложные положительные (False Positives, FP): 51
* Ложные отрицательные (False Negatives, FN): 89
* Верные положительные (True Positives, TP): 232

1. Random Forest Classification Report (Отчет о классификации Random Forest): Этот отчет предоставляет различные метрики для каждого класса и их взвешенные средние. Для этого случая:

* Precision (Точность): Доля верно предсказанных положительных примеров. Для класса 0 - 86%, для класса 1 - 82%.
* Recall (Полнота): Доля фактически положительных примеров, которые были успешно предсказаны. Для класса 0 - 92%, для класса 1 - 72%.
* F1-Score (F1-мера): Среднее гармоническое между точностью и полнотой. Для класса 0 - 89%, для класса 1 - 77%.

Эти метрики позволяют оценить, насколько хорошо модель справляется с задачей классификации и идентификации ухода сотрудников.

**Gradient Boosting**

Также для определения лучшей модели я применяю третий метод, который называется градиентный бустинг (Gradient Boosting). В результате получил следующее:

1. Gradient Boosting Accuracy (Точность модели): Доля правильных предсказаний относительно общего числа предсказаний. Здесь точность модели Gradient Boosting составляет примерно 86.04%, что означает, что примерно 86% предсказаний были верны.
2. Gradient Boosting Confusion Matrix (Матрица ошибок Gradient Boosting): Эта матрица показывает количество верных и неверных предсказаний для каждого класса. В данном случае:

* Верные отрицательные (True Negatives, TN): 581
* Ложные положительные (False Positives, FP): 29
* Ложные отрицательные (False Negatives, FN): 101
* Верные положительные (True Positives, TP): 220

1. Gradient Boosting Classification Report (Отчет о классификации Gradient Boosting): Этот отчет предоставляет различные метрики для каждого класса и их взвешенные средние. Получаем в нашем случае:

* Precision (Точность): Доля верно предсказанных положительных примеров. Для класса 0 - 85%, для класса 1 - 88%.
* Recall (Полнота): Доля фактически положительных примеров, которые были успешно предсказаны. Для класса 0 - 95%, для класса 1 - 69%.
* F1-Score (F1-мера): Среднее гармоническое между точностью и полнотой. Для класса 0 - 90%, для класса 1 - 77%.

Эти метрики обобщают производительность модели Gradient Boosting, помогая оценить ее способность правильно предсказывать уход сотрудников и общую эффективность в сравнении с другими моделями.

**Результаты**

Применив три метода для определения оптимальной модели, я решил сравнить результаты. Я применил столбчатую диаграмму (см. Приложение 1, Рисунок 5), в котором четко видно, какой метод дал лучшие результаты. В итоге мы получаем:

Логистическая регрессия – 71%;

Случайный лес – 85%;

Градиентный бустинг – 86%.

Отсюда мы делаем выводы, что метод Градиентный бустинг выдал нам лучший результат(86%), что означает, что наша модель успешно классифицировала правильно примерно 86% случаев в нашем наборе данных.

Также для сравнения результатов точности модели, я решил использовать балансировку классов и SMOTE (Synthetic Minority Over-sampling Technique), которые являются методами для работы с дисбалансом классов в задачах машинного обучения, где количество примеров в различных классах существенно различается.

Балансировка классов: В случае, когда один класс представлен значительно большим количеством примеров по сравнению с другим, модель может стать смещенной и предсказывать этот "больший" класс с высокой точностью, но не уделять достаточного внимания "меньшему" классу. Балансировка классов позволяет уровнять распределение, чтобы обе категории имели близкое количество примеров.

SMOTE (Synthetic Minority Over-sampling Technique): Этот метод направлен на устранение дисбаланса путем создания синтетических примеров для класса, который представлен в меньшем количестве. Вместо простого дублирования существующих примеров, SMOTE генерирует новые, исходя из анализа ближайших соседей. Это помогает улучшить представление меньшего класса в обучающем наборе данных. Эти методы полезны в ситуациях, когда дисбаланс классов может привести к нежелательным результатам, и они способствуют обучению модели, которая более уравновешенно учитывает оба класса.

В итоге, после применения данных метрик, я получил следующие результаты и вывел их для сравнения в виде столбчатой диаграмме (см. Приложение1, Рисунок 6):

Логистическая регрессия – 67%;

Случайный лес – 86%;

Градиентный бустинг – 85%.

Исходя из результатов, мы делаем вывод, что модели случайного леса и градиентного бустинга показывают более высокую точность по сравнению с логистической регрессией.

Ну, и в конце, для подведения итогов, я решил сделать столбчатую диаграмму (см. Приложение 1, Рисунок 7), в которой я вывел все результаты данных методов, которые я использовал. В итоге, я получил такие выводы, что точность логистической регрессии уменьшилась после применения балансировки классов и SMOTE. Это может быть вызвано тем, что модель столкнулась с более сложными данными после уравновешивания классов. Точность случайного леса осталась неизменной на уровне 86%, что может свидетельствовать о его устойчивости к дисбалансу классов. Это подтверждает его эффективность в работе с разнородными данными. Точность градиентного бустинга также осталась стабильной на уровне 85%, что подчеркивает его устойчивость к дисбалансу классов и способность обобщать данные.

Также я решил сделать график AUC-ROC(см. Приложение1, Рисунок 8). В нашем случае "ROC curve (area=0.89)" означает, что площадь под ROC-кривой для нашей модели составляет 0.89. Это является положительным показателем, так как AUC ближе к 1 указывает на более точный и эффективный классификатор. Высокое значение AUC свидетельствует о том, что модель хорошо различает между классами и обладает высокой способностью распознавать положительные и отрицательные случаи.

**Выводы**

В рамках исследования был проведен анализ различных факторов, влияющих на решение сотрудников об уходе из компании. Произведено построение нескольких моделей машинного обучения, включая логистическую регрессию, случайный лес и градиентный бустинг. Каждая из моделей была оценена на основе метрик производительности, таких как точность (accuracy), что позволяет выбрать наиболее подходящую модель для предсказания ухода сотрудников.

Была проведена оценка важности признаков, что позволило выделить те факторы, которые оказывают наибольшее влияние на прогноз ухода сотрудников. Для улучшения обучения моделей на несбалансированных данных были применены методы балансировки классов и синтетического увеличения данных (SMOTE). Это позволило улучшить производительность моделей и сделать их более устойчивыми к классам с меньшим количеством данных. Были проведены статистические тесты для проверки предположений о влиянии образования и возраста на уход сотрудников. Общий результат проделанной работы — разработка инструмента для предсказания ухода сотрудников с использованием машинного обучения и выделение важных факторов, оказывающих влияние на это решение.

**Приложение 1**

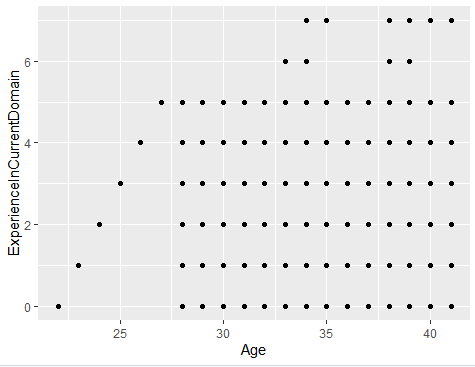


Рисунок 1. Визуализация связи между возрастом и опытом работы в текущей области.

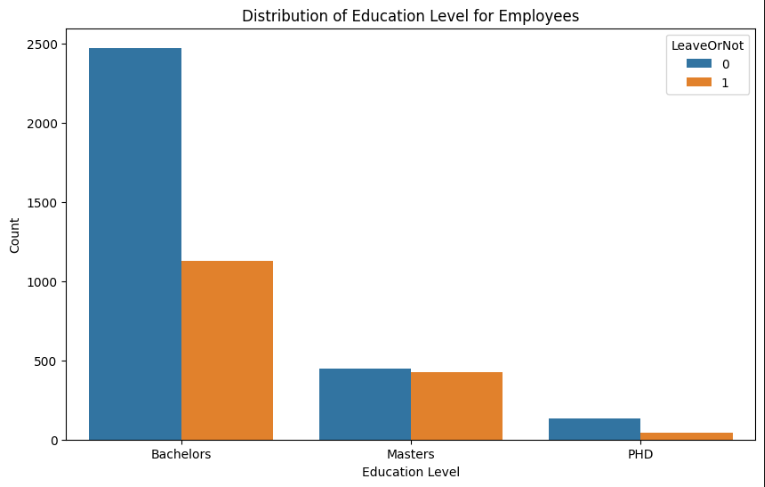


Рисунок 2. Связь между образованием и уходом сотрудника из компании.

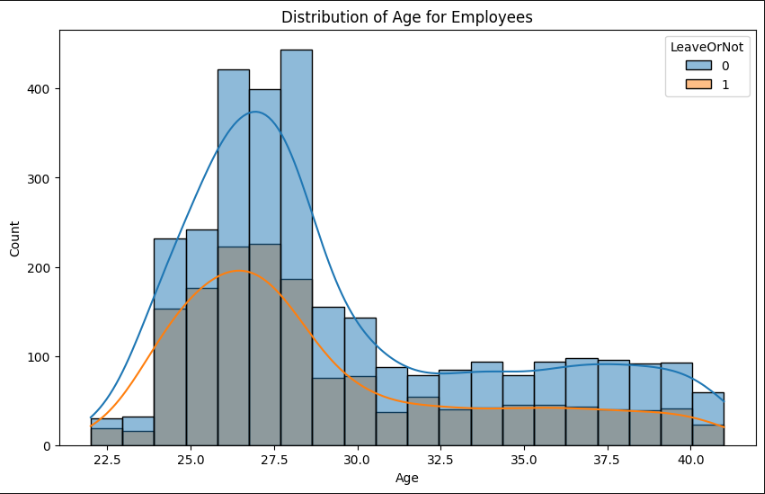


Рисунок 3. Связь между возрастом и уходом сотрудника из компании.

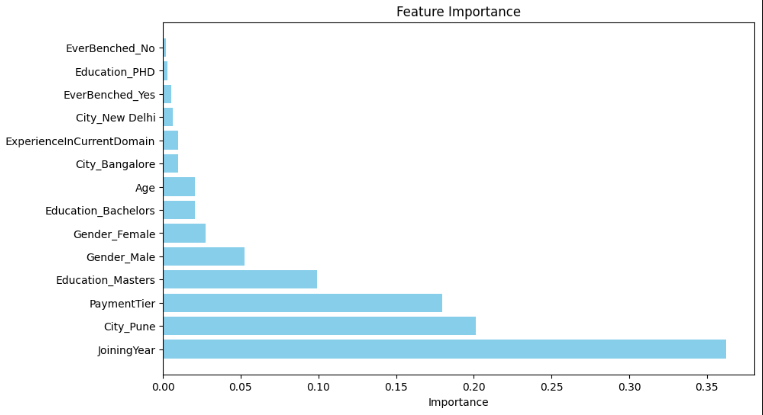


Рисунок 4. Важность переменных.

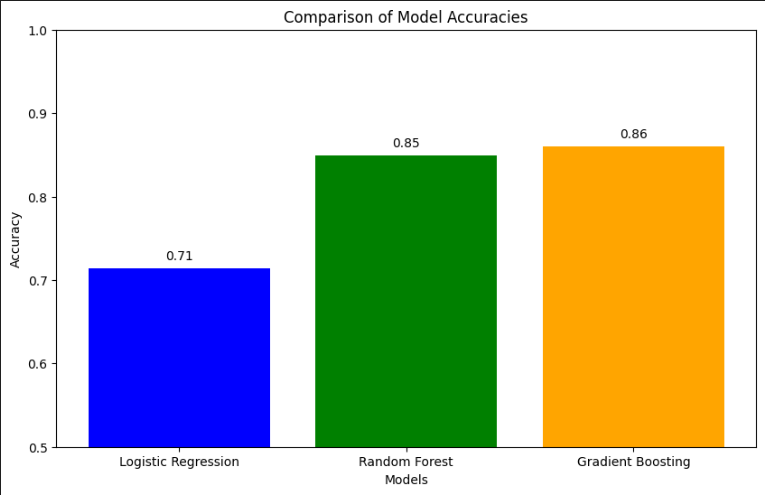


Рисунок 5. Диаграмма сравнения результатов.

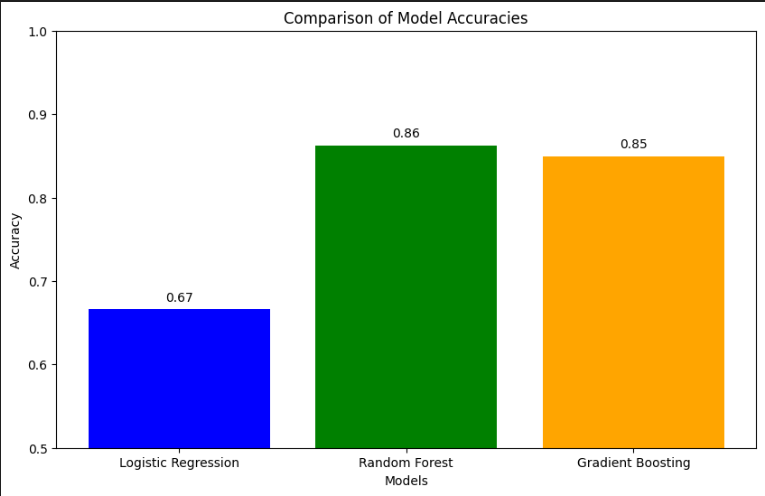


Рисунок 6. Диаграмма сравнения результатов после балансировки и SMOTE.

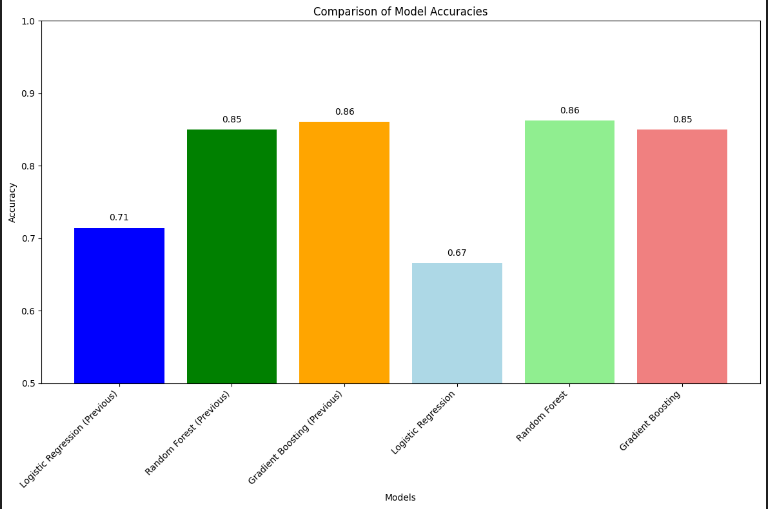


Рисунок 7. Диаграмма сравнения всех результатов.

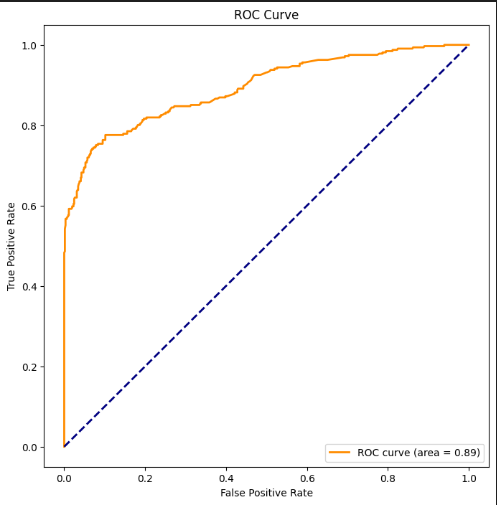


Рисунок 8.График AUC-ROC.

**Библиография**

1. Loginom © 2023, Логистическая регрессия и ROC-анализ — математический аппарат, прочитано [18.12.2023]. Взято из <https://loginom.ru/blog/logistic-regression-roc-auc#:~:text>
2. Градиентый бустинг — просто о сложном, прочитано[18.12.2023]. Взято из <https://neurohive.io/ru/osnovy-data-science/gradientyj-busting/#pll_switcher>
3. © 2023, Proglib Машинное обучение для начинающих: алгоритм случайного леса (Random Forest), прочитано [18.12.2023]. Взято из <https://proglib.io/p/mashinnoe-obuchenie-dlya-nachinayushchih-algoritm-sluchaynogo-lesa-random-forest-2021-08-12>
4. SAP Help Portal, Площадь под ROC-кривой (AUC), прочитано [18.12.2023]. Взято из<https://help.sap.com/docs/SAP_ANALYTICS_CLOUD/00f68c2e08b941f081002fd3691d86a7/235c79933a7b4f398369e23a04520a3e.html>
5. © 2023 Skillfactory media, Критерий хи-квадрат, прочитано [18.12.2023]. Взято из <https://blog.skillfactory.ru/glossary/kriteriy-hi-kvadrat/>
6. © 2008–2023 Unisender, Виды диаграмм: 5 популярных способов визуализации данных, прочитано [18.12.2023]. Взято из <https://www.unisender.com/ru/blog/idei/vidy-diagramm-sposoby-vizualizacii-dannyh>
7. © 2023 Oracle, Что такое визуализация данных?, прочитано [18.12.2023]. Взято из <https://www.oracle.com/cis/business-analytics/what-is-data-visualization>
8. Severino Ribecca, Столбиковая диаграмма, прочитано [18.12.2023]. Взято из <https://datavizcatalogue.com/RU/metody/stolbikovaja_diagramma.html>