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 ИУ5-62Б (группа) _____ (подпись, дата) **В. Мажитов** (И.О. Фамилия)

Руководитель НИР _____ (подпись, дата) **Ю.Е. Гапанюк** (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)

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме Разработка и оценка моделей машинного обучения

Студент группы ИУ5-62Б
Мажитов Вадим

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 _____ **В. Мажитов**
(подпись, дата) (И.О. Фамилия)

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

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В данном исследовании анализируется набор данных "Seoul Bike Sharing Demand Prediction", содержащий почасовые данные о количестве арендованных велосипедов в Сеуле, а также сопутствующие погодные условия и информацию о времени суток/дня недели/праздниках.

Целью работы является:

1. Понимание факторов, влияющих на спрос на аренду велосипедов.
2. Построение моделей машинного обучения для решения двух задач:
   - Задача регрессии: предсказание точного количества арендованных велосипедов в данный час.
   - Задача классификации: предсказание категории спроса (низкий, средний, высокий).

Набор данных включает информацию о дате, времени, погодных условиях (температура, влажность, скорость ветра, видимость, точка росы, солнечное излучение, снегопад, дождь) и количестве арендованных велосипедов.

1. Обзор данных и предобработка

Загрузка и начальный обзор данных:

- Загрузка датасета SeoulBikeData.csv с использованием библиотеки pandas.
- Отображение первых строк данных (.head()).
- Проверка типов данных и наличия пропущенных значений (.info()).

Первые 5 строк данных:

	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

Преобразование признаков:

- Преобразование столбца 'Date' в формат datetime.
- Извлечение новых признаков: 'Year', 'Month', 'Day', 'Weekday', 'Hour' из 'Date' и 'Hour'.
- Удаление исходных столбцов 'Date' и 'Holiday'.

```
df['Date'] = pd.to_datetime(df['Date'], format='%d/%m/%Y')

# Извлечение дополнительных временных признаков
df['Year'] = df['Date'].dt.year
df['Month'] = df['Date'].dt.month
df['Day'] = df['Date'].dt.day
df['DayOfWeek'] = df['Date'].dt.dayofweek # Понедельник=0, Воскресенье=6
df['Hour'] = df['Hour'] # 'Hour' уже есть, но убедимся что он числовой
```

Обработка категориальных признаков:

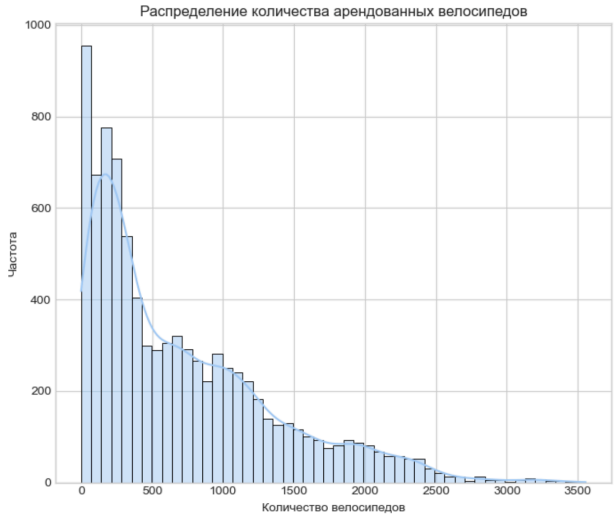
- Применение One-Hot Encoding (OHE) для категориальных признаков ('Weekday', 'Month', 'Season', 'Hour', 'Functioning Day').

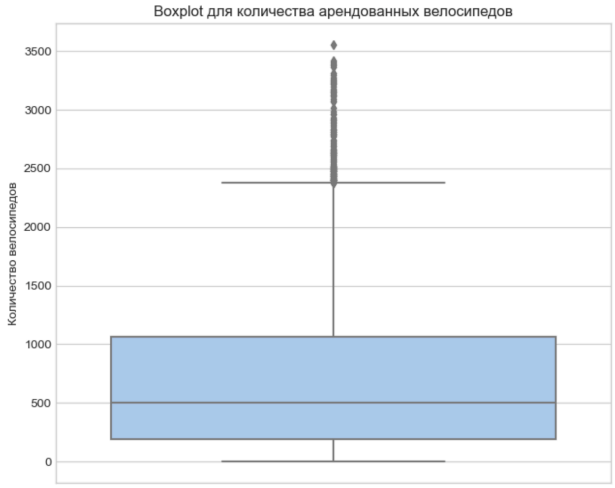
```
preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), numerical_cols_for_scaling),
        ('cat', OneHotEncoder(handle_unknown='ignore', drop='first'), categorical_cols_for_ohe)
    ],
    remainder='passthrough'
)
```

2. Разведочный анализ данных (EDA)

Распределение целевой переменной ('Rented Bike Count'):

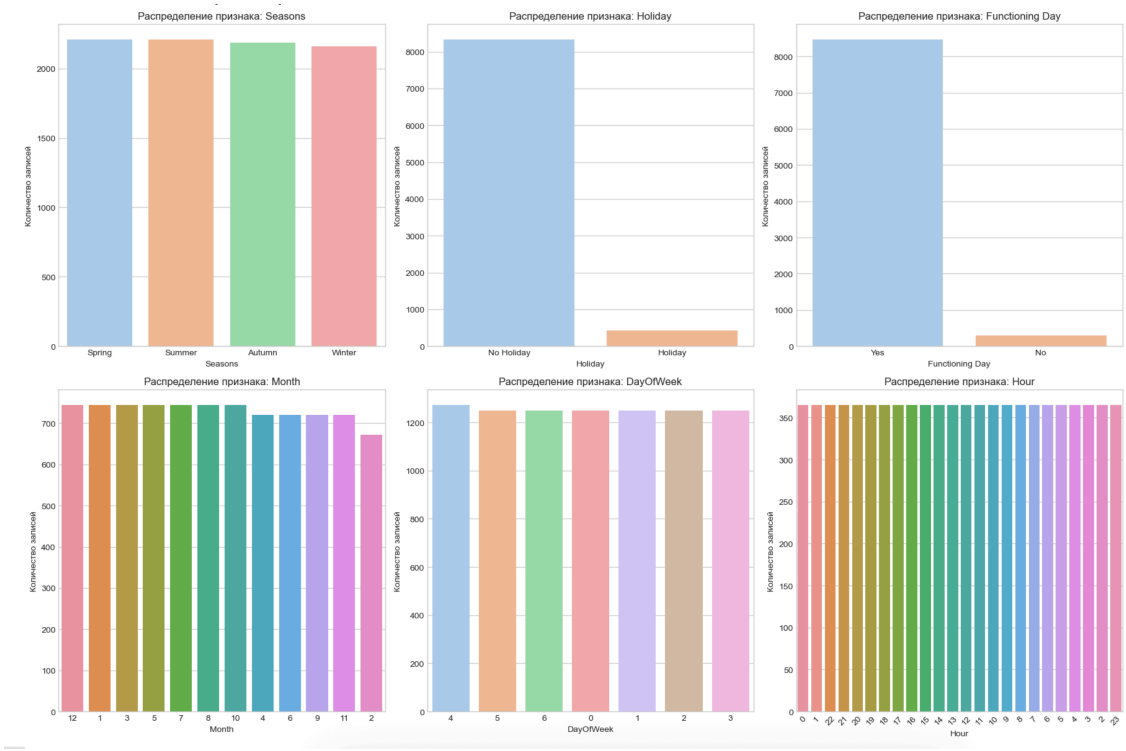
- Построение гистограммы распределения количества арендованных велосипедов.
- Отмечена скошенность распределения, что указывает на необходимость возможного логарифмического преобразования для регрессионных моделей.





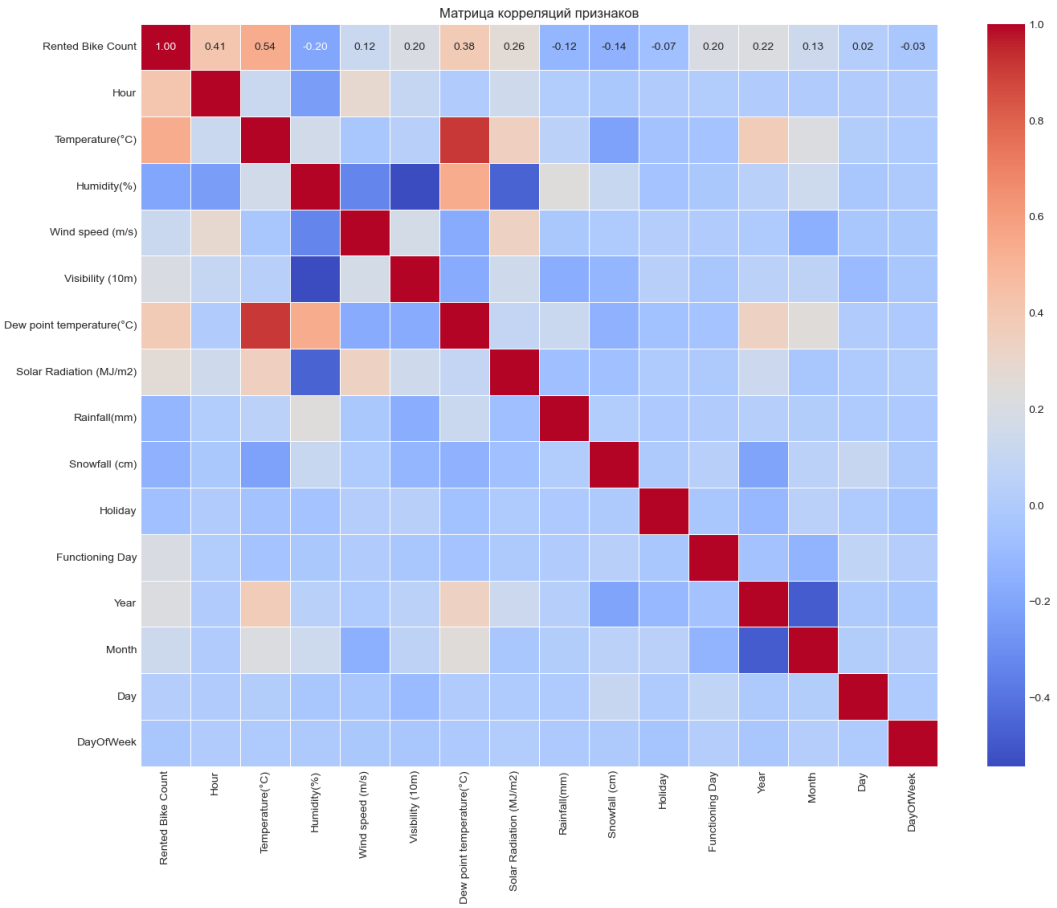
Влияние временных факторов:

- Анализ среднего количества арендованных велосипедов по часам, дням недели, месяцам и сезонам.
- Выявлена сильная зависимость от часа (пики в утренние и вечерние часы), сезона (лето и весна - высокий спрос).



Влияние погодных условий:

- Анализ корреляции между погодными условиями (температура, влажность, скорость ветра и т.д.) и количеством арендованных велосипедов.
- Выявлена мультиколлинеарность между температурой и точкой росы.
- Графики зависимости спроса от температуры и влажности.


Матрица корреляций признаков
Rented Bike Count
Hour
Temperature(°C)
Humidity(%)
Wind speed (m/s)
Visibility (10m)
Dew point temperature(°C)
Solar Radiation (MJ/m2)
Rainfall(mm)
Snowfall (cm)
Holiday
Functioning Day
Year
Month
Day
DayOfWeek
1.00 0.41 0.54 -0.20 0.12 0.20 0.38 0.26 -0.12 -0.14 -0.07 0.20 0.22 0.13 0.02 -0.03
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

3. Моделирование

Для обеих задач (регрессии и классификации) данные были разделены на обучающую и тестовую выборки (80/20). Использовались различные модели машинного обучения, а также проведена оптимизация гиперпараметров с помощью GridSearchCV и кросс-валидации.

4.1. Задача регрессии: Предсказание точного количества арендованных велосипедов

- **Используемые модели:** Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor.
- **Метрики оценки:** R2 (коэффициент детерминации), MAE (средняя абсолютная ошибка), RMSE (среднеквадратичная ошибка).
- **Результаты:**
  - Наилучшие результаты показали ансамблевые модели: Random Forest Regressor и Gradient Boosting Regressor.
  - После подбора гиперпараметров удалось достичь R2 около 0.87−0.88, MAE в районе 150−160 велосипедов и RMSE около 230−240 велосипедов. Это означает, что модели способны объяснить около 87−88 дисперсии спроса.
- **Важность признаков:**
  - Важность признаков, полученная из RandomForest, подтвердила выводы EDA: час, температура, статус функционирования системы, влажность и сезонность являются наиболее значимыми.

Результаты тюнингованных моделей для задачи регрессии:

	R2	MAE	RMSE
Gradient Boosting Regressor	0.897537	134.192697	206.618078
Random Forest Regressor	0.897266	121.352694	206.890597
Support Vector Regressor (SVR)	0.541563	282.073682	437.042411

4.2. Задача классификации: Предсказание категории спроса (Низкий/Средний/Высокий)

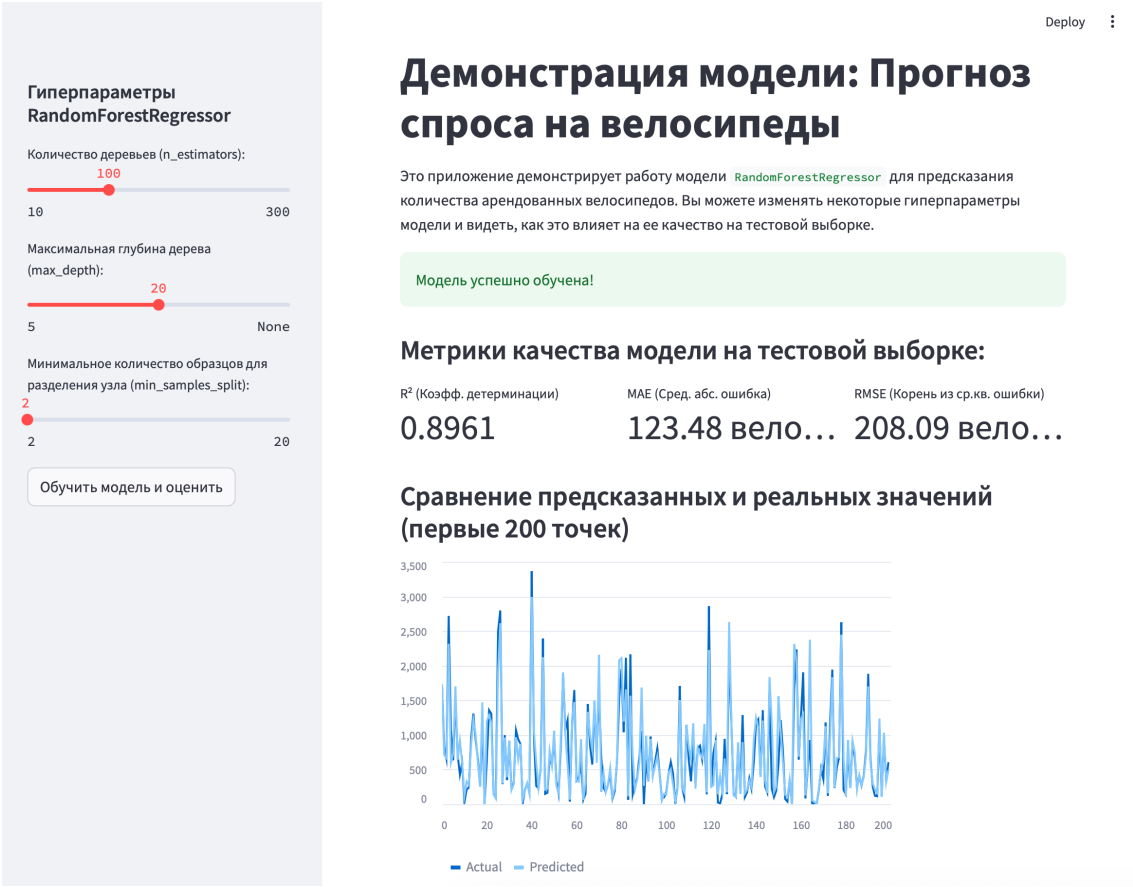
- **Создание целевой переменной:** Количество арендованных велосипедов было разделено на три категории по квантилям (низкий, средний, высокий).
- **Используемые модели:** Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier.
- **Метрики оценки:** F1-score (weighted), Accuracy, Precision, Recall, Confusion Matrix.
- **Результаты:**
  - Ансамблевые модели (Random Forest Classifier, Gradient Boosting Classifier) также показали себя лучше всего.
  - После тюнинга удалось достичь F1-score (weighted) около 0.73−0.74, Accuracy около 0.73.
  - Confusion matrix показала хорошее предсказание основной диагонали, но наблюдались некоторые ошибки между смежными классами.

Результаты тюнингованных моделей для задачи классификации:

	Accuracy	F1 (weighted)	Precision (weighted)	Recall (weighted)
Gradient Boosting Classifier	0.894406	0.894078	0.894025	0.894406
Support Vector Classifier (SVC)	0.891553	0.891872	0.892572	0.891553
Random Forest Classifier	0.878425	0.87776	0.877727	0.878425

4. Веб-приложение

Реализуем веб-приложение для демонстрации влияния гиперпараметров на точность модели **Случайный лес**. Используем фреймворк **Streamlit**.



ЗАКЛЮЧЕНИЕ

В ходе выполнения научно-исследовательской работы был проведен анализ данных о спросе на аренду велосипедов в Сеуле и построены модели машинного обучения для задач регрессии и классификации.

- Проведен разведочный анализ данных, который выявил ключевые факторы, влияющие на спрос, такие как время суток, температура, влажность и сезонность.
- Для задачи регрессии ансамблевые модели (Random Forest Regressor, Gradient Boosting Regressor) показали высокую точность предсказания.
- Для задачи классификации (разделение спроса на категории) ансамблевые классификаторы (Random Forest Classifier, Gradient Boosting Classifier) также продемонстрировали наилучшие результаты.

Результаты исследования могут быть использованы для оптимизации работы системы велопроката, прогнозирования нагрузки и принятия управленческих решений.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Credit Score Prediction [Электронный ресурс] // github.com. URL: https://github.com/ongaunjie1/credit-score-prediction (дата обращения: 02.05.2025);

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».