|  |  |
| --- | --- |
|  | **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 |  |  |  | В.Е. Пичурин |
|  | (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 | | | |  | | | | | | | |
|  | | | | Пичурин Вадим Евгеньевич | | | | | | | | | | | |
| 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** |  |  |  | В.Е. Пичурин |
|  |  | (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 сегодня является одним из наиболее динамично развивающихся направлений в сфере анализа данных. Его методы широко применяются для решения разнообразных задач — от классификации и регрессии до кластеризации и прогностического моделирования.

В рамках данной работы рассматривается задача классификации: на основе данных о человеке необходимо предсказать, зарабатывает ли он больше 50 000$.

Целью исследования является разработка, обучение и сравнительный анализ нескольких моделей машинного обучения, включая ансамблевые алгоритмы. В ходе работы выполняется полный цикл подготовки и обработки данных, проводится оценка качества моделей с использованием различных метрик, осуществляется подбор гиперпараметров, формируются выводы о качестве построенных решений.

Дополнительно, для демонстрации результатов создаётся интерактивное веб-приложение, позволяющее пользователю взаимодействовать с моделью в реальном времени.

1. ПОСТАНОВКА ЗАДАЧИ

В рамках данной работы рассматривается задача бинарной классификации на основе открытого датасета Adult Income (Census Income Dataset), содержащего социодемографическую информацию о респондентах.

Цель — спрогнозировать, превышает ли ежегодный доход человека порог в 50 000 долларов США. Целевая переменная — income, принимающая значения <=50K и >50K.

Для решения поставленной задачи необходимо:

1. Провести разведочный анализ данных (EDA): изучить распределения признаков, выявить пропуски и выбросы, оценить сбалансированность классов;
2. Выполнить обработку пропусков и кодирование категориальных признаков
3. Провести масштабирование числовых признаков;
4. Разделить данные на обучающую и тестовую выборки;
5. Построить не менее пяти моделей машинного обучения, включая как минимум две ансамблевые модели (например, Random Forest и Gradient Boosting);
6. Оценить качество моделей с помощью нескольких метрик: accuracy, F1-score, ROC AUC;
7. Выполнить подбор гиперпараметров с использованием GridSearchCV и кросс-валидации;
8. Сравнить результаты моделей до и после оптимизации и обосновать выбор финальной модели;
9. Создать интерактивное веб-приложение, позволяющее пользователю либо регулировать гиперпараметры модели и визуализировать метрики, либо вводить значения признаков и получать прогноз модели.

# ПОДГОТОВКА ДАННЫХ

Выбранный для исследования набор данных Adult Income (Census Income Dataset) включает следующие ключевые признаки:

* age — возраст респондента,
* education — уровень образования,
* marital-status — семейное положение,
* occupation — род профессиональной деятельности,
* hours-per-week — количество рабочих часов в неделю,
* sex — пол,
* native-country — страна происхождения,
* capital-gain — прирост капитала,
* capital-loss — убыток капитала,
* workclass — тип занятости,
* relationship — семейный статус в семье,
* race — расовая принадлежность.

Из исходного датасета были исключены такие признаки, как:

* fnlwgt — весовая переменная, не несущая прямой аналитической ценности,
* education-num — числовая версия признака
* education, дублирующая категориальную переменную,
* native-country — оставлена только в качестве категориального признака после обработки.

Также была удалена информация о уникальных идентификаторах, которая отсутствовала в исходном датасете, и любые другие служебные признаки, не влияющие непосредственно на результат классификации.

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 победителей.

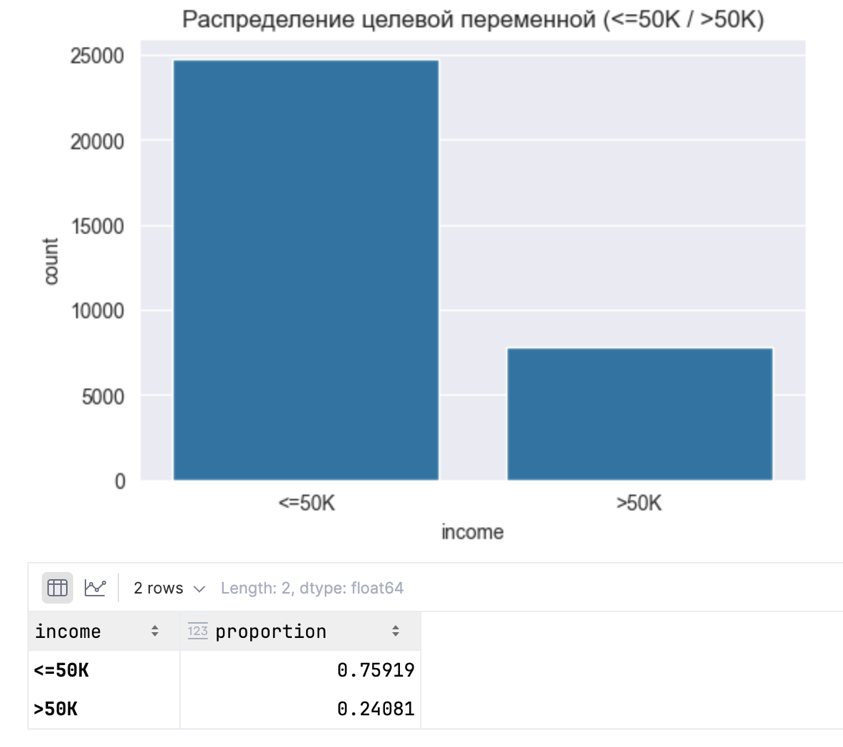


Рисунок 1 – Распределение целевой переменной

Проанализируем распределение всех признаков, как категориальных, так и числовых:

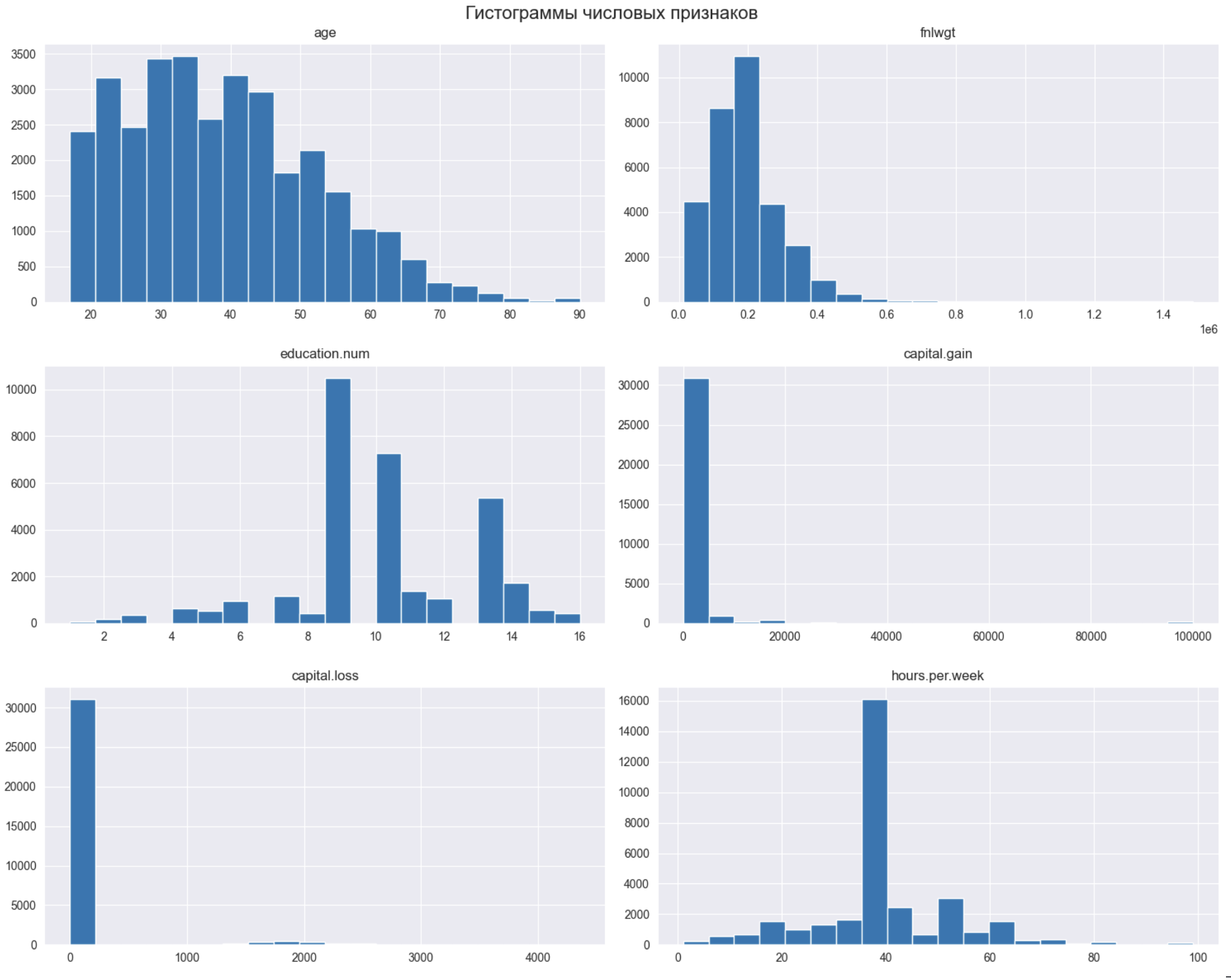


Рисунок 2 – Распределение числовых признаков

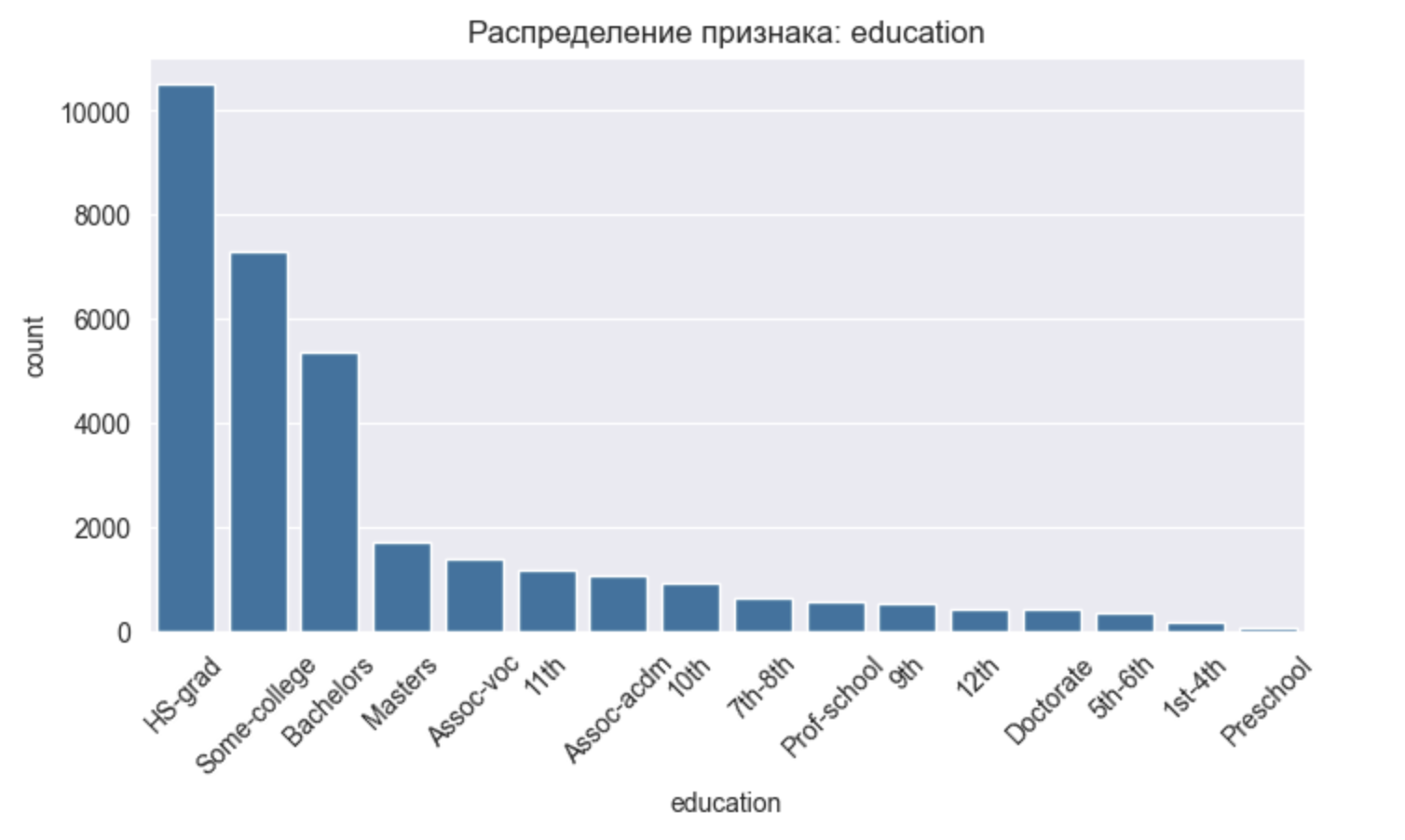


Рисунок 3 – Пример распределения категориального признака

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

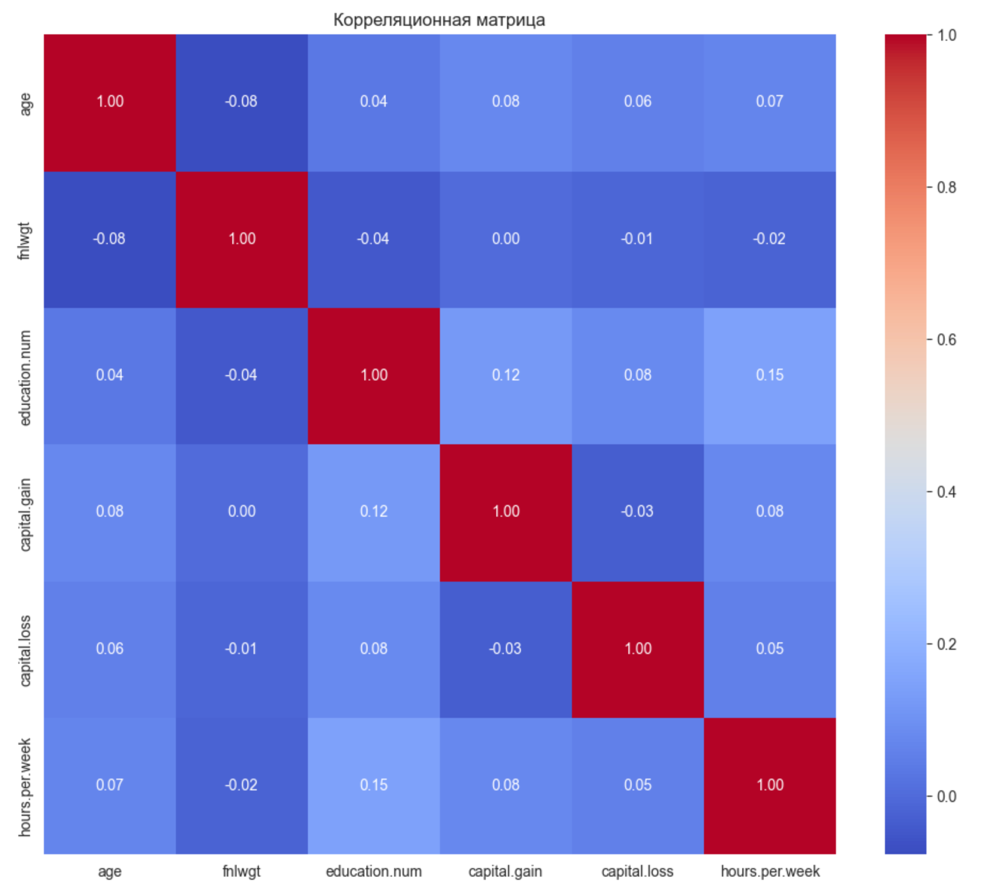


Рисунок 4 – Корреляционная матрица признаков

Матрица корреляций показывает взаимосвязи между признаками в датасете.

Для более подробного анализа влияния конкретного значения категориального признака на целевую переменную построим графики для каждого признака:

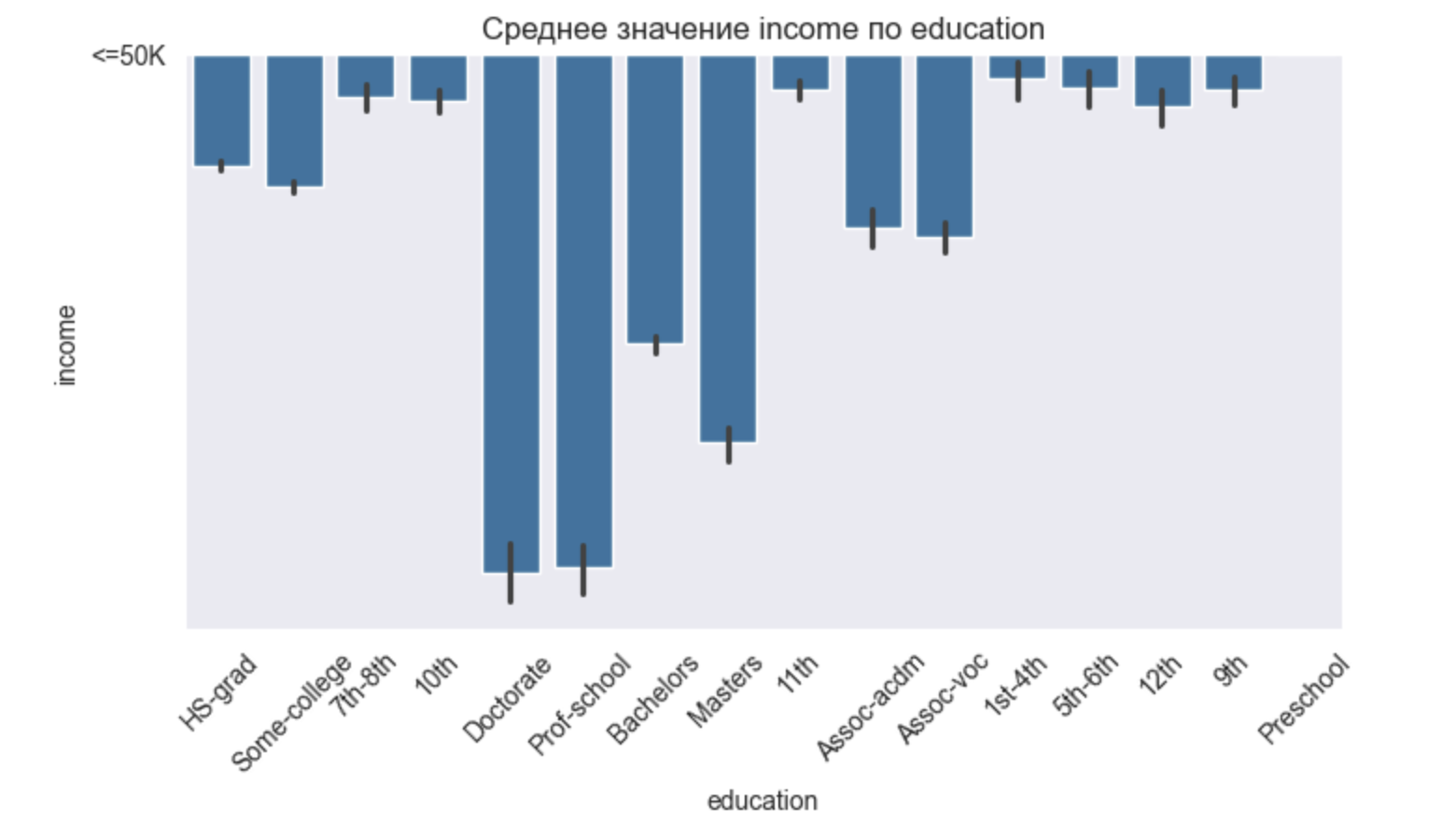


Рисунок 5 – Пример корреляции целевой переменной с категориальным признаком

1. ОБРАБОТКА И ПРЕОБРАЗОВАНИЕ ПРИЗНАКОВ

Перед построением моделей был проведён ряд подготовительных шагов по обработке датасета:

* преобразование целевой переменной income в бинарный формат (0 — доход ≤ 50K$, 1 — доход > 50K$);
* удаление неинформативных признаков (например, fnlwgt, education-num, а также дублирующих или служебных признаков);
* обработка пропусков в категориальных признаках (удаление строк с пропущенными значениями);
* преобразование категориальных признаков с помощью one-hot encoding;
* масштабирование числовых признаков с использованием стандартизации.

5. ПОСТРОЕНИЕ И СРАВНЕНИЕ МОДЕЛЕЙ

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

1. Логистическая регрессия (LogisticRegression)
2. Дерево решений (DecisionTreeClassifier)
3. Случайный лес (RandomForestClassifier)
4. Градиентный бустинг (GradientBoostingClassifier)
5. Метод К-ближайших соседей (KNN)

Все модели были обучены на одной и той же выборке (80% обучающая, 20% тестовая), и поначалу использовались базовые параметры без настройки.

По результатам базового обучения была получена следующая статистика:

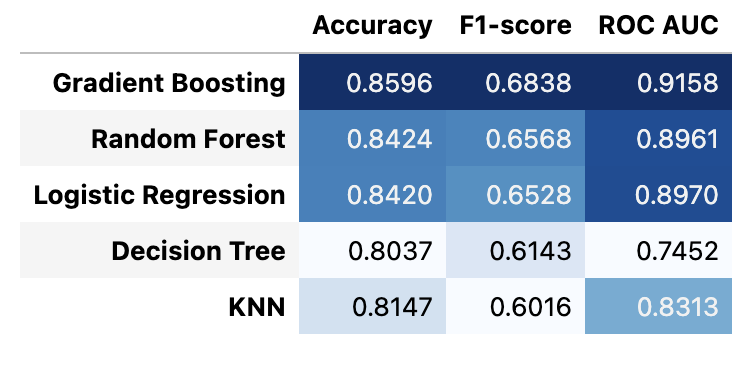


Рисунок 6 – Таблица результатов базового обучения

1. НАСТРОЙКА ГИПЕРПАРАМЕТРОВ

После проведения GridSearchCV и повторной оценки моделей наблюдается улучшение производительности у всех алгоритмов:

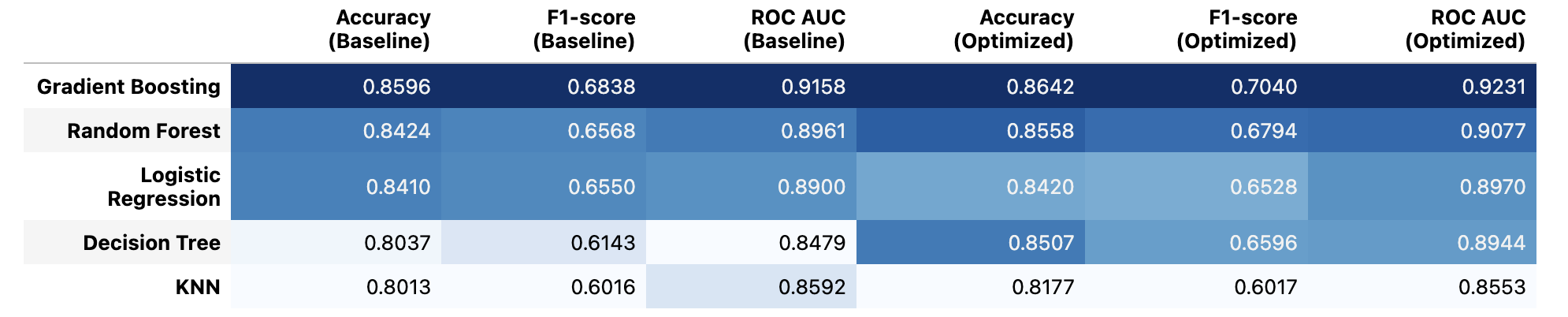


Рисунок 7 – Таблица с результатами оптимальных моделей

Построим сравнительные гистограммы:

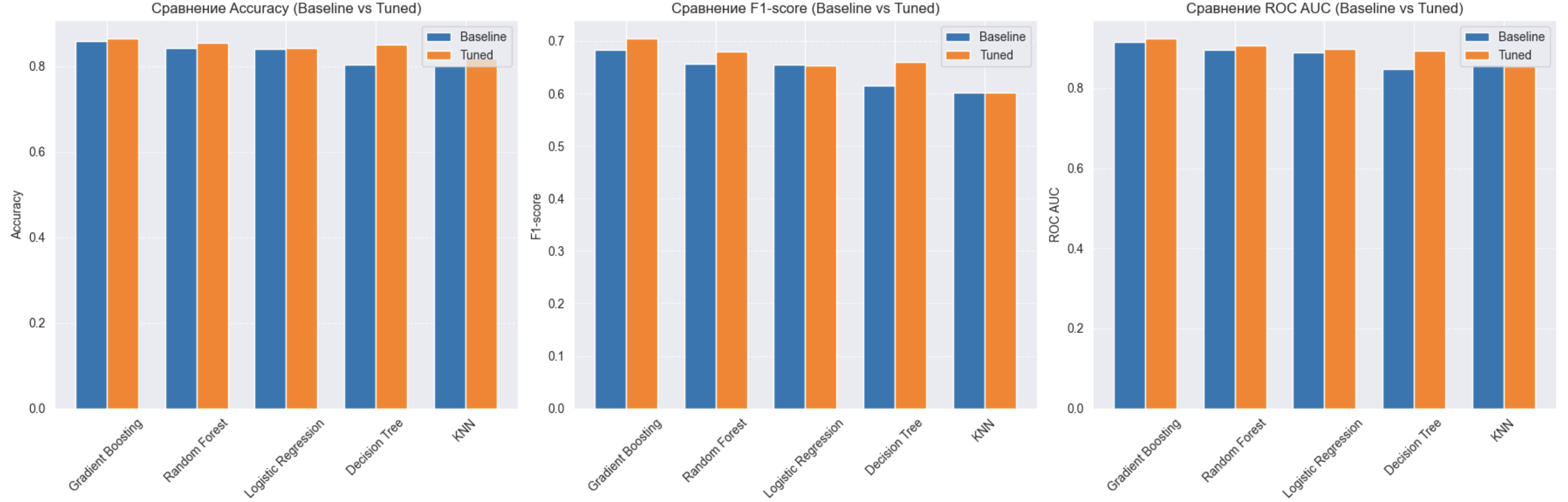


Рисунок 8 – Сравнение моделей по F1-Score

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

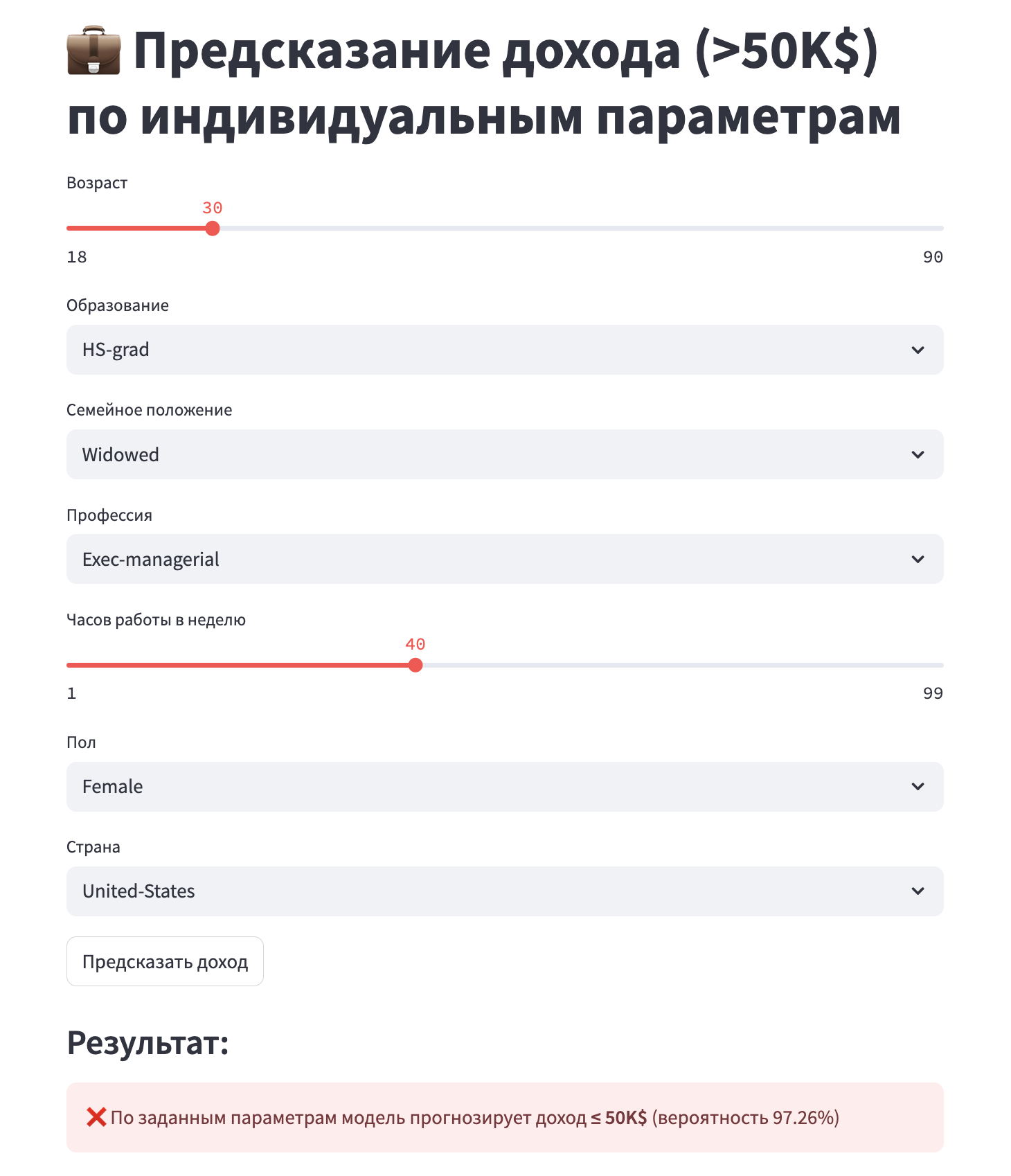
1. ВЕБ-ПРИЛОЖЕНИЕ

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



**Рисунок 9 – 1 страница веб-приложения**

Изменяя параметры, на 1 странице мы можем переобучать модель и сравнивать получившиеся параметры, подбирая оптимальные.



**Рисунок 10 – 2 страница веб-приложения**

На 2 странице веб-приложения мы можем с помощью обученной модели по введённому ряду параметров сделать прогноз целевого значения.

ЗАКЛЮЧЕНИЕ

Результаты данной работы продемонстрировали, что при системном и поэтапном подходе к задаче бинарной классификации можно значительно улучшить качество предсказаний за счёт следующих факторов:

* Проведение полноценного разведочного анализа и осмысленная работа с признаками;
* Корректная обработка пропусков, кодирование категориальных переменных и масштабирование числовых данных;
* Использование современных моделей машинного обучения, включая ансамблевые методы, и тщательный подбор гиперпараметров;
* Оценка результатов по нескольким метрикам (Accuracy, F1-score, ROC AUC), что позволяет делать более взвешенные и обоснованные выводы.

Наилучшие результаты были достигнуты с использованием модели градиентного бустинга. После оптимизации гиперпараметров она продемонстрировала F1-score выше 0.7, а также высокую сбалансированность между точностью и полнотой, что делает её наиболее подходящим решением для поставленной задачи.

Построенные модели и весь цикл работы могут быть адаптированы для решения других задач классификации, что подтверждает универсальность и воспроизводимость подхода.

Дополнительно, был реализован интерактивный веб-интерфейс с использованием Streamlit, позволяющий как настраивать гиперпараметры, так и вводить пользовательские данные для получения прогноза в реальном времени. Это открывает возможности для внедрения модели в прикладные решения и использования в продуктивной среде.

**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Kaggle: Chess Game Dataset (https://www.kaggle.com/datasets/datasnaek/chess)
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/>