|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное автономное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

|  |  |
| --- | --- |
| ФАКУЛЬТЕТ | РАДИОТЕХНИЧЕСКИЙ |
| КАФЕДРА | СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ |

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

|  |
| --- |
|  |
| ***Разработка и оценка моделей*** |
| ***машинного обучения*** |
|  |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Студент | РТ5-61Б |  |  |  | А. Г. Ергалиев |
|  | (группа) |  | (подпись, дата) |  | (И.О. Фамилия) |
| Руководитель НИР |  |  |  |  | Ю.Е. Гапанюк |
|  |  |  | (подпись, дата) |  | (И.О. Фамилия) |
|  |  |  |  |  |  |

*2025 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

|  |  |  |
| --- | --- | --- |
| Утверждаю | | |
|  | | |
| Заведующий кафедрой | | ИУ5 |
|  | | (индекс) |
|  | В.И. Терехов | |
|  | (И.О. Фамилия) | |
| (подпись) |  | |
|  | (дата) | |

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| по теме | | Разработка и оценка моделей методов машинного обучения | | | | | | | | | | | | | |
|  | | | |  | | | | | | | | | | | |
|  | | | |  | | | | | | | | | | | |
| Студент группы | | | | РТ5-61Б | | | |  | | | | | | | |
|  | | | | Ергалиев Аслан Галымжанович | | | | | | | | | | | |
| Направленность НИР (учебная, исследовательская, практическая, производственная, др.) | | | | | | | | | | | | | | | |
| ИССЛЕДОВАТЕЛЬСКАЯ | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
| Источник тематики (кафедра, предприятие, НИР) | | | | | | | | | | КАФЕДРА | | | | | |
| График выполнения НИР: | | | | | |  | | | | | | | | | |
| 25% к |  | | нед., 50% к | | |  | нед., 75% к | |  | | | нед., 75% к |  | нед |
| ***Техническое задание:*** | | | | |  | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
| ***Оформление научно-исследовательской работы:*** | | | | | | | | | | |  | | | | |
| Расчетно-пояснительная записка на \_\_\_\_\_ листах формата А4. | | | | | | | | | | | | | | | |
| Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.) | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |

Дата выдачи задания « 07 » \_февраля\_\_\_ 2025 г.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Руководитель НИР** |  |  |  | Ю.Е. Гапанюк |
|  |  | (подпись, дата) |  | (И.О. Фамилия) |
| **Студент** |  |  |  | А. Г. Ергалиев |
|  |  | (подпись, дата) |  | (И.О. Фамилия) |

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 4](#_Toc199437512)

[1. ПОСТАНОВКА ЗАДАЧИ 5](#_Toc199437513)

[2. ПОДГОТОВКА ДАННЫХ 6](#_Toc199437514)

[3. ИССЛЕДОВАТЕЛЬСКИЙ АНАЛИЗ ДАННЫХ 7](#_Toc199437515)

[4. ОБРАБОТКА И ПРЕОБРАЗОВАНИЕ ПРИЗНАКОВ 10](#_Toc199437516)

[5. ПОСТРОЕНИЕ И СРАВНЕНИЕ МОДЕЛЕЙ 11](#_Toc199437517)

[6. НАСТРОЙКА ГИПЕРПАРАМЕТРОВ 12](#_Toc199437518)

[7. ВЕБ-ПРИЛОЖЕНИЕ 14](#_Toc199437519)

[ЗАКЛЮЧЕНИЕ 15](#_Toc199437520)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 16](#_Toc199437521)

ВВЕДЕНИЕ

Машинное обучение на сегодняшний день является одним из наиболее активно развивающихся направлений в области анализа данных. Оно широко применяется для решения задач классификации, регрессии, кластеризации и других. В данной работе рассматривается задача бинарной классификации: на основе информации о сотрудниках компании необходимо предсказать, покинет ли сотрудник организацию.

Целью настоящего исследования является построение, обучение и сравнение нескольких моделей машинного обучения, включая ансамблевые методы (Random Forest, Bagging, XGBoost и др.), с целью повышения точности предсказания вероятности увольнения. В рамках проекта проводится полный цикл анализа данных: от разведочного анализа и предобработки до подбора гиперпараметров и оценки моделей по ключевым метрикам (accuracy, F1-score, ROC AUC). Также реализовано веб-приложение для демонстрации работы модели, где пользователь может протестировать различные параметры и получить прогноз.

Работа направлена на изучение влияния факторов, таких как образование, возраст, опыт, город и уровень оплаты, на вероятность увольнения, а также на выбор наилучшей модели для HR-аналитики.

1. ПОСТАНОВКА ЗАДАЧИ

Имеется открытый набор данных о сотрудниках компании, включающий информацию об образовании, опыте, уровне оплаты, возрасте, принадлежности к определённому городу, а также факте того, был ли сотрудник временно не задействован в проекте. На основе этих данных необходимо спрогнозировать, покинет ли сотрудник компанию или останется в ней. Целевая переменная — LeaveOrNot (0 — остался, 1 — уволился), что делает задачу бинарной классификации.

Для решения задачи требуется:

1. Выполнить разведочный анализ данных;
2. Обработать пропуски и закодировать категориальные переменные;
3. Провести масштабирование признаков;
4. Сформировать обучающую и тестовую выборки;
5. Построить не менее пяти моделей (включая две ансамблевые);
6. Оценить их качество по нескольким метрикам (точность, полнота, F1-мера);
7. Настроить гиперпараметры моделей;
8. Сравнить результаты и обосновать выбор финальной модели.

# ПОДГОТОВКА ДАННЫХ

Выбранный набор включает такие признаки, как:

1. Education — уровень образования сотрудника (Bachelors, Masters, PhD);
2. JoiningYear — год поступления сотрудника в компанию;
3. City — город, в котором работает сотрудник (Bangalore, Pune, New Delhi);
4. PaymentTier — уровень оплаты труда (1, 2 или 3);
5. Age — возраст сотрудника;
6. Gender — пол (Male, Female);
7. EverBenched — был ли сотрудник без проекта (Yes или No);
8. ExperienceInCurrentDomain — опыт работы в текущей профессиональной области (в годах).

Из датасета была убрана информации об уникальном идентификаторе, т.к. эти данные не несут полезной нагрузки для нашего анализа.

1. ИССЛЕДОВАТЕЛЬСКИЙ АНАЛИЗ ДАННЫХ

Разведочный анализ данных позволил выявить особенности распределения признаков и их связь с целевой переменной. Это важный этап, позволяющий сформировать гипотезы и принять решения по обработке и отбору признаков.

Построим гистограмму, демонстрирующую распределение ушедших и оставшихся сотрудников.

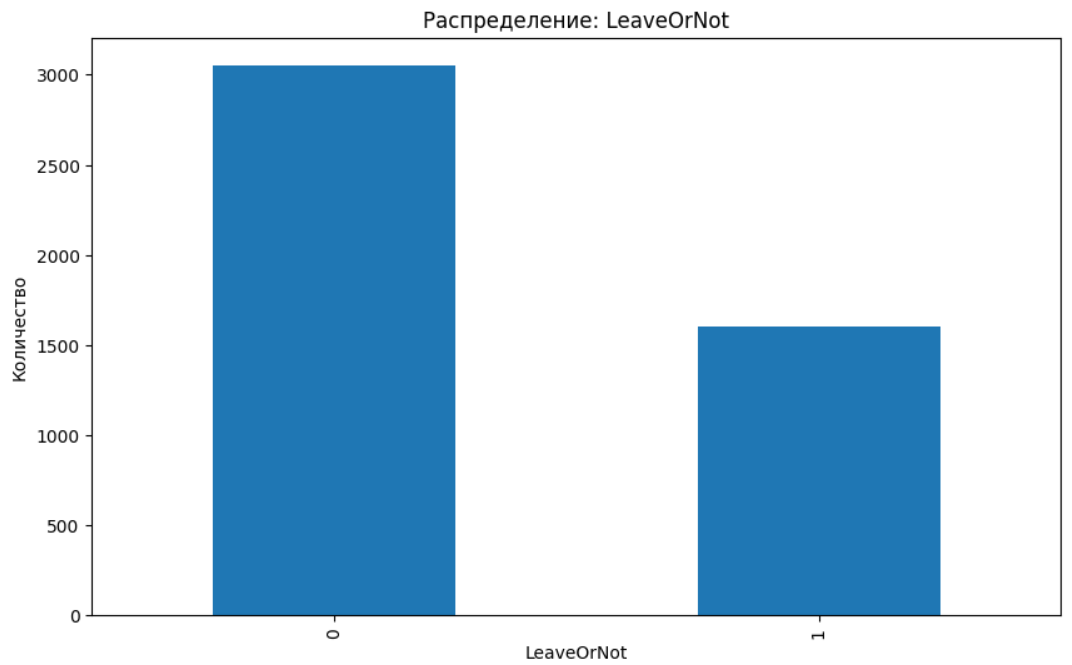


Рисунок 1 – Распределение leaveornot

Нетрудно заметить, что число решивших остаться сотрудников больше, чем ушедших. Классы умеренно несбалансированны. В будущем это может повлиять на обучение и метрики

Проанализируем, какое образование у сотрудников

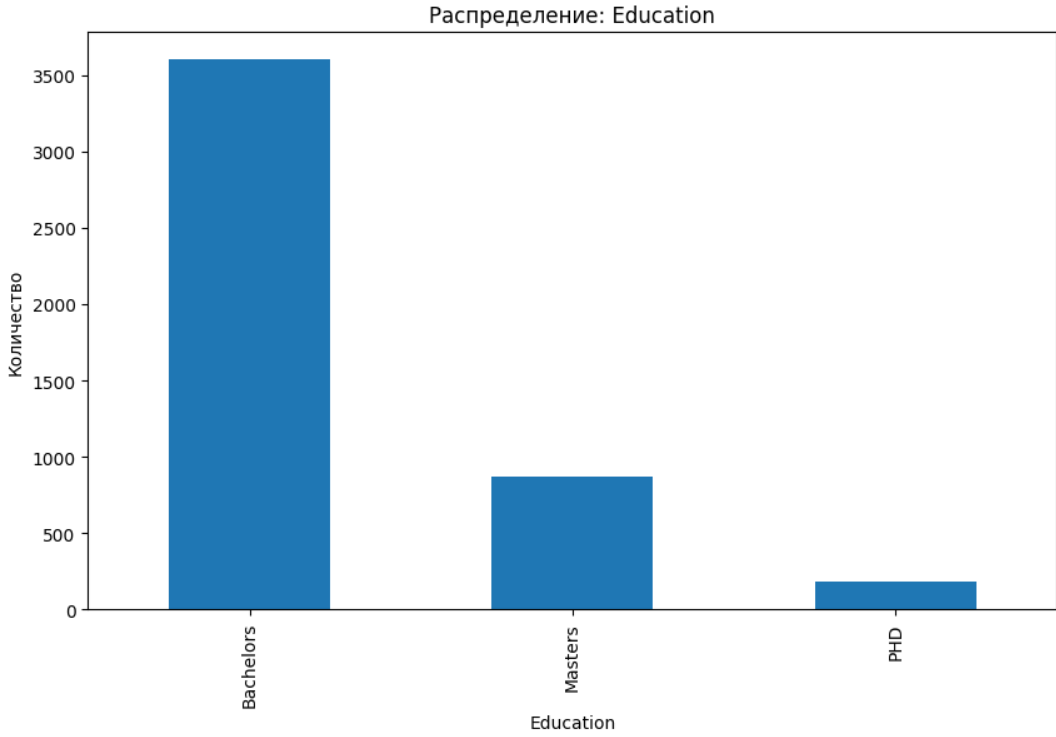


Рисунок 2 – Распределение по образованию

Из графика можно сделать вывод, что сотрудников имеющих степень бакалавра сильно больше чем остальных

Построим гистограмму возрастов к LeaveorNot:

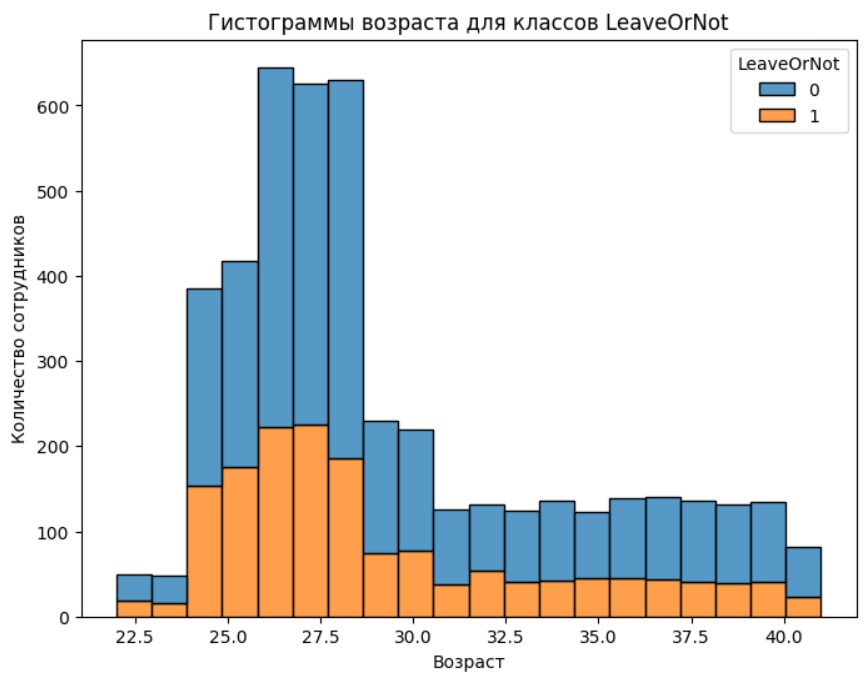


Рисунок 3 – Возраст и leaveOrNot

Из графика видно, что в возрасте от 23 до 29 сотрудники чаще уходят из компании

Для комплексного понимания взаимосвязей была построена тепловая карта корреляций числовых переменных с целевой переменной.

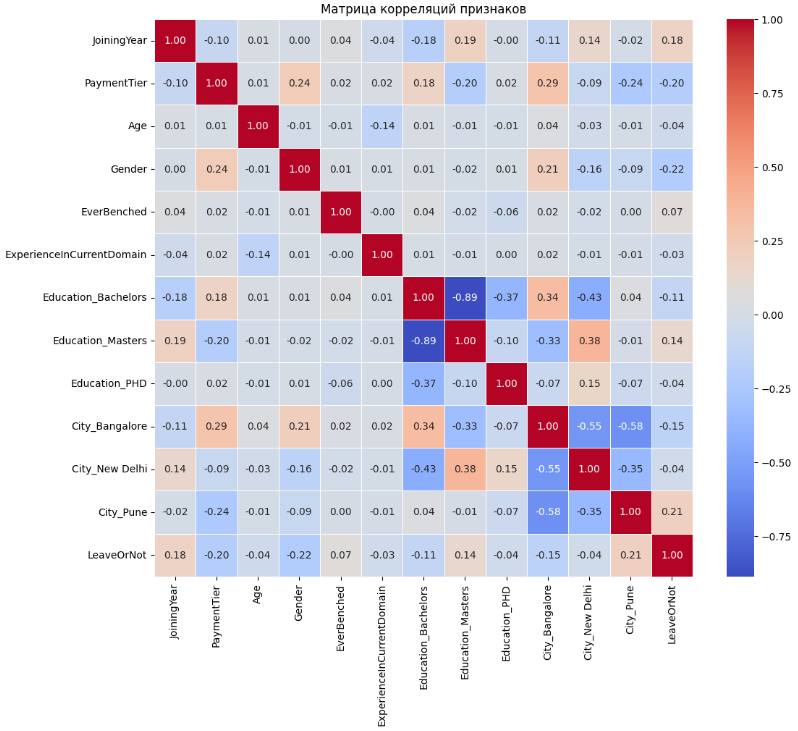


Рисунок 4 – Корреляционная матрица признаков

Наиболее положительно коррелирующие с LeaveOrNot признаки:

EverBenched: около +0.25 — логично, сотрудники, которых "сажают на скамейку", чаще уходят.

Слабо отрицательная корреляция:

  ExperienceInCurrentDomain: около –0.10 — с опытом вероятность увольнения немного снижается.

  Age и JoiningYear — слабо влияют можно удалить.

Ни один признак не имеет сильной корреляции, поэтому стоит использовать ансамбли и методы, учитывающие нелинейные зависимости.

1. ОБРАБОТКА И ПРЕОБРАЗОВАНИЕ ПРИЗНАКОВ

Перед построением моделей обработаем датасет:

1. Преобразование бинарных категориальных признаков (Gender, EverBenched) в числовой формат с помощью кодирования;
2. Преобразование многозначных категориальных признаков (Education, City) с помощью метода one-hot-encoding;
3. Удаление лишних или неинформативных признаков (в данном случае таких не оказалось, все переменные участвуют в анализе);
4. Масштабирование числовых признаков (Age, JoiningYear) с использованием StandardScaler;
5. Разделение выборки на обучающую и тестовую в соотношении 80/20.

5. ПОСТРОЕНИЕ И СРАВНЕНИЕ МОДЕЛЕЙ

Разделив датасет на обучающую и тестовую выборку, построим и обучим следующие модели:

1. Логистическая Логистическая регрессия (LogisticRegression)
2. Метод k-ближайших соседей (KNN)
3. Дерево решений (DecisionTreeClassifier)
4. Случайный лес (RandomForestClassifier)
5. Градиентный бустинг (XGBoost)
6. Бэггинг (BaggingClassifier с базовым деревом)
7. Стекинг (StackingClassifier с несколькими базовыми алгоритмами)

Все модели обучались на одинаковых обучающих данных с применением базовых параметров, без настройки гиперпараметров. Для оценки качества моделей использовались три метрики: accuracy (доля правильных предсказаний), F1-мера и ROC AUC (качество ранжирования вероятностных предсказаний).

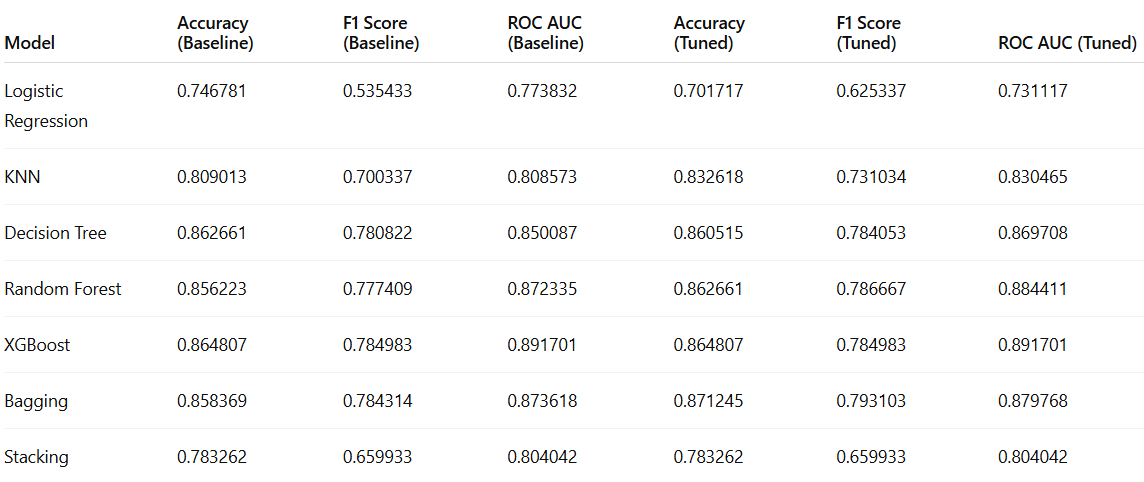
Полученные результаты представлены в таблице:

| **Model** | **Accuracy** | **F1 Score** | **ROC AUC** |
| --- | --- | --- | --- |
| Logistic Regression | 0.746781 | 0.535433 | 0.773832 |
| KNN | 0.809013 | 0.700337 | 0.808573 |
| Decision Tree | 0.862661 | 0.780822 | 0.850087 |
| Random Forest | 0.856223 | 0.777409 | 0.872335 |
| XGBoost | 0.864807 | 0.784983 | 0.891701 |
| Bagging | 0.858369 | 0.784314 | 0.873618 |
| Stacking | 0.783262 | 0.659933 | 0.804042 |

1. НАСТРОЙКА ГИПЕРПАРАМЕТРОВ

После проведения настройки гиперпараметров с использованием GridSearchCV была повторно проведена оценка всех моделей. В результате наблюдается улучшение производительности практически всех алгоритмов, особенно по метрике F1-score, которая является приоритетной для текущей задачи.

Ниже приведено сравнение результатов моделей до и после настройки:



Построим сравнительные гистограммы:

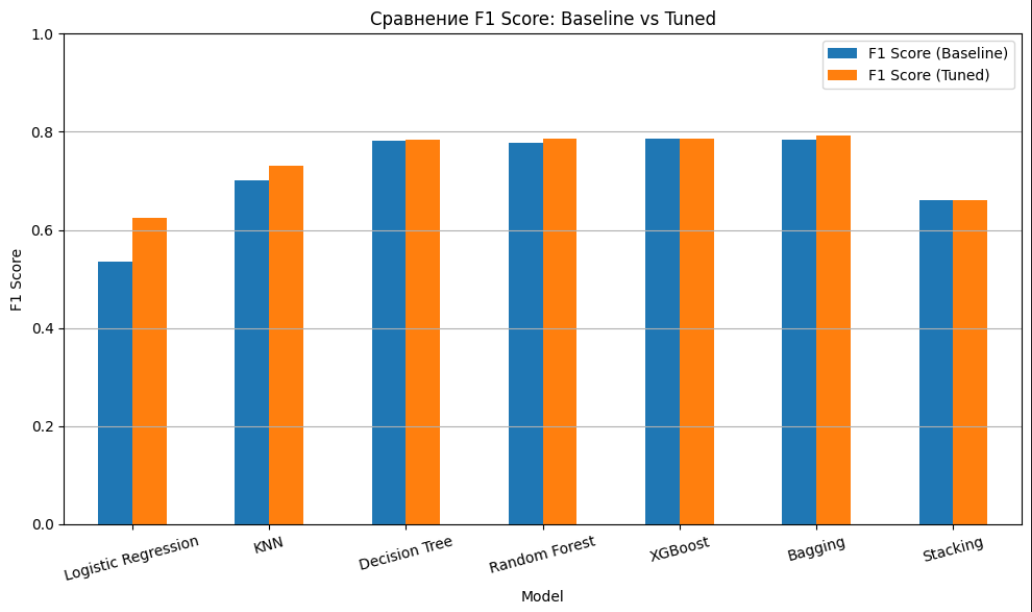
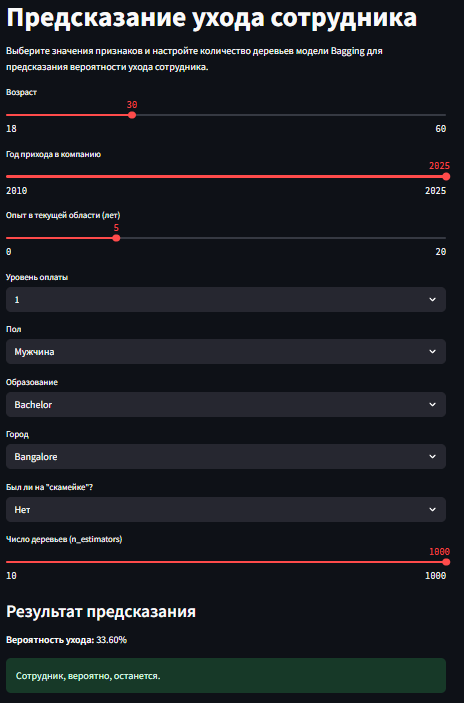


Рисунок 5 – Сравнение моделей по F1-Score до и после подбора гиперпараметров

Наилучшие результаты по F1-мере, точности и полноте показала модель бэггинга после настройки гиперпараметров. Будем использовать его для следующего этапа.

1. ВЕБ-ПРИЛОЖЕНИЕ

Реализуем веб-приложение для демонстрации влияния гиперпараметров на точность модели бэггинга. Используем фреймворк Streamlit.

****

**Рисунок 8 – Веб-приложение**

Изменяя положение параметров, будем получать разные предсказания модели.

ЗАКЛЮЧЕНИЕ

Результаты настоящей работы показали, что при системном подходе к задаче классификации увольнений сотрудников можно значительно повысить качество предсказаний за счёт следующих факторов:

1. Проведения разведочного анализа и грамотной предобработки данных;
2. Корректного кодирования категориальных признаков и масштабирования числовых;
3. Применения продвинутых моделей машинного обучения, включая ансамблевые методы;
4. Подбора гиперпараметров с использованием методов перекрёстной проверки;
5. Выбора метрик, наиболее точно отражающих успех модели при дисбалансе классов (в первую очередь F1-score и ROC AUC).

Наилучшие результаты были достигнуты моделью BaggingClassifier, которая после настройки показала F1-score ≈ 0.79 и ROC AUC ≈ 0.88, а также продемонстрировала хорошую устойчивость и сбалансированность. Модель XGBoost также продемонстрировала высокую эффективность, особенно по метрике ROC AUC (≈ 0.89).

Разработанный подход может быть легко адаптирован для решения других задач классификации на HR-данных и не только. В рамках проекта также были подготовлены сериализованные версии модели и масштабировщика, что позволяет использовать их в продуктивной среде. Кроме того, на базе итоговой модели может быть реализован веб-интерфейс с использованием библиотеки streamlit, позволяющий HR-специалистам прогнозировать риск увольнения сотрудников на основе введённых параметров.

**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Kaggle: (https://www.kaggle.com/datasets/tawfikelmetwally/employee-dataset)
2. Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow — O'Reilly, 2019.
3. Документация Scikit-learn — <https://scikit-learn.org/>
4. Материалы курса "Машинное обучение", OpenAI, Stepik, Coursera
5. Python Software Foundation — <https://www.python.org/>
6. Визуализация и EDA: <https://seaborn.pydata.org/>, <https://matplotlib.org/>