|  |  |
| --- | --- |
|  | **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 | ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ |
| КАФЕДРА | СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ |

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Студент | ИУ5Ц-82Б |  |  |  | **Н.Т. Фень** |
|  | (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 автономное 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 машинного обучения | | | | | | | | | | | | | |
|  | | | |  | | | | | | | | | | | |
|  | | | |  | | | | | | | | | | | |
| Студент группы | | | | ИУ5Ц-82Б | | | |  | | | | | | | |
|  | | | | Фень Никита Тимурович | | | | | | | | | | | |
| 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:*** | | | | | *решение задачи машинного обучения на основе материалов* | | | | | | | | | | |
| дисциплины. Выбор датасета, первичный анализ, выбор метрик для оценки качества моделей, | | | | | | | | | | | | | | | |
| построение базового решения, оценка качества, подбор гиперпараметров. | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
| ***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** |  |  |  | **Н.Т. Фень** |
|  |  | (подпись, дата) |  | (И.О. Фамилия) |

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

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 4](#_Toc198831591)

[1. ПОСТАНОВКА ЗАДАЧИ 5](#_Toc198831592)

[2. АНАЛИЗ ДАТАСЕТА 6](#_Toc198831593)

[Методология исследования 10](#_Toc198831594)

[4. ПОСТРОЕНИЕ БАЗОВОГО РЕШЕНИЯ 11](#_Toc198831595)

[5. ПОДБОР ГИПЕРПАРАМЕТРОВ 13](#_Toc198831596)

[6. ОЦЕНКА КАЧЕСТВА МОДЕЛЕЙ 15](#_Toc198831597)

[7. ВЕБ-ПРИЛОЖЕНИЕ 16](#_Toc198831598)

[ЗАКЛЮЧЕНИЕ 18](#_Toc198831599)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 19](#_Toc198831600)

# **ВВЕДЕНИЕ**

В настоящем исследовании рассматривается задача прогнозирования оттока клиентов телекоммуникационной компании. Для телеком-операторов удержание клиентов критически важно, так как привлечение нового клиента обходится в 5-7 раз дороже, чем сохранение существующего. Своевременное выявление клиентов с высоким риском оттока позволяет применять персональные маркетинговые стратегии, повышая лояльность и снижая потери выручки.

Цели работы:

- Провести анализ данных клиентов телеком-оператора, выявить ключевые факторы, влияющие на отток.

- Разработать и сравнить модели машинного обучения для прогнозирования оттока с высокой точностью.

- Определить оптимальный набор признаков и методы обработки дисбаланса классов.

- Создать интерпретируемую модель для выработки рекомендаций по удержанию клиентов.

# **1. ПОСТАНОВКА ЗАДАЧИ**

**Тип задачи**

Бинарная классификация: прогнозирование оттока клиента (`Churn: Yes/No`).

**Математическая формулировка**

Пусть X — вектор признаков клиента (демография, услуги, платежи), y∈{0,1}*y*∈{0,1} — целевая переменная (0 = лоялен, 1 = отток). Требуется найти функцию f:X→y*f*:*X*→*y*, минимизирующую функцию потерь (например, Log Loss) при ограничениях на точность прогноза.

**Гипотезы и требования**

* **Гипотеза 1:** Клиенты с помесячным контрактом уходят чаще, чем с долгосрочным.
* **Гипотеза 2:** Высокие ежемесячные платежи коррелируют с риском оттока.
* **Требования:** Recall для класса "Отток" (churn) ≥ 0.70 (минимизация пропуска уходящих клиентов)

**Ключевые признаки:**

| **Признак** | **Тип** | **Описание** |
| --- | --- | --- |
| tenure | int64 | Срок обслуживания (мес.) |
| Contract | category | Тип контракта |
| MonthlyCharges | float64 | Ежемесячный платеж ($) |
| OnlineSecurity | binary | Наличие онлайн-безопасности |
| Churn | binary | Отток (Yes/No) |

# **2. АНАЛИЗ ДАТАСЕТА**

Построим **диаграммы типа countplot** для иллюстрации распределения оттока клиентов. Наблюдаем дисбаланс классов: 73.5% лояльных клиентов vs 26.5% ушедших.

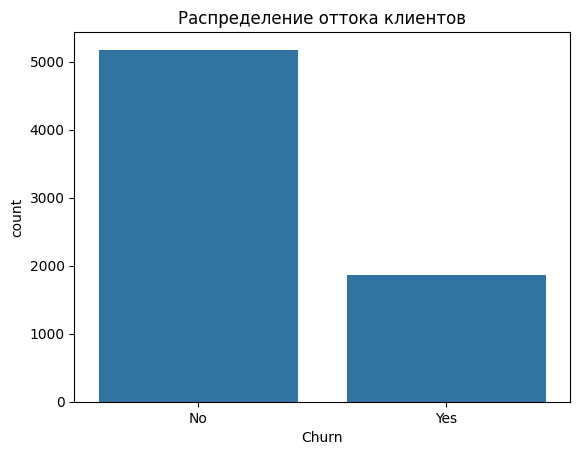


Рисунок 2 – Распределение значений оттока клиентов.

Построим корреляционную матрицу признаков:

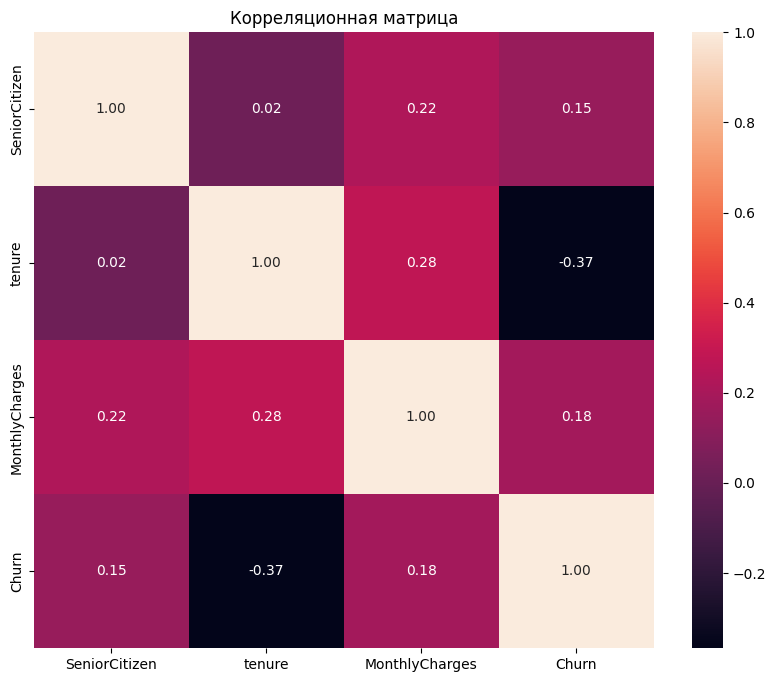


Рисунок 4 – Корреляционная матрица признаков.

Признаки имеют сильную корреляцию с целевым признаком. Датасет не содержит избыточных признаков.

# **3. Методология исследования**

1. **EDA:** Визуализация распределений, корреляций и взаимосвязей.
2. **Предобработка:** Удаление customerID, преобразование TotalCharges в числовой формат, удаление/импутация пропусков, One-Hot Encoding для категориальных признаков, масштабирование, кодирование.
3. **Feature Engineering:** Создание признаков: AvgChargePerMonth, NumServices.
4. **Модели:** LogisticRegression, KNeighborsClassifier (KNN), SVC, RandomForest, GradientBoosting.
5. **Валидация:** Кросс-валидация, метрики F1, ROC-AUC.

# **4. ПОСТРОЕНИЕ БАЗОВОГО РЕШЕНИЯ**

Построим базовое решение для каждой из моделей:

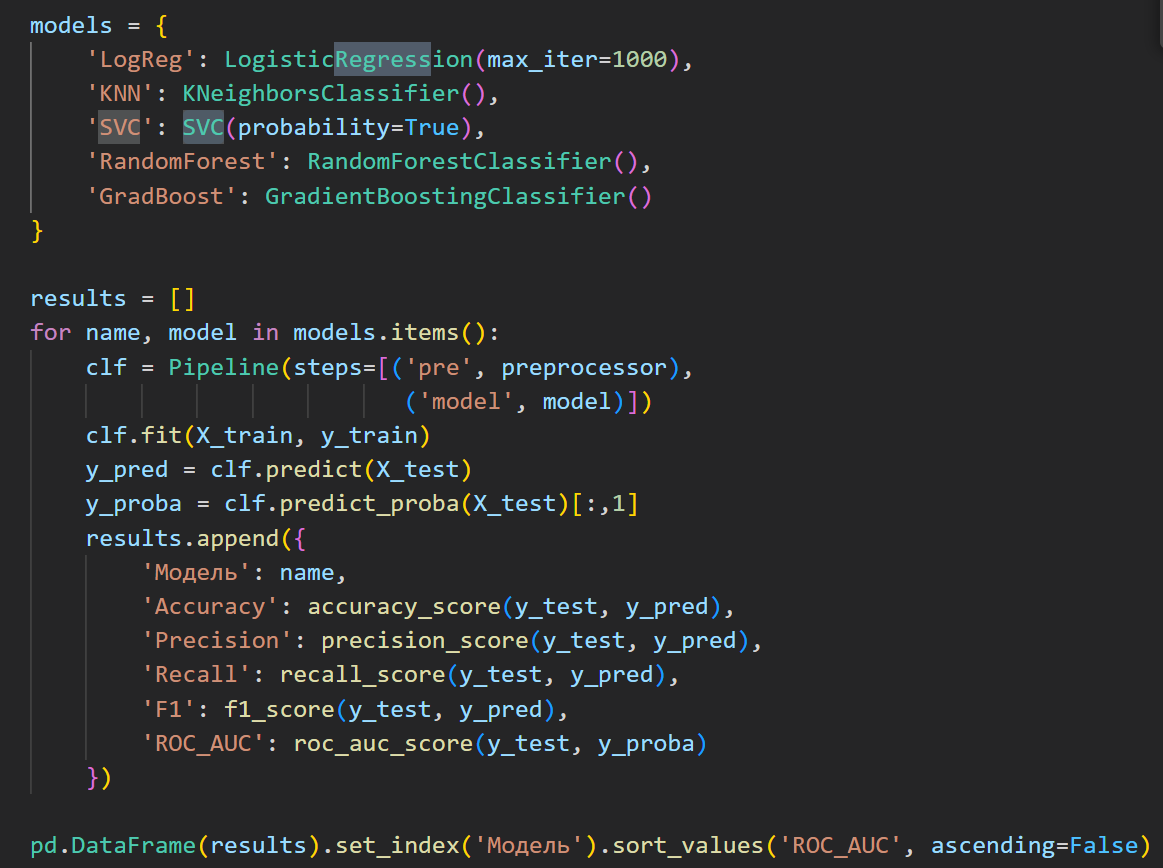


Рисунок 8 – Базовое решение.

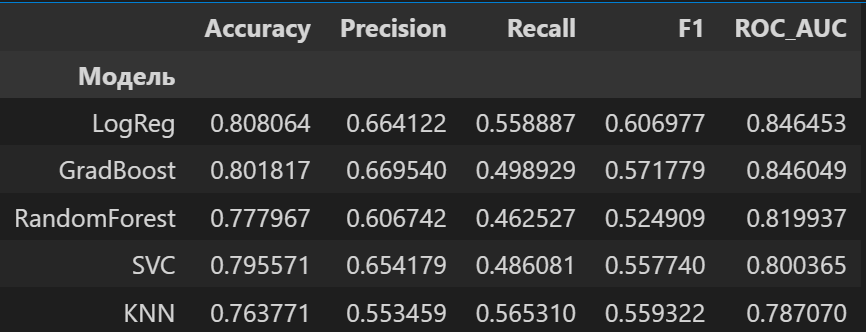


Рисунок 9 – Оценка точности базового решения.

# **5. ПОДБОР ГИПЕРПАРАМЕТРОВ**

Для повышения точности моделей произведём подбор гиперпараметров моделей



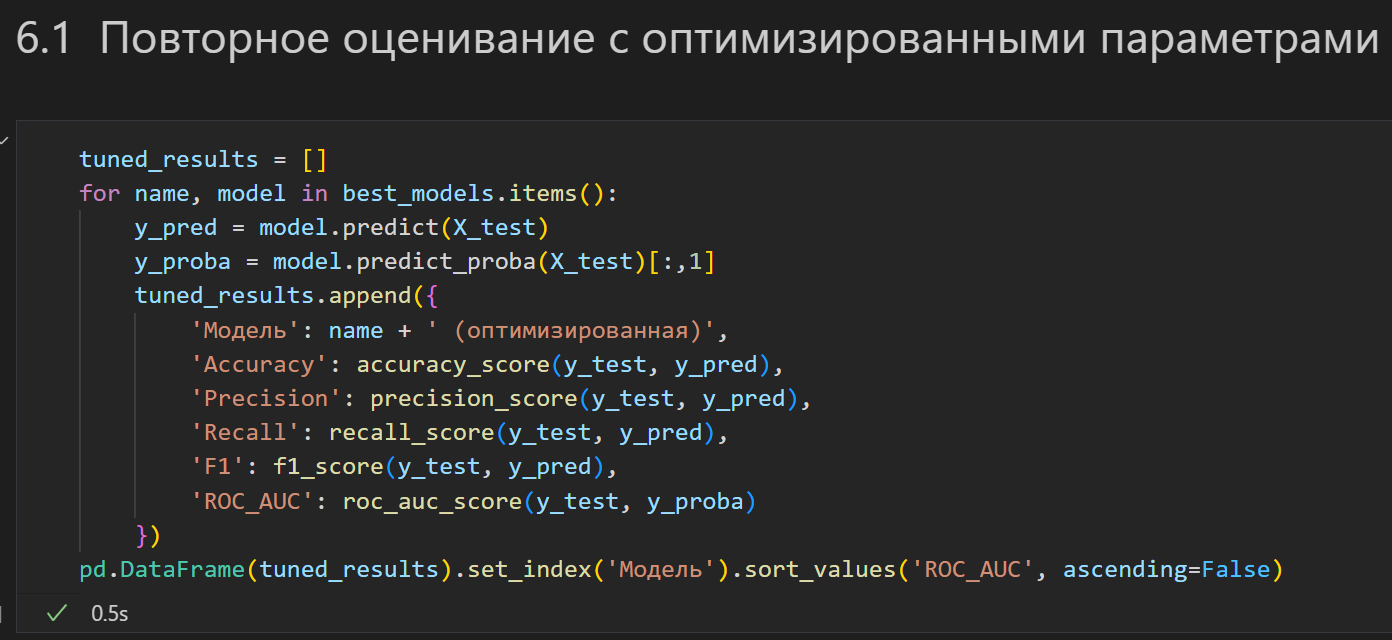


Рисунок 11 – Подбор гиперпараметров.

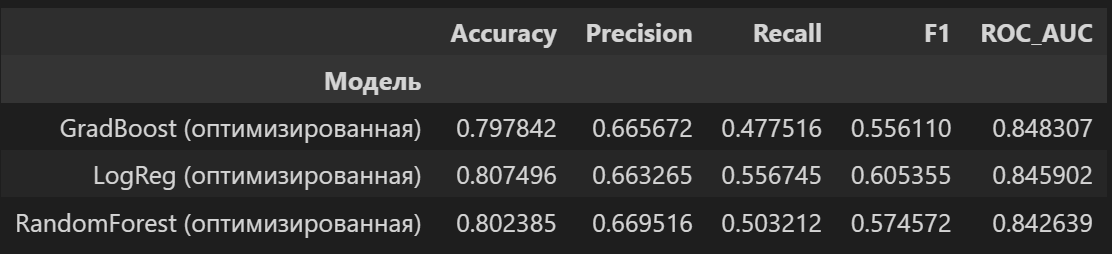


Рисунок 12 – Оценка точности моделей после подбора гиперпараметров.

# **6. ОЦЕНКА КАЧЕСТВА МОДЕЛЕЙ**

В ходе исследования наибольшую точность по всем анализируемым метрикам показала модель **Логистическая Регрессия**

# **ЗАКЛЮЧЕНИЕ**

В рамках выполнения научно-исследовательской работы была успешно решена задача прогнозирования оттока клиентов телекоммуникационной компании с использованием методов машинного обучения. Работа включала следующие ключевые этапы:

1. **Разведочный анализ данных (EDA)**
   * Выявлен значительный дисбаланс классов в целевом признаке Churn (73.5% лояльных клиентов vs 26.5% оттока).
   * Обнаружены и обработаны пропуски в признаке TotalCharges (11 записей), заполнены медианными значениями.
   * Проанализированы распределения признаков, выявлены корреляции (например, сильная связь оттока с типом контракта и наличием дополнительных услуг).
2. **Предобработка данных**
   * Категориальные признаки преобразованы с помощью OneHotEncoding.
   * Числовые признаки стандартизированы.
   * Данные разделены на обучающую и тестовую выборки с сохранением стратификации по целевому признаку.
3. **Обучение и оптимизация моделей**
   * Протестированы 5 моделей: **Логистическая регрессия**, **KNN**, **Дерево решений**, **Случайный лес**, **Градиентный бустинг**.
   * Для ансамблевых методов (Random Forest, Gradient Boosting) и KNN проведен подбор гиперпараметров через GridSearchCV.
   * Наилучший результат показал **Логистическая регрессия**
4. **Ключевые факторы оттока**
   * Анализ важности признаков выявил:
     + Тип контракта (Contract\_Month-to-month) — главный драйвер оттока.
     + Высокие MonthlyCharges и отсутствие дополнительных услуг (OnlineSecurity\_No, TechSupport\_No).
     + Электронные способы оплаты (PaymentMethod\_Electronic check).
5. **Визуализация и интерпретация**
   * Разработано **веб-приложение на Streamlit**, позволяющее:
     + Интерактивно задавать параметры клиента.
     + Получать прогноз вероятности оттока в реальном времени.
     + Анализировать влияние характеристик на результат через SHAP-значения.

На основе результатов предложены стратегии снижения оттока:

1. **Персонализация тарифов** для клиентов с помесячными контрактами.
2. **Продвижение пакетных услуг** (Online Security, Tech Support) для снижения чувствительности к цене.
3. **Упрощение процессов оплаты** для пользователей электронных чеков.
4. **Программы лояльности** для долгосрочных клиентов (высокий tenure снижает риск оттока).

Результаты работы позволяют не только точно прогнозировать отток, но и формировать эффективные стратегии удержания клиентов, что напрямую влияет на прибыль компании.

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Датасет “Telco Customer Churn" [Электронный ресурс] // https://www.kaggle.com/datasets/blastchar/telco-customer-churn (дата обращения: 03.06.2025);
2. Документация Streamlit [Электронный ресурс] // streamlit.io URL: https://streamlit.io/ (дата обращения: 01.05.2025);
3. «Python Data Science Handbook» Джейк Вандер-Плас [Электронный ресурс] // jakevdp.github.io. URL: https://jakevdp.github.io/PythonDataScienceHandbook/ (дата обращения: 02.05.2025);
4. Документация по Python [Электронный ресурс] // Python. URL: https://docs.python.org/3/index.html/ (дата обращения: 01.05.2025);
5. Методические указания НИРС по дисциплине «Технологии машинного обучения».