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 ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ

КАФЕДРА СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

Предсказательная модель числа арендованных велосипедов

Студент ИУ5-63Б _____ **Кузнецов В.А.**
(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.

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

по теме _____Предсказательная модель числа арендованных велосипедов_____
_____

Студент группы _ИУ5-63Б__

_____Кузнецов Владислав Алексеевич_____
(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 _____
_____Исследовать методы машинного обучения для решения задачи регрессии_____
_____
_____
_____

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на _19_ 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 _____ _____Кузнецов В.А._____
(Подпись, дата) (И.О.Фамилия)

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

Оглавление

Введение

Концепция развития велосипедного проката набирает популярность по всему миру, предлагая удобное и экологически чистое средство передвижения. Однако одна из ключевых задач в управлении системой аренды велосипедов — это предсказание числа арендованных велосипедов в разные временные периоды. Точное предсказание позволяет оптимально распределять ресурсы, обеспечивать доступность велосипедов и снижать эксплуатационные издержки.

В данной работе мы будем использовать данные об аренде велосипедов, включающие временные параметры, погодные условия, данные о праздничных днях и другие факторы, чтобы построить модель машинного обучения, способную предсказывать число арендованных велосипедов. Мы применим алгоритмы регрессии для определения основных факторов, влияющих на спрос на велосипеды, и создания точных предсказаний.

Целью данной работы является разработка эффективной модели, которая может помочь операторам велосипедных прокатов быстро и точно прогнозировать спрос, что позволит улучшить качество обслуживания пользователей и повысить эффективность работы системы проката.

Для достижения поставленной цели были определены следующие этапы:

1. Поиск и выбор набора данных для построения моделей машинного обучения для решения задачи регрессии или классификации.
2. Проведение разведочного анализа данных.
3. Выбор признаков, подходящих для построения моделей.
4. Кодирование категориальных признаков. Масштабирование данных. Формирование вспомогательных признаков, улучшающих качество моделей.
5. Проведение корреляционного анализа данных. Формирование промежуточных выводов о возможности построения моделей машинного обучения.
6. Выбор метрик для последующей оценки качества моделей.

7. Выбор наиболее подходящих моделей для решения задачи классификации или регрессии.
8. Формирование обучающей и тестовой выборок на основе исходного набора данных.
9. Построение базового решения (baseline) для выбранных моделей без подбора гиперпараметров и оценка качества моделей на основе тестовой выборки.
10. Подбор гиперпараметров для выбранных моделей. Построение оптимальных моделей.
11. Формирование выводов о качестве построенных моделей на основе выбранных метрик.

Постановка задачи

Данная работа по машинному обучению направлена на решение задачи регрессии, а именно, предсказание числа арендованных велосипедов в определенные временные периоды. Имеются данные об аренде велосипедов, которые включают информацию о таких факторах, как временные параметры, погодные условия, праздничные и выходные дни, а также сезонные особенности.

Имеются данные об аренде велосипедов, которые включают информацию о таких факторах, как время аренды, количество арендованных велосипедов, действительная температура, температура по ощущениям, влажность, скорость ветра, категория погоды, наличие праздничного дня, выходной день, и время года.

Целью задачи является создание модели машинного обучения, которая будет использовать имеющиеся данные для предсказания числа арендованных велосипедов. Для этого мы будем использовать различные алгоритмы регрессии, такие как линейная регрессия, метод опорных векторов, градиентный бустинг, бэггинг и дерево решений. Модель должна обучаться на тренировочных данных и проверяться на тестовых данных для оценки её точности и эффективности.

Результатом работы должна быть модель, которая сможет точно предсказывать количество арендованных велосипедов в зависимости от временных и погодных условий, а также других факторов. Это поможет операторам велосипедных прокатов оптимизировать распределение велосипедов и улучшить обслуживание клиентов.

Выполнение работы

Для решения задачи регрессии был выбран набор данных содержащий информацию о поездках.

В наборе данных присутствуют следующие столбцы:

- "timestamp" - время
- "cnt" - количество арендованных велосипедов
- "t1" - действительная температура в градусай Цельсия
- "t2" - температура по ощущениям в граудсах Цельсия
- "hum" - влажность
- "wind_speed" - скорость ветра в км/ч
- "weather_code" - категория погоды (насколько плохая погода)
- "is_holiday" - является ли день праздничным
- "is_weekend" - является ли день выходным
- "season" - время года: 0-весна ; 1-лето; 2-очень; 3-зима.

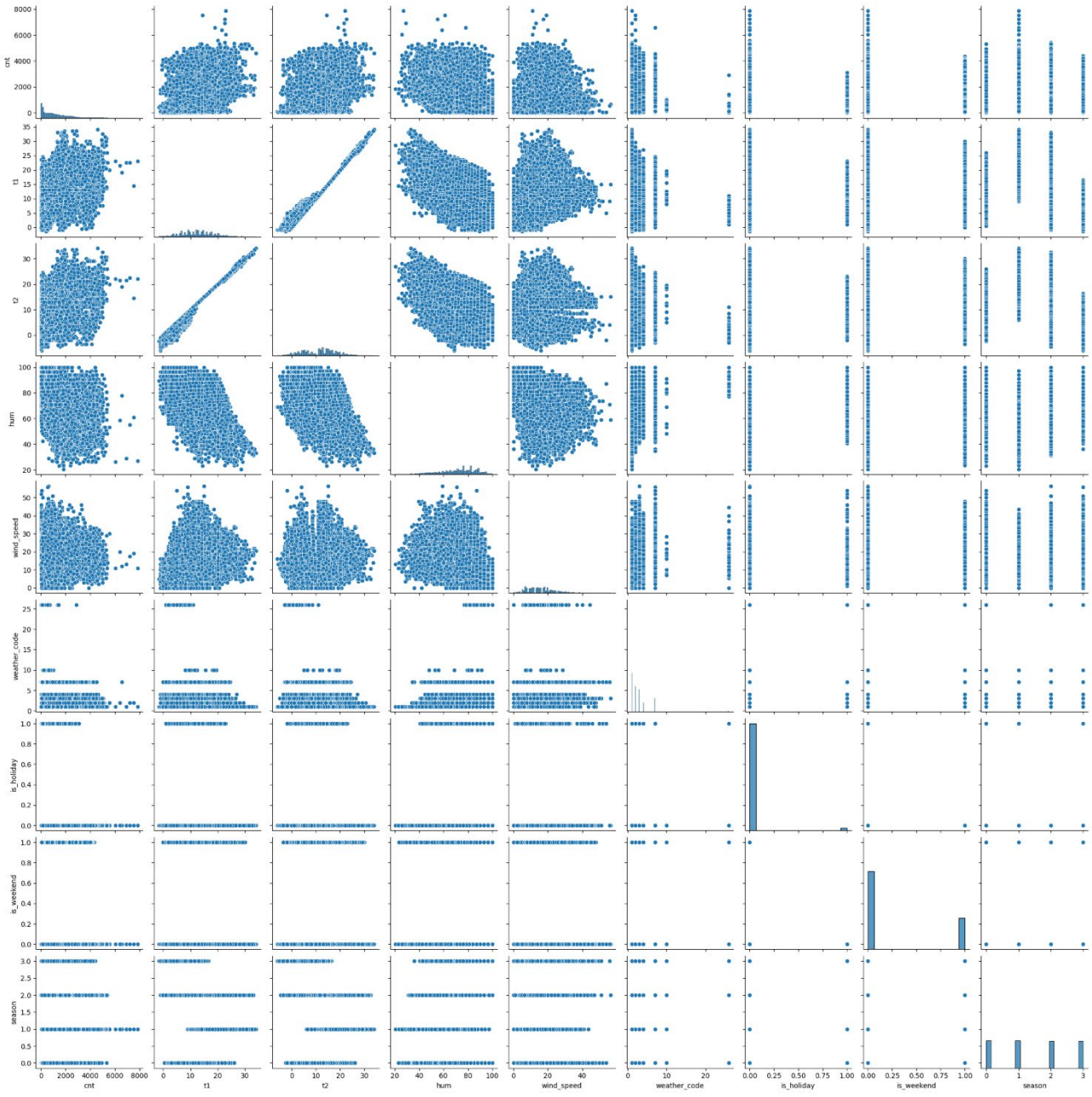
Данный датасет использован для решения задачи регрессии - предсказания числа поездок.

Загружаем данные, получаем общую информацию о датасете и делаем предположения о влиянии признаков на целевую переменную. В наборе данных содержится 17415 строк и 10 столбцов, из которых 8 типа float64, 1 типа int64 и 1 типа object.

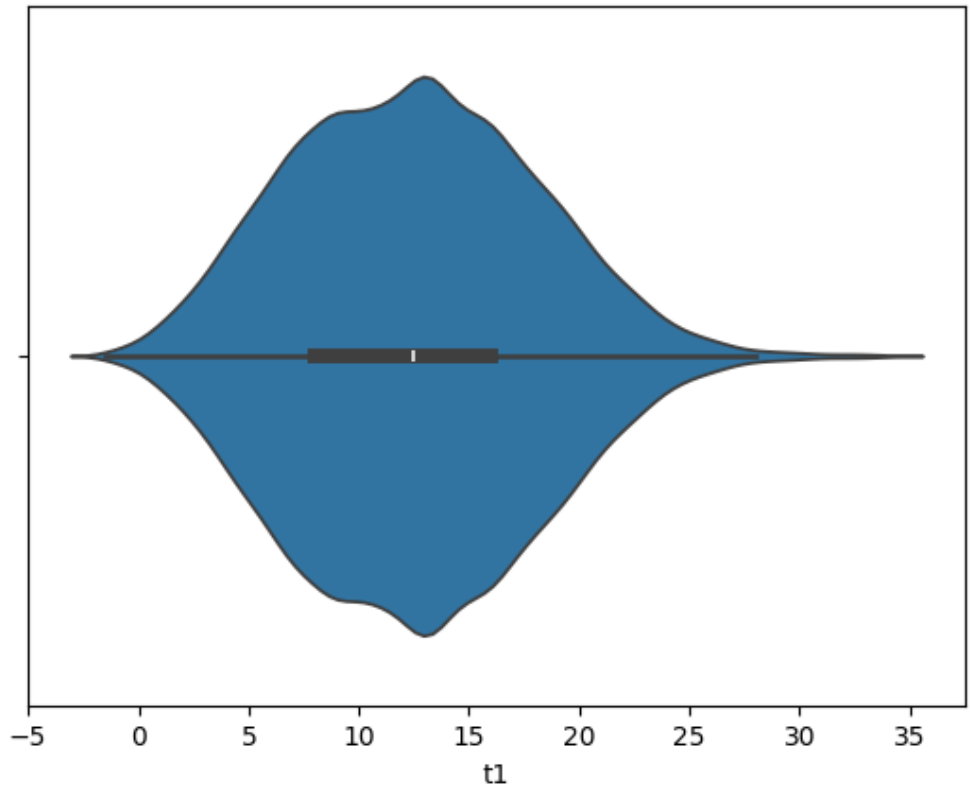
Меняем тип колонки timestamp на ‘datetime, так как она очевидно содержит время.

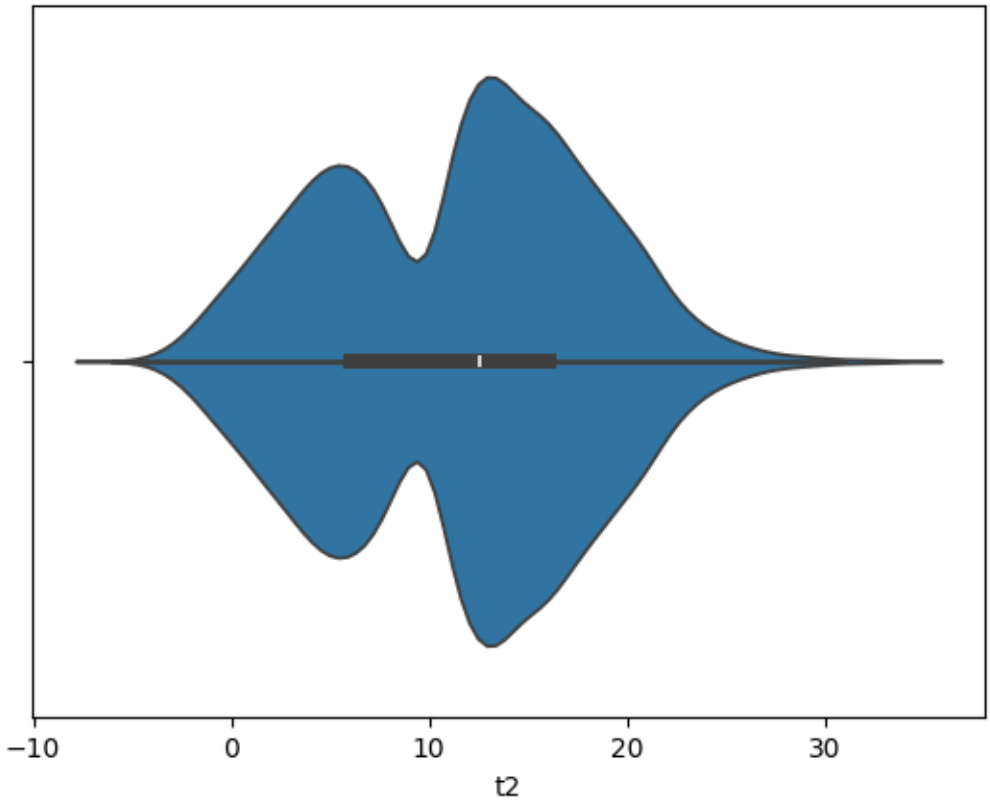
Пропусков не было обнаружено.

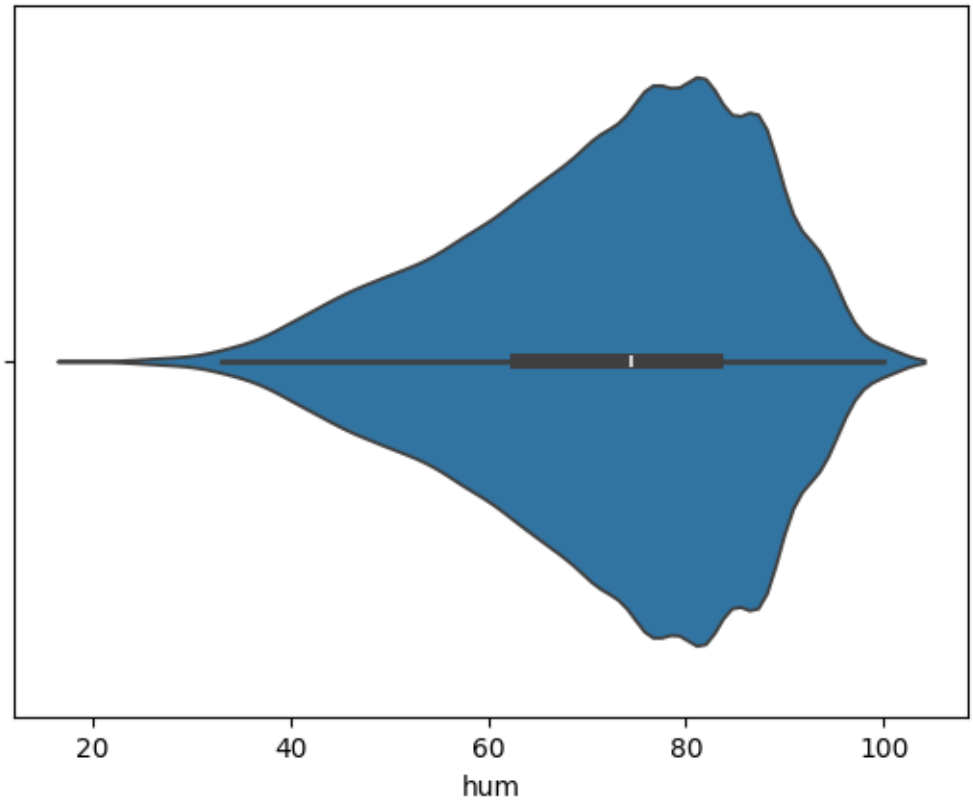
Строим график pairplot для визуализации распределения данных попарно для множества колонок.

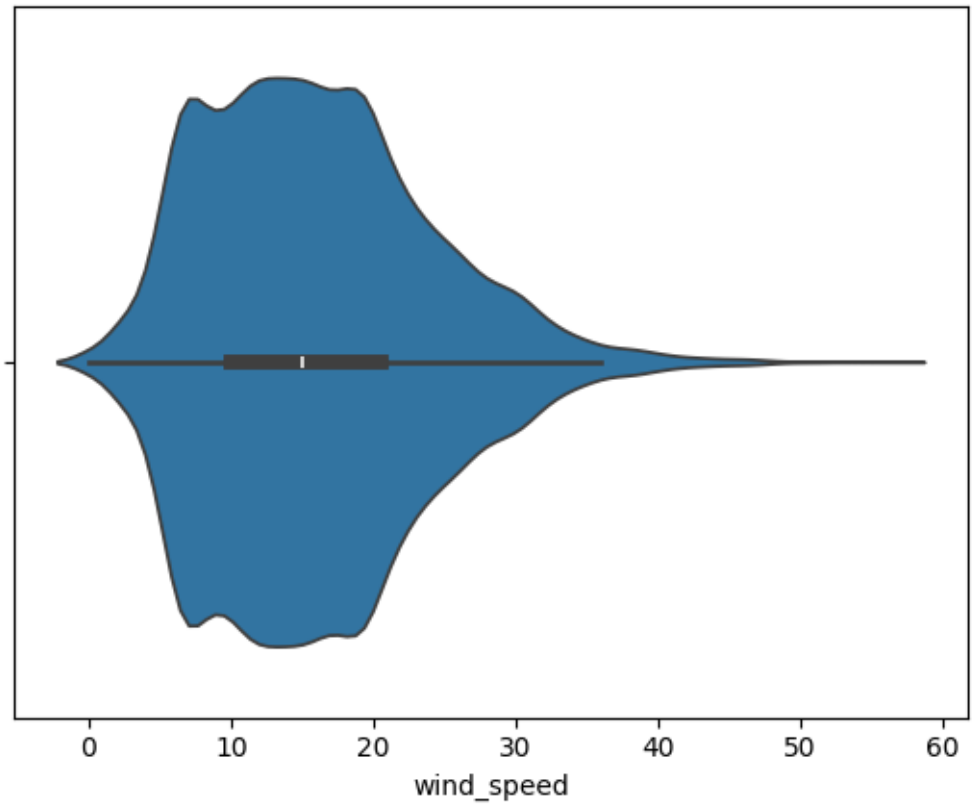

cnt
t1
t2
hum
wind_speed
weather_code
is_holiday
is_weekend
season

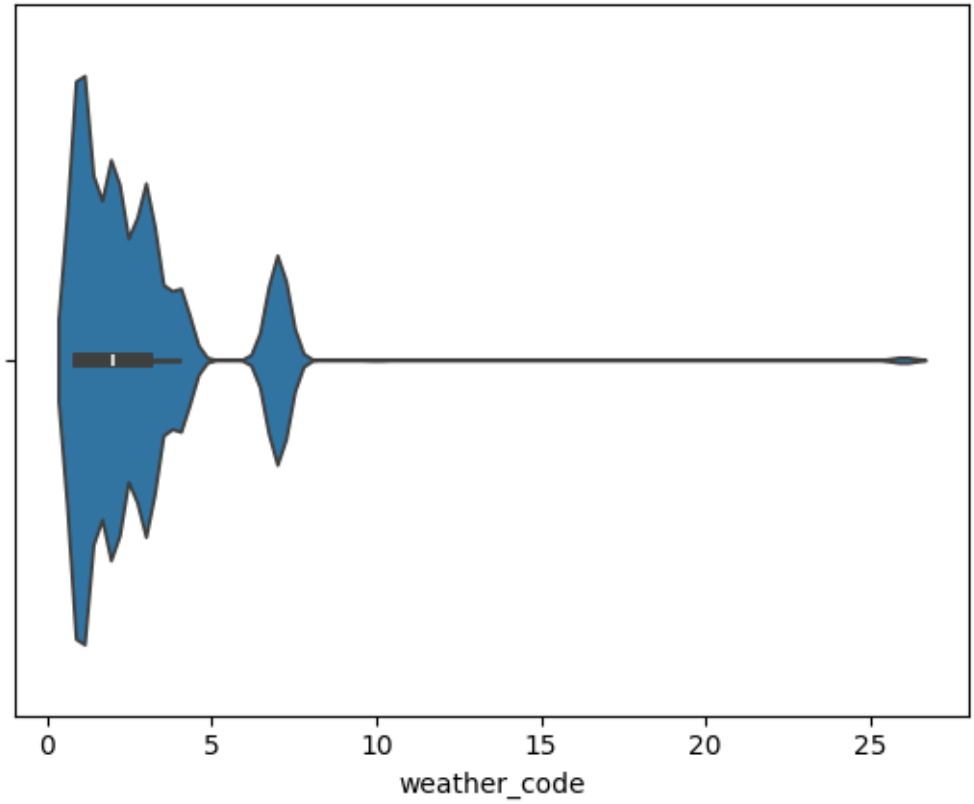
Далее были построены несколько скрипичных-диаграмм, отражающие распределение некоторых признаков:






20
40
60
80
100
hum


0
10
20
30
40
50
60
wind_speed

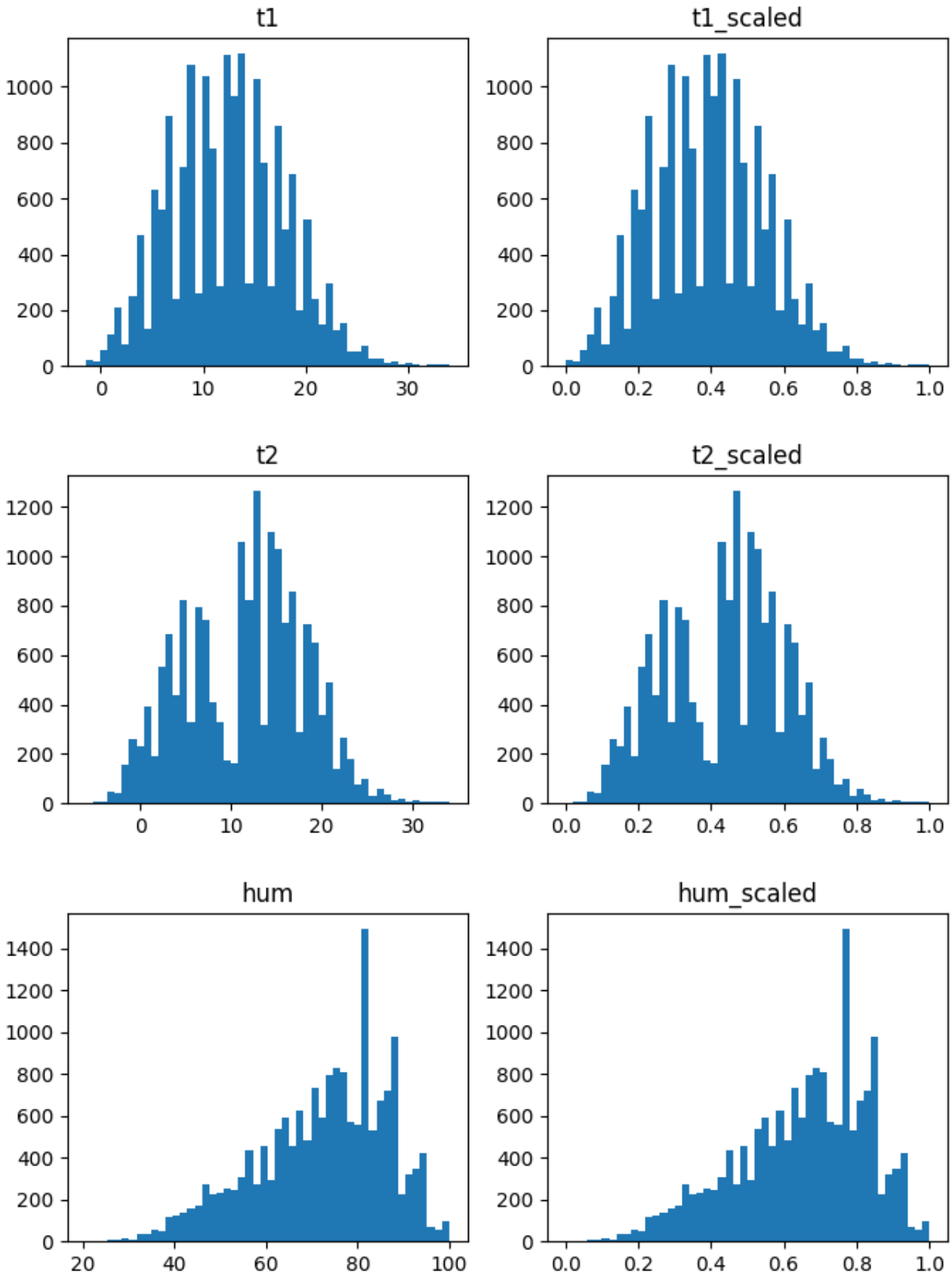


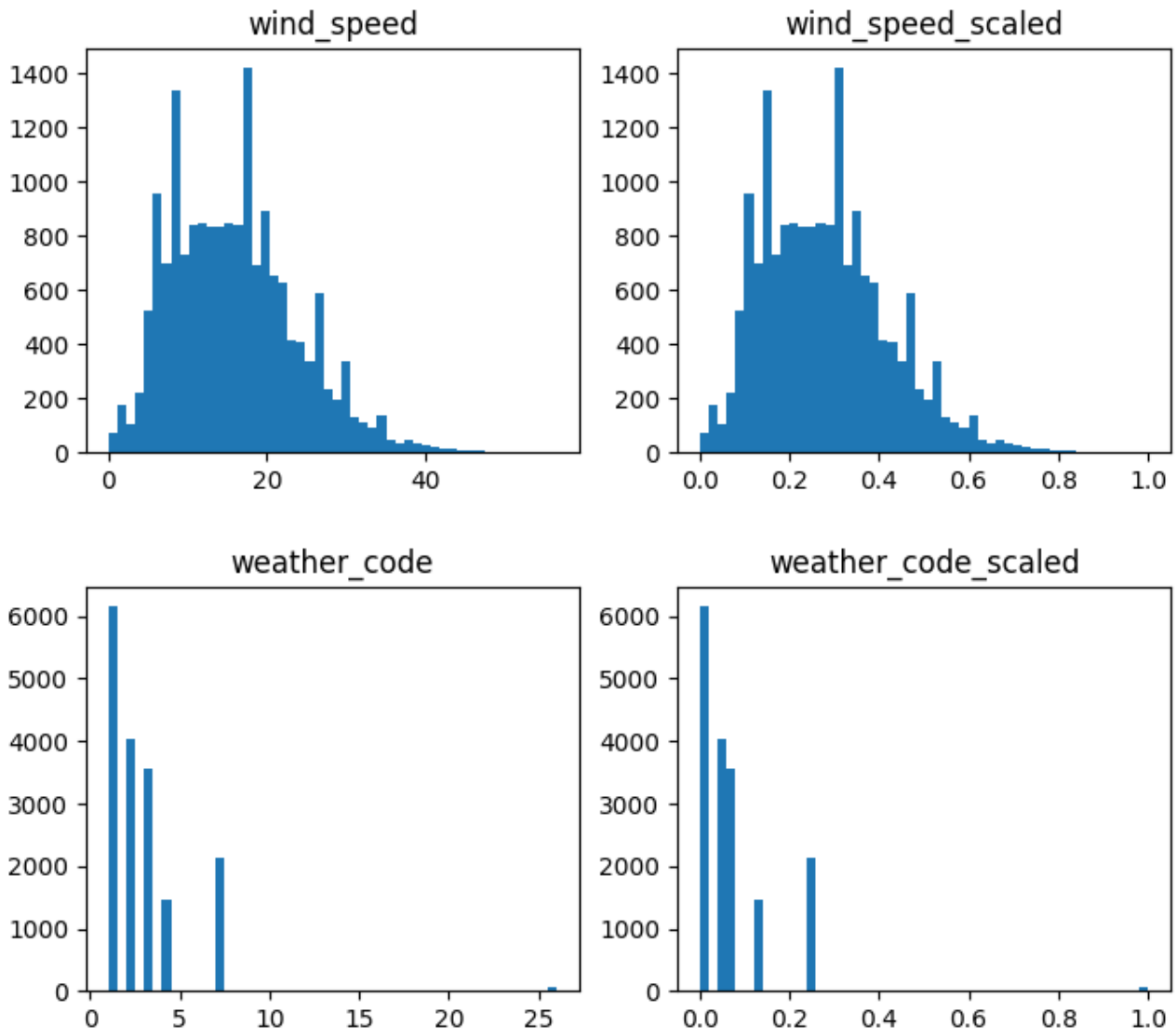
Для построения моделей были использованы все признаки кроме признака timestamp, потому что мы не рассматриваем наши данные как временной ряд.

Категориальные признаки были уже закодированы на основе метода LabelEncoding. Перекодировали признак season при помощи метода OneHotEncoding, чтобы избежать неправильной интерпретации данных, как упорядоченных.

Построил вспомогательный признак времени суток "hour" на основе timestamp.

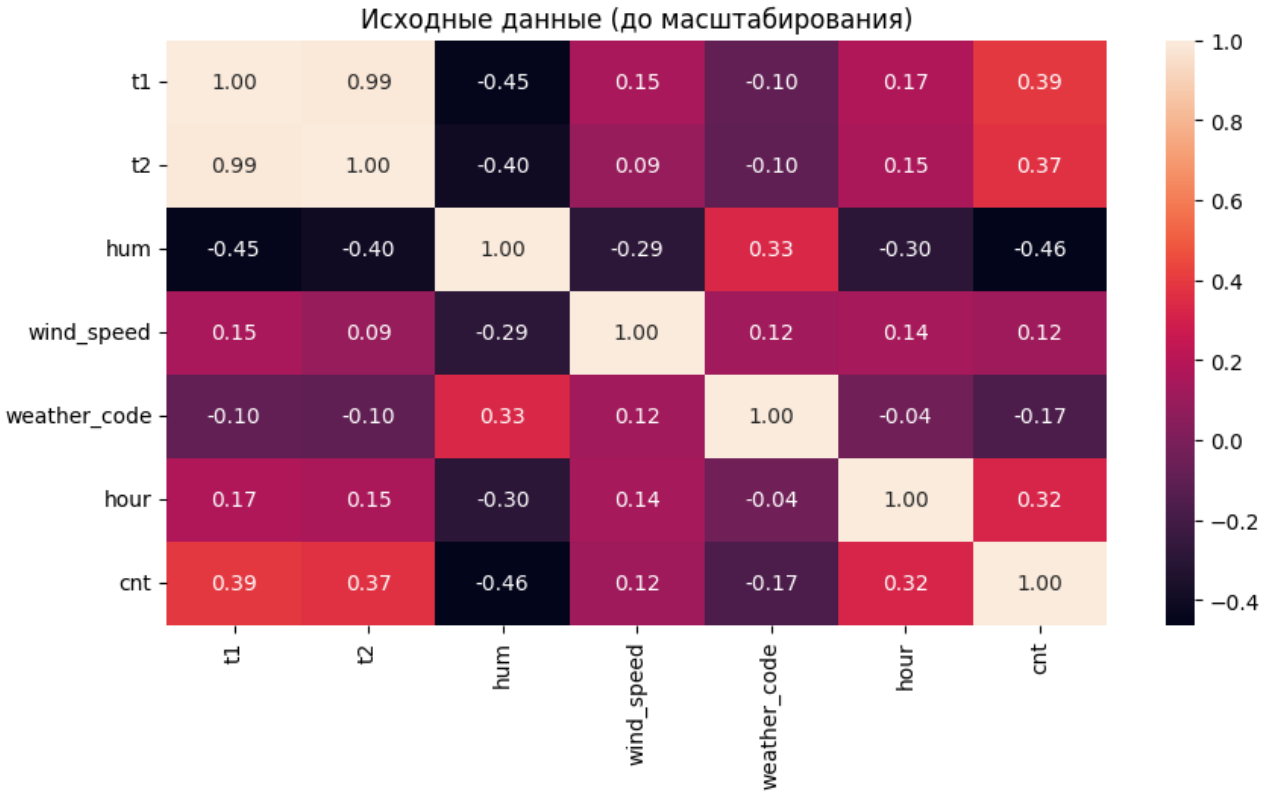
Выполнил масштабирование данных.

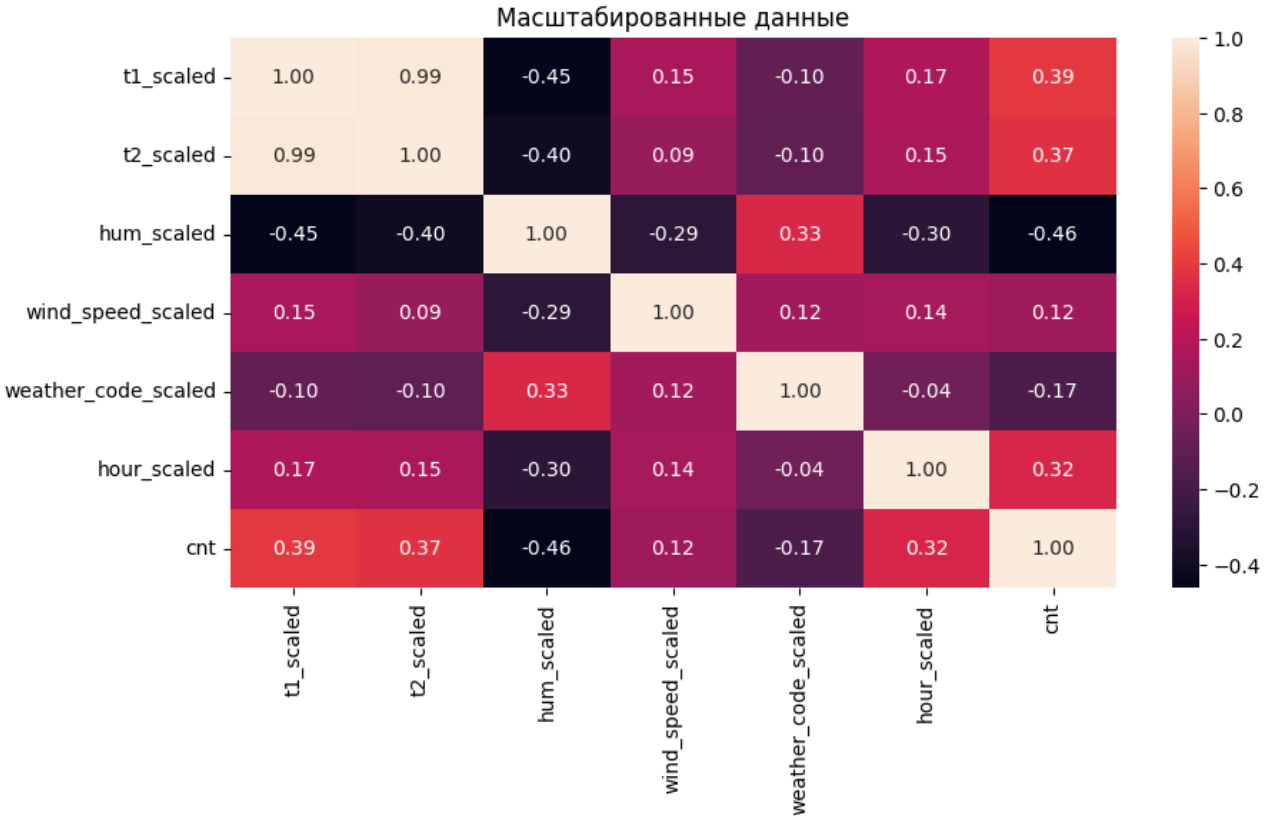

t1
t1_scaled
t2
t2_scaled
hum
hum_scaled



Распределение не изменилось.

Проводим корреляционный анализ данных. Строим тепловую карту корреляций.





На основе корреляционной матрицы можно сделать следующие выводы:

- Корреляционные матрицы для исходных и масштабированных данных совпадают.

- Целевой признак регрессии "cnt" наиболее сильно коррелирует с "hum" (-0.46), температурой (0.39 и 0.37) и временем суток. Эти признаки обязательно следует оставить в модели регрессии.
- Признаки "t1" и "t2" имеют корреляцию, близкую по модулю к 1, поэтому оба признака не следуют включать в модели. Будем использовать признак "t1", так как он лучше чем "t2" коррелирует с остальными признаками.
- 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.

Выберем метрики для оценки качества модели:

- Mean absolute error – средняя абсолютная ошибка.
- Mean squared error – средняя квадратичная ошибка.
- Метрика R2 или коэффициент детерминации.

Выберем модели для решения задачи регрессии:

- Линейная регрессия
- Метод опорных векторов
- Градиентный бустинг
- Бэггинг
- Дерево решений

Формируем обучающую и тестовую выборку в соотношении 80/20.

Строим базовое решения, выводим значения метрик:

```
LR_b        MAE=672.91      MSE=826178.44       R2=0.299
SVC_b       MAE=706.62      MSE=1075450.38      R2=0.088
GB_b        MAE=235.2       MSE=122396.78       R2=0.896
Baggin_b    MAE=149.01      MSE=64463.44        R2=0.945
Tree_b      MAE=197.98      MSE=121027.4        R2=0.897
```

Используем GridSearch для поиска оптимальных гиперпараметров для каждой модели.

Линейная регрессия:

Лучшие параметры: {'fit_intercept': False}

Метод опорных векторов:

Лучшие параметры: {'C': 10, 'degree': 4, 'kernel': 'poly'}

Градиентный бустинг:

Лучшие параметры: {'n_estimators': 200}

Бэггинг

Лучшие параметры: {'n_estimators': 100}

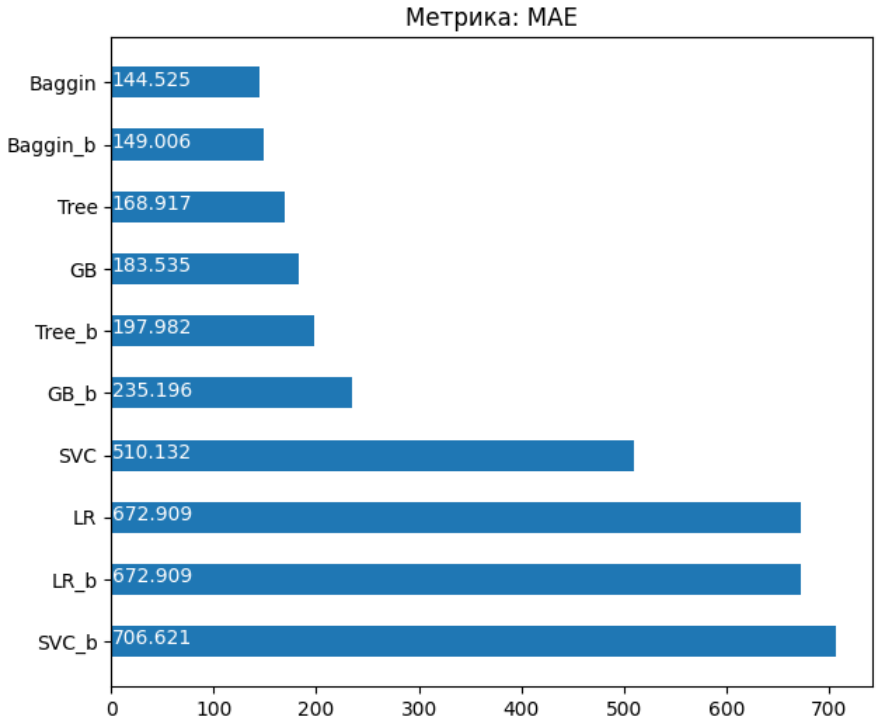
Дерево решений:

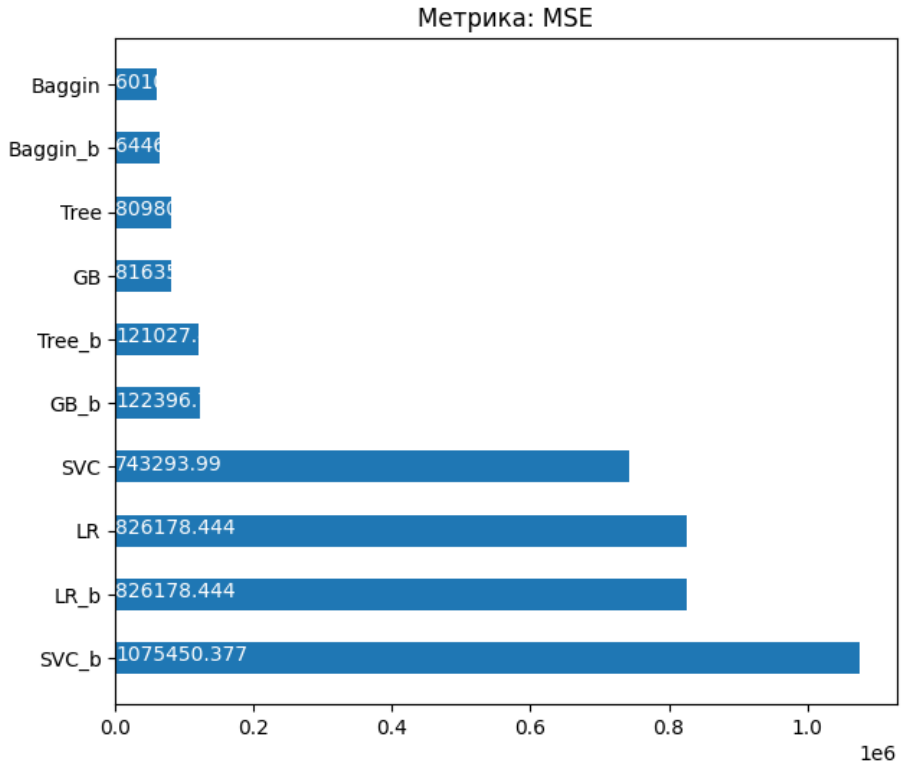
Лучшие параметры: {'criterion': 'poisson', 'max_features': None, 'min_samples_leaf': 5, 'min_samples_split': 20}

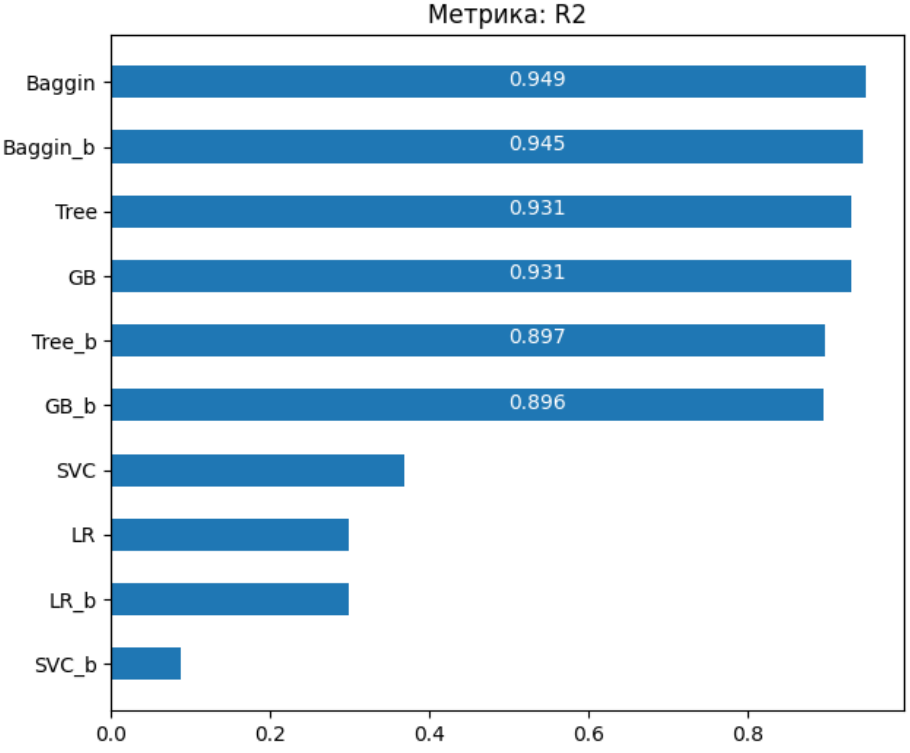
Обучаем и получаем новые значения метрик:

```
LR          MAE=672.91      MSE=826178.44       R2=0.299
SVC         MAE=510.13      MSE=743293.99       R2=0.37
GB          MAE=183.54      MSE=81635.63        R2=0.931
Baggin      MAE=144.53      MSE=60108.67        R2=0.949
Tree        MAE=168.92      MSE=80980.91        R2=0.931
```

Сравним результаты моделей, построив диаграммы для каждой из метрик:




Метрика: MSE
Baggin
Baggin_b
Tree
GB
Tree_b
GB_b
SVC
LR
LR_b
SVC_b
743293.99
826178.444
826178.444
1075450.377
0.0
0.2
0.4
0.6
0.8
1.0
1e6


Метрика: R2
Baggin
Baggin_b
Tree
GB
Tree_b
GB_b
SVC
LR
LR_b
SVC_b
0.949
0.945
0.931
0.931
0.897
0.896
0.0
0.2
0.4
0.6
0.8

Заключение

Предсказание числа арендованных велосипедов с помощью методов машинного обучения является актуальной и перспективной задачей в области управления городским транспортом.

В рамках НИРС была разработана эффективная модель, которая может помочь операторам велосипедных прокатов быстро и с высокой точностью прогнозировать спрос на велосипеды в различные временные периоды. Данные были проанализированы, визуализированы и подготовлены к обучению. Были применены различные алгоритмы, такие как линейная регрессия, метод опорных векторов, дерево решений, бэггинг и градиентный бустинг.

В результате исследования было показано, что только некоторые из использованных методов могут достичь хороших результатов, но самым точным оказался бэггинг на основе дерева решений с подобранными гиперпараметрами.

Список использованной литературы

1. Machine Learning Metrics in simple terms // Medium URL: https://medium.com/analytics-vidhya/machine-learning-metrics-in-simple-terms-d58a9c85f9f6
2. London bike sharing dataset //Kaggle URL: https://www.kaggle.com/datasets/hmavrodiev/london-bike-sharing-dataset?select=london_merged.csv
3. Опорный пример для выполнения проекта по анализу данных. // Jupyter nbviewer URL: https://nbviewer.org/github/ugapanyuk/courses_current/blob/main/notebooks/ml_project_example/project_classification_regression.ipynb
4. Репозиторий курса "Технологии машинного обучения", бакалавриат, 6 семестр. // GitHub URL: https://github.com/ugapanyuk/courses_current/wiki/COURSE_TMO_SPRING_2024/