|  |  |
| --- | --- |
|  | **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)** |

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\_\_\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_ Н.Д. Голумин\_\_\_\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** **\_\_\_\_\_****Ю.Е. Гапанюк\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2024 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)

«\_07\_» \_\_\_\_февраля\_\_\_\_\_\_\_ 2024 г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_\_Разработка и оценка моделей машинного обучения\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_РТ5-61Б\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Голумин Никита Дмитриевич\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

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., 100% 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\_\_\_\_\_\_\_ 2024 г.

**Руководитель НИР**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_Ю.Е. Гапанюк\_\_\_\_

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** \_\_\_\_\_\_\_\_\_Н. Д. Голумин**\_\_**

(Подпись, дата) (И.О.Фамилия)

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

СОДЕРЖАНИЕ

[Введение 3](#_Toc166636917)

[Основная часть 5](#_Toc166636918)

[Заключение 14](#_Toc166636919)

### Введение

В современном мире конкуренция за места в аспирантуре становится все более жесткой, особенно в области науки и технологий. Успех в конкурсе за место в аспирантуре зависит от многих факторов, включая академические достижения, мотивацию, опыт исследований и даже личные качества кандидата. В этом контексте, понимание того, какие факторы оказывают наибольшее влияние на принятие решения о приеме, может значительно повысить шансы абитуриентов на успешное поступление.

В данной работе мы сосредоточимся на анализе набора данных по поступлению в аспирантуру в Индии, предоставленного Kaggle. Этот набор данных включает информацию о выпускных баллах GRE и TOEFL, рейтинге университета, мотивации, среднем балле бакалавра, опыте исследований и шансе на поступление. Целью работы является разработка и оценка моделей машинного обучения для прогнозирования шансов на поступление в аспирантуру на основе этих факторов.

Модели машинного обучения могут предложить ценный инструмент для анализа сложных взаимосвязей между различными признаками и их влиянием на итоговое решение о приеме. Они могут помочь выявить, какие факторы являются наиболее важными для комитета по приему, и таким образом, помочь абитуриентам лучше подготовиться к конкурсу.

В ходе работы будет детальный анализ данных, включая предварительный исследовательский анализ, выбор и кодирование признаков, корреляционный анализ и подбор гиперпараметров для моделей машинного обучения. Я также сравню различные модели, чтобы определить, какая из них наиболее эффективна в прогнозировании шансов на поступление.

Результаты работы могут служить полезным руководством для абитуриентов, стремящихся улучшить свои шансы на поступление в аспирантуру, а также для исследователей и специалистов в области образования, интересующихся вопросами, влияющими на процесс приема в аспирантуру.

### Основная часть

**Постановка Задачи**

Цель этой работы — разработать и оценить модели машинного обучения для прогнозирования вероятности поступления в аспирантуру в Индии, используя набор данных с Kaggle: https://www.kaggle.com/datasets/mohansacharya/graduate-admissions. В качестве целевой переменной для задачи регрессии выбрана вероятность поступления ("Chance of Admit").

Датасет состоит из одного файла Admission\_Predict со следующими колонками

GRE Scores ( out of 340 ) - выпускной балл

TOEFL Scores ( out of 120 ) - балл ТОЕФЛ

University Rating ( out of 5 ) - рейтинг университета

Statement of Purpose ( out of 5 ) - мотивация

Undergraduate GPA ( out of 10 ) - средний балл бакалавра

Research Experience ( either 0 or 1 ) - опыт исследований

Chance of Admit ( ranging from 0 to 1 ) - шанс поступления

**Последовательность действий**

1. Проведение разведочного анализа данных. Построение графиков, необходимых для понимания структуры данных. Анализ и заполнение пропусков в данных.

Размер датасета - 400 строк, 9 колонок,

Список колонок с типами данных:

Serial No. int64

GRE Score int64

TOEFL Score int64

University Rating int64

SOP float64

LOR float64

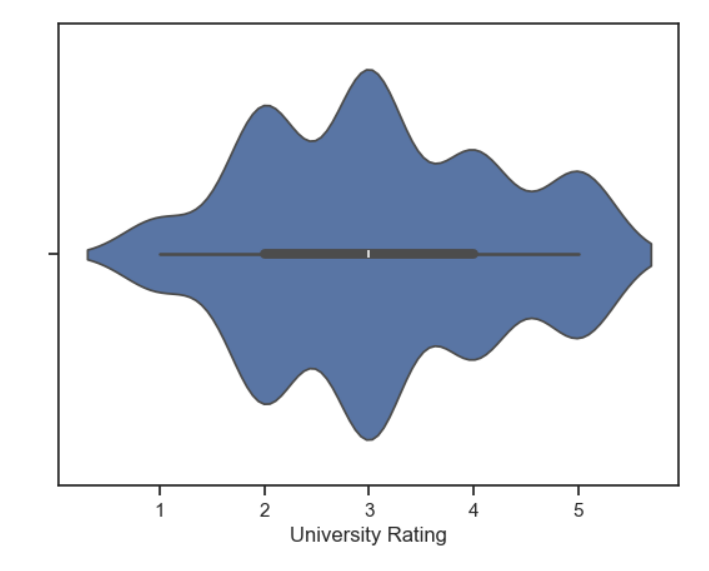
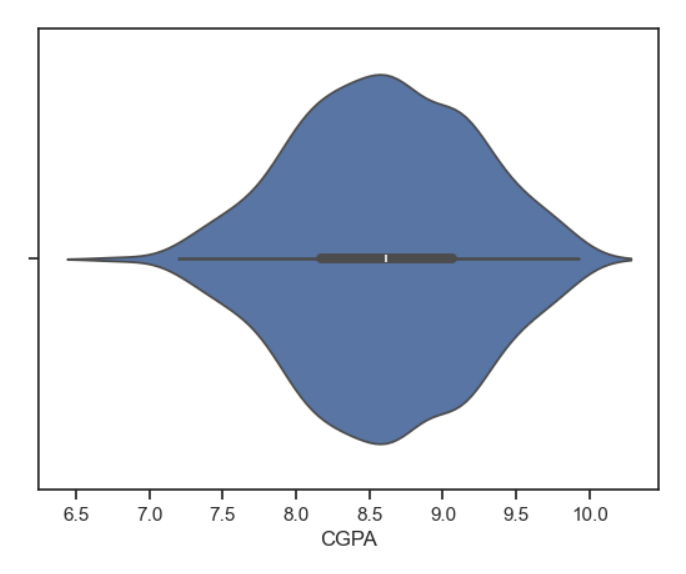
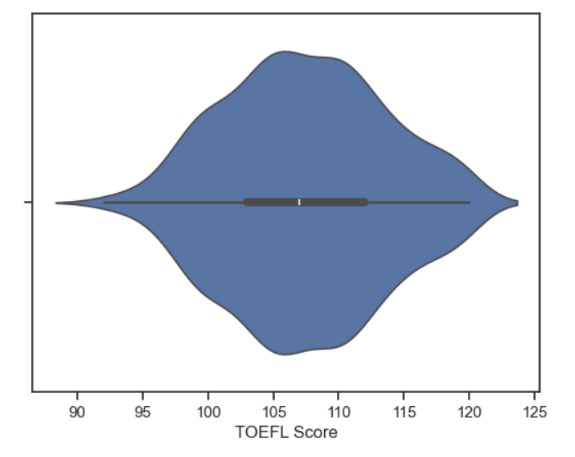
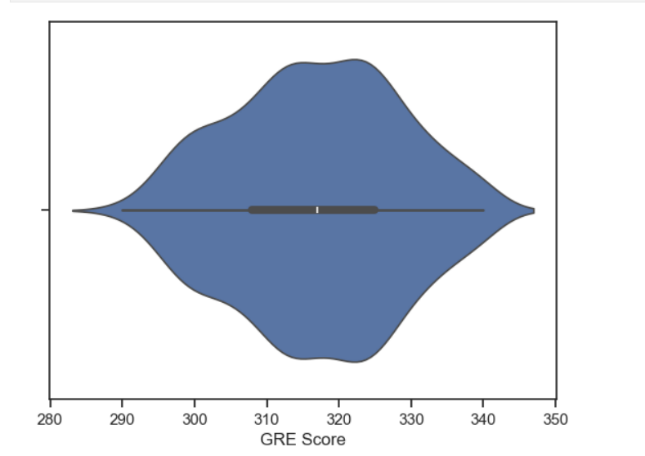
CGPA float64

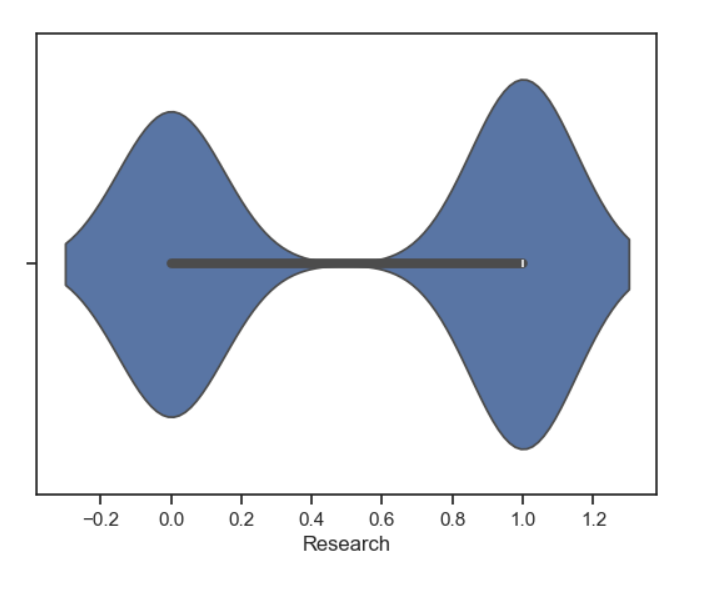
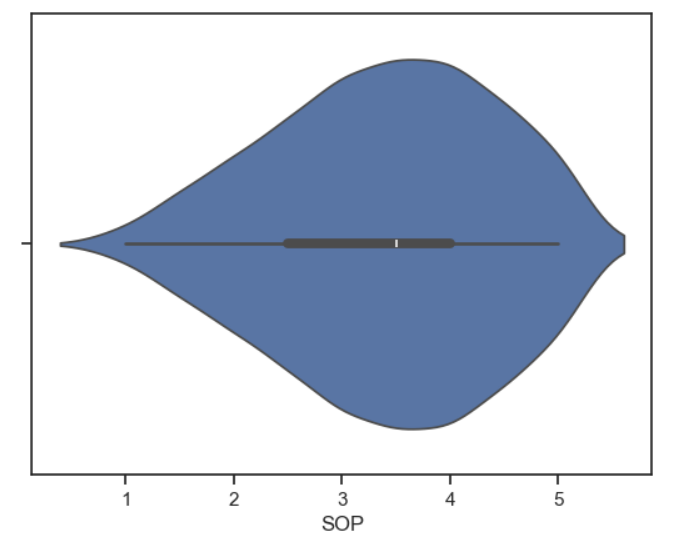
Research int64

Chance of Admit float64,

Пустых значений не обнаружено, все категориальные признаки уже были закодированы

Скрипичные диаграммы:



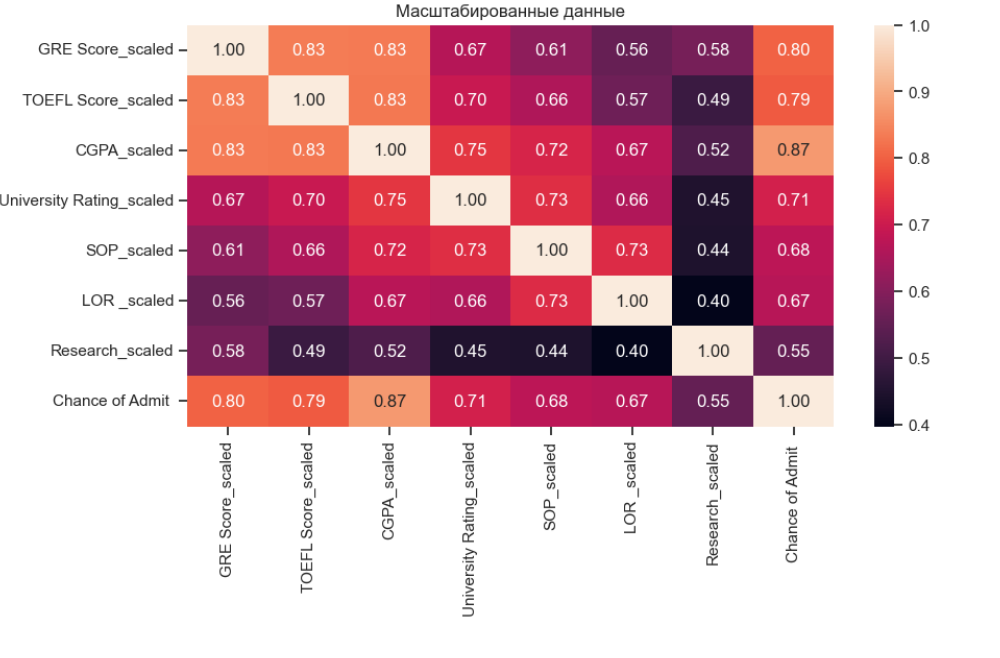


1. Подготовка Данных: В процессе подготовки данных были выполнены следующие действия:
   * Масштабирование числовых признаков, чтобы обеспечить равномерное распространение значений и улучшить стабильность обучения моделей.
2. Выбор Признаков: Был проведен анализ корреляции между признаками и целевой переменной "Chance of Admit" для определения наиболее значимых признаков, которые будут использоваться для построения моделей.

На основе корреляционной матрицы можно сделать следующие выводы:

- Корреляционные матрицы для исходных и масштабированных данных совпадают.

- Целевой признак классификации "Chance of Admit" наиболее сильно коррелирует с средниv баллом бакалавра (0.87), баллом TOEFL (0,79) и выпускным баллом (0.79). Однако между этими тремя признаками есть сильная корреляция, поэтому возьмем только признак "GRE Score", так как он имеет самую большую корреляцию. Также возьмем признак "University Rating" (0.71)

- Большие по модулю значения коэффициентов корреляции свидетельствуют о значимой корреляции между исходными признаками и целевым признаком. На основании корреляционной матрицы можно сделать вывод о том, что данные позволяют построить модель машинного обучения.

1. Выбор метрик для последующей оценки качества моделей.

Mean absolute error - средняя абсолютная ошибка.

Mean squared error - средняя квадратичная ошибка

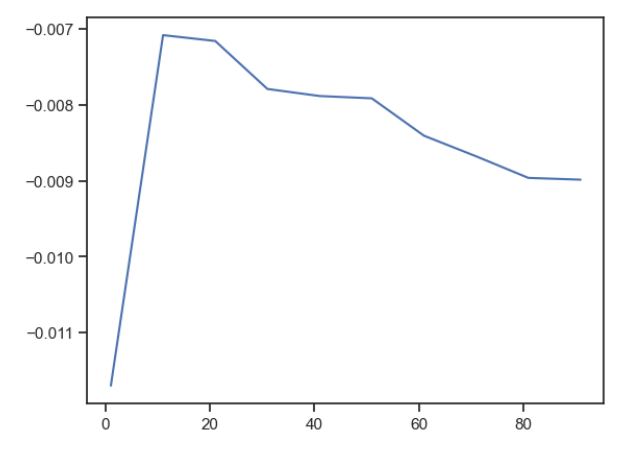
Метрика R2

MAE и MSE взяты как как важна направленность ошибки, R2 так как важно, насколько хорошо модель объясняет вариацию в данных

1. Выбор наиболее подходящих моделей для решения задачи регрессии.

* Линейная регрессия
* Метод ближайших соседей
* Метод опорных векторов
* Решающее дерево
* Случайный лес
* Градиентный бустинг

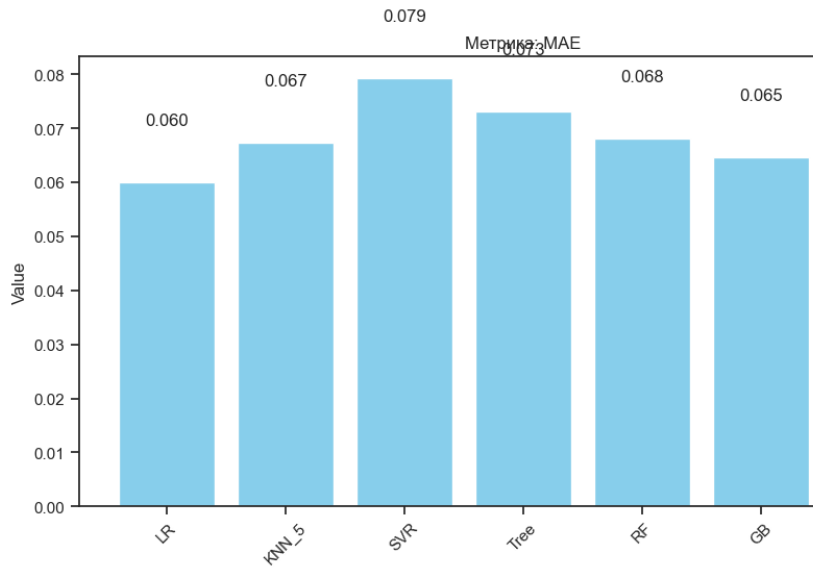
1. Формирование обучающей и тестовой выборок на основе исходного набора данных.
2. Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров.
3. LR : MAE=0.06, MSE=0.007, R2=0.708
4. KNN\_5: MAE=0.067, MSE=0.009, R2=0.621
5. SVR : MAE=0.079, MSE=0.011, R2=0.499
6. MAE=0.073, MSE=0.01, R2=0.569
7. RF :MAE=0.068, MSE=0.009, R2=0.626
8. GB : MAE=0.065, MSE=0.008, R2=0.643
9. Подбор гиперпараметров для выбранных моделей. Анализ Результатов

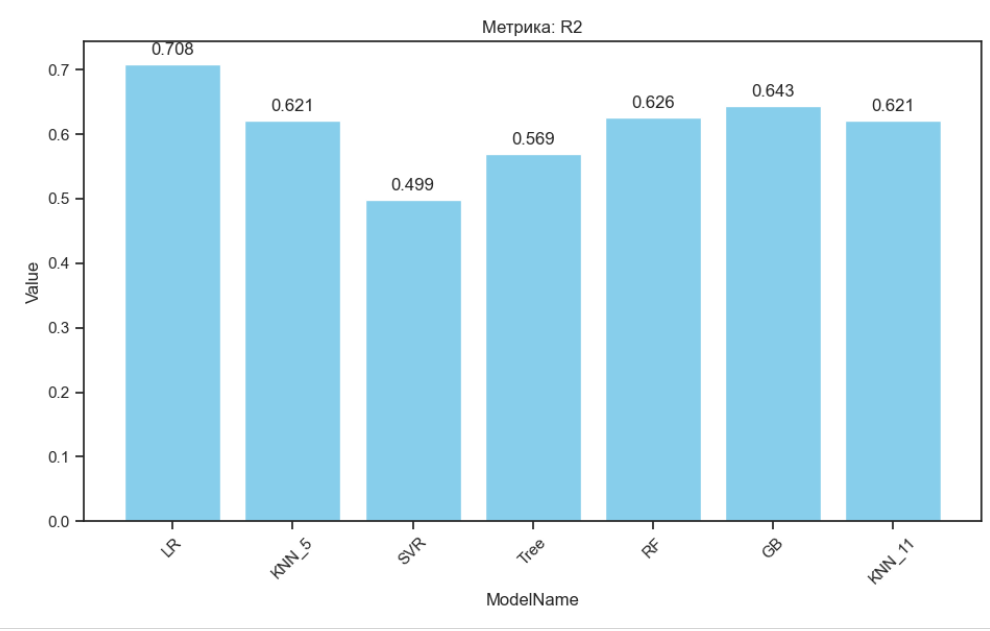


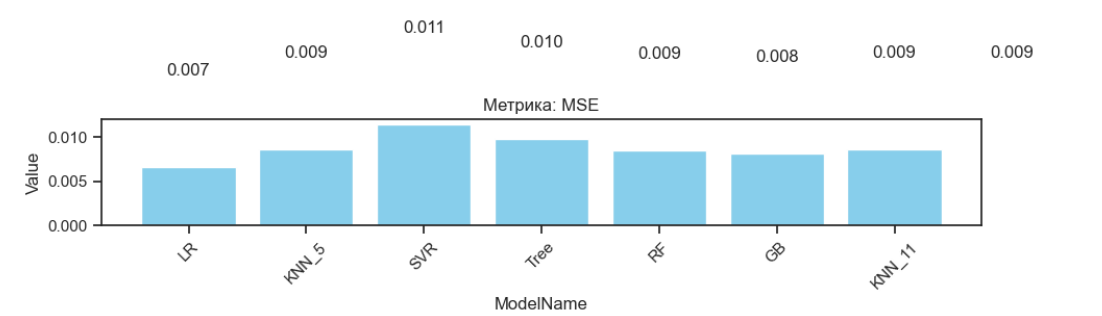
1. Сравнение значений базового и с гиперпараметром
2. KNN\_5 : MAE=0.067, MSE=0.009, R2=0.621
3. KNN\_11 : MAE=0.065, MSE=0.009, R2=0.62
4. LR : MAE=0.067, MSE=0.009, R2=0.621
5. LR\_True : MAE=0.06, MSE=0.007, R2=0.708
6. RF : MAE=0.068, MSE=0.008, R2=0.633
7. RF\_BestParams : MAE=0.059, MSE=0.007, R2=0.706
8. SVR\_Base : MAE=0.079, MSE=0.011, R2=0.499
9. SVR\_Optimized : MAE=0.064, MSE=0.007, R2=0.696
10. GB\_Base : MAE=0.065, MSE=0.008, R2=0.643
11. GB\_Optimized : MAE=0.062, MSE=0.007, R2=0.675

Итог: гиперпараметры помогли незначительно улучшить результаты метрик

1. Формулирование Выводов.







По графикам видно, что лучше всего себя показала линейная регрессия.

### Заключение

В ходе данной работы проведено всестороннее исследование для прогнозирования шансов на поступление в аспирантуру в Индии с использованием набора данных с Kaggle. Разработаны и оценены различные модели машинного обучения, включая как базовые, так и оптимизированные с применением подбора гиперпараметров.

Результаты показали, что подбор гиперпараметров незначительно улучшает точность прогнозирования, делая модели более точными и надежными. Особенно эффективно себя показала модель линейной регрессии.

Таким образом, созданные модели машинного обучения и проведенный анализ позволяют не только прогнозировать шансы на поступление в аспирантуру, но и выявлять ключевые факторы, которые следует учитывать при подготовке к конкурсу. Это может стать ценным инструментом как для абитуриентов, так и для образовательных учреждений, стремящихся оптимизировать процесс приема.

В будущем, расширение анализа за счет дополнительных признаков и использование более продвинутых методов машинного обучения может еще больше повысить точность прогнозирования и дать более глубокое понимание механизмов принятия решений в процессе поступления в аспирантуру.