**МИНОБРНАУКИ РОССИИ**

**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ  
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**

**Кафедра вычислительной техники**

**ОТЧЁТ  
по дисциплине «Машинное обучение»**

| Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| --- | --- |
| Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Гатауллин Р.И. |

**Лабораторная работа № 4  
Тема: «Регрессия»**

Санкт-Петербург 2023

**Оглавление**

[Цель работы 3](#_30j0zll)

[Задачи 3](#_1fob9te)

[Ход работы 4](#_2et92p0)

[Теория 4](#_tyjcwt)

Подготовка данных5

Линейная регрессия 8

LASSO11

Ridge регрессия [1](#_26in1rg)4

[Вывод 1](#_lnxbz9)7

**Цель работы**

Целью данной работы служит получение и закрепление навыков предобработки данных и применения методов машинного обучения для решения задач регрессии на датасете, полученном из исследования производительности ноутбуке.

# **Задачи**

В рамках данной работы необходимо выполнить следующие задачи:

1. Подготовить датасет для обучения моделей: преобразовать данные, отобразить значимые признаки, нормализовать данные.
2. Обучить модели и подобрать параметры тремя способами: линейная регрессия, LASSO, ридж-регрессия.
3. Визуализировать предсказанные значения.

**Ход работы**

# **Теория**

*Линейная регрессия* — это метод статистического моделирования, который позволяет предсказывать зависимую переменную на основе линейной комбинации независимых переменных. Он предполагает, что есть линейная связь между переменными и пытается найти оптимальные коэффициенты для построения линейной модели.

Для обучения модели линейной регрессии используется метод наименьших квадратов, который минимизирует сумму квадратов разностей между фак-тическими и предсказанными значениями зависимой переменной.

*Lasso (Least Absolute Shrinkage and Selection Operator)* — это метод регуляризации линейной регрессии, который вводит штрафы на сумму абсолютных значений коэффициентов, для того чтобы сократить количество участвующих переменных. Это позволяет получить более простую и интерпретируемую модель.

*Ridge (Гребневая регрессия)* — это другой метод регуляризации линейной регрессии, который вводит штрафы на сумму квадратов коэффициентов. Ridge-регрессия помогает справиться с проблемой мультиколлинеарности (когда между независимыми переменными существует высокая корреляция), поскольку она стимулирует коэффициенты быть близкими к нулю.

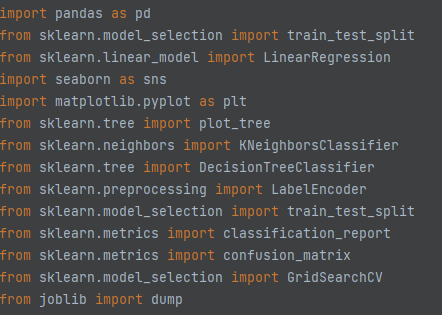
# **Подготовка данных**

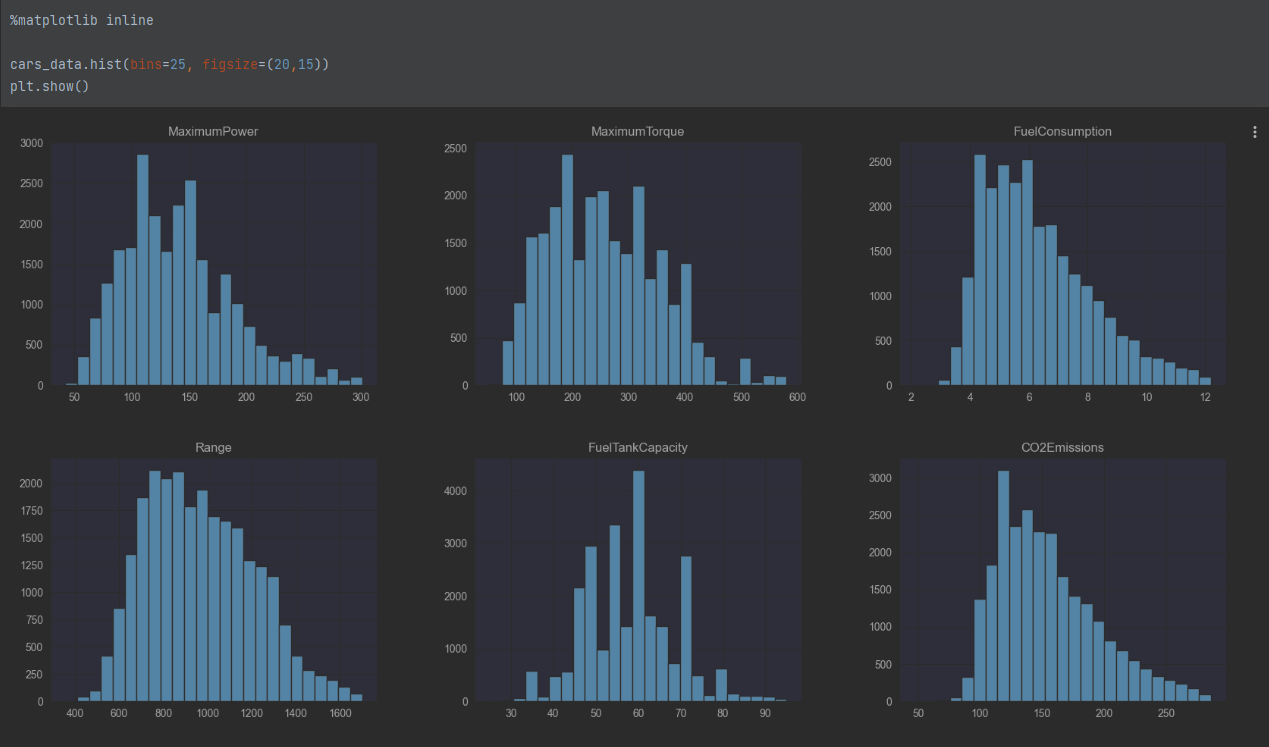
В ходе 1-ой лабораторной работы из исходного набора данных были удалены все дубликаты и пропуски, а также была проведена нормализация исходных данных.

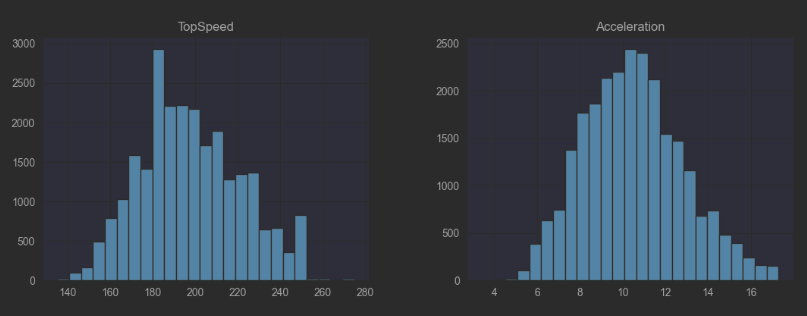
Также в ходе 1 лабораторной работы были построены диаграммы рассеяния, гистограммы и ящики с усами, приведенные ниже, по ним удобно определить какой параметр брать в качестве целевой функции.

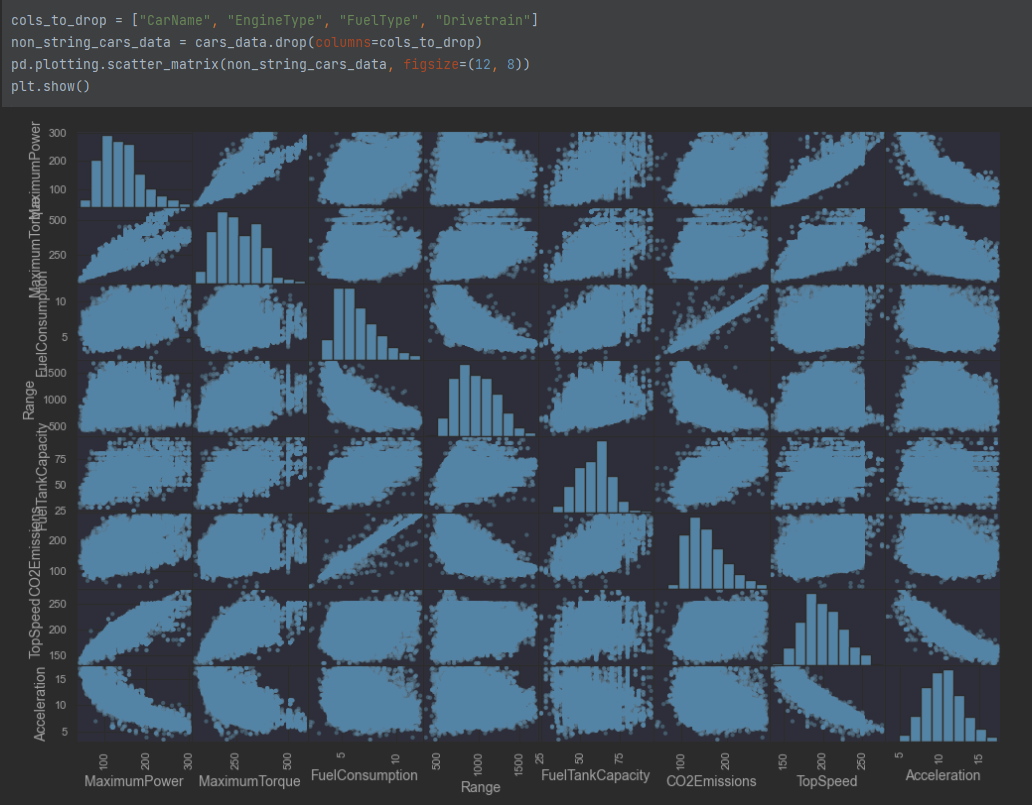
В нашем случае был взят параметр - TopSpeed (Максимальная скорость), так как данный параметр имеет большее кол-во “положительных” зависимостей с другими параметрами.

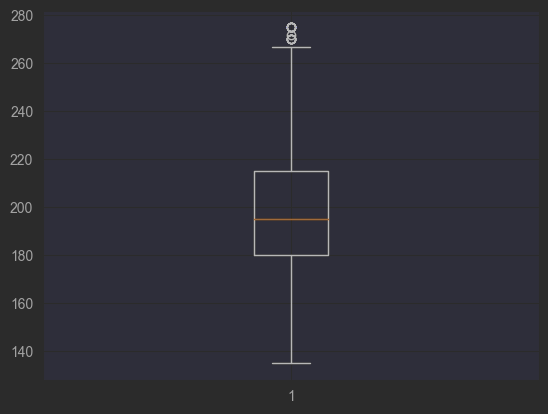
Используемые библиотеки:



Гистограммы исходных данных

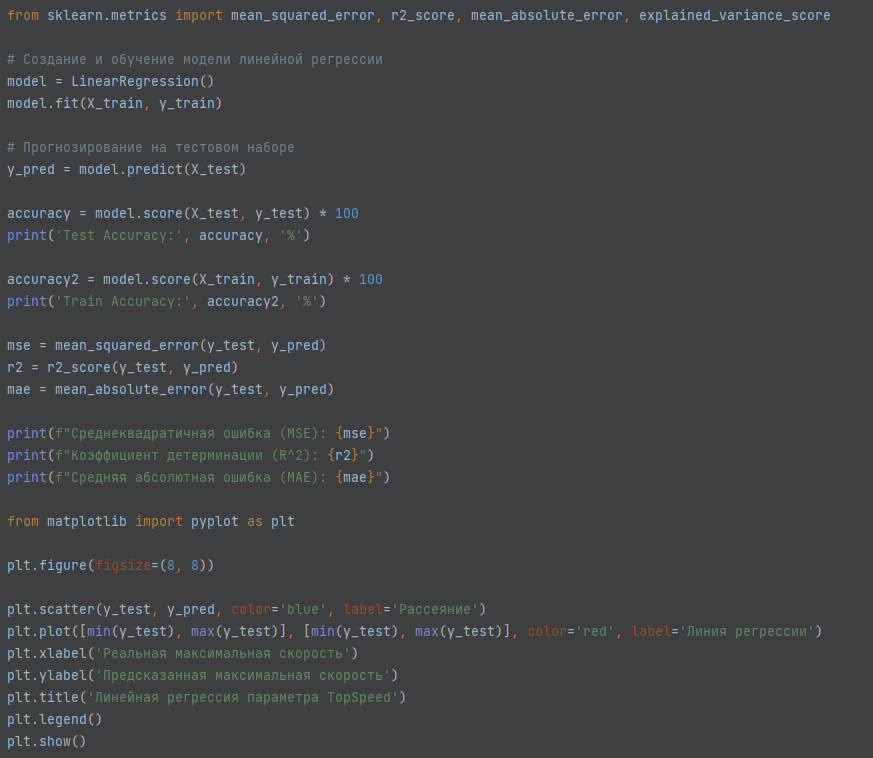


Диаграммы рассеяния

“Ящик с усами” - для параметра TopSpeed

# **Линейная регрессия**

Код метода



Результаты:

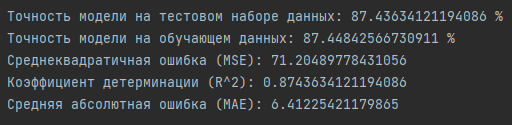
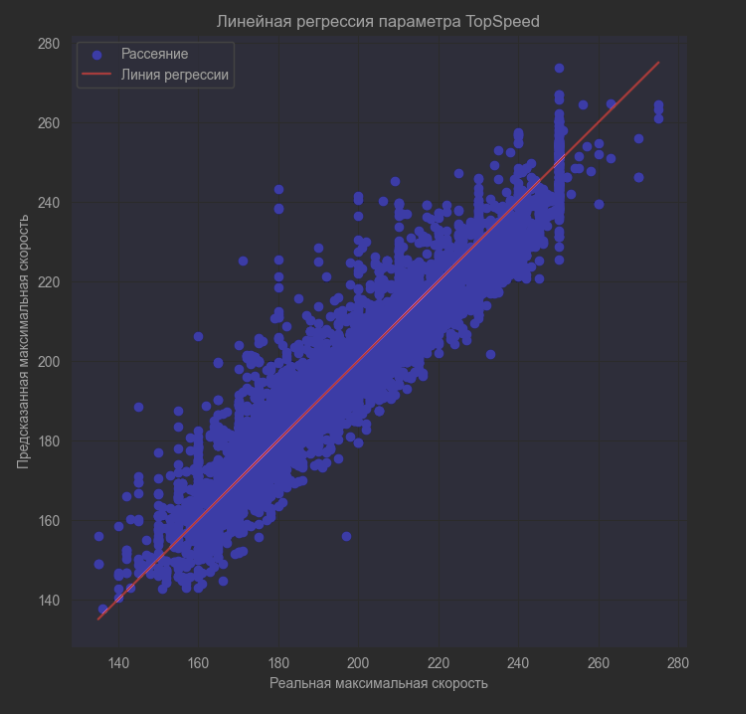


График:



Точность на тестовом наборе: 87,43% — это точность модели на отложенном наборе данных (тестовом наборе). Значение в 87.43% говорит о том, что модель правильно предсказывает зависимую переменную для 87.44% наблюдений в тестовом наборе.

Точность на обучающем наборе: 87,44% — это точность модели на обучающем наборе данных. Значение в 87.44% означает, что модель правильно предсказывает зависимую переменную для 87.44% наблюдений в обучающем наборе.

Значение Mean Squared Error (MSE) указывает на среднеквадратичную ошибку модели в предсказании зависимой переменной. Чем ниже значение MSE, тем ближе предсказанные значения к реальным значениям. В данном случае, значение MSE равно 71.20489778431056, что означает, что средняя квадратичная ошибка предсказания составляет примерно 71.2.

Коэффициент детерминации (R^2) - это мера, которая оценивает, насколько хорошо модель соответствует данным. Значение в 0.8743634121194086 указывает на то, что примерно 87.44% вариации зависимой переменной может быть объяснено моделью. Чем ближе коэффициент детерминации к единице, тем лучше соответствие модели данным.

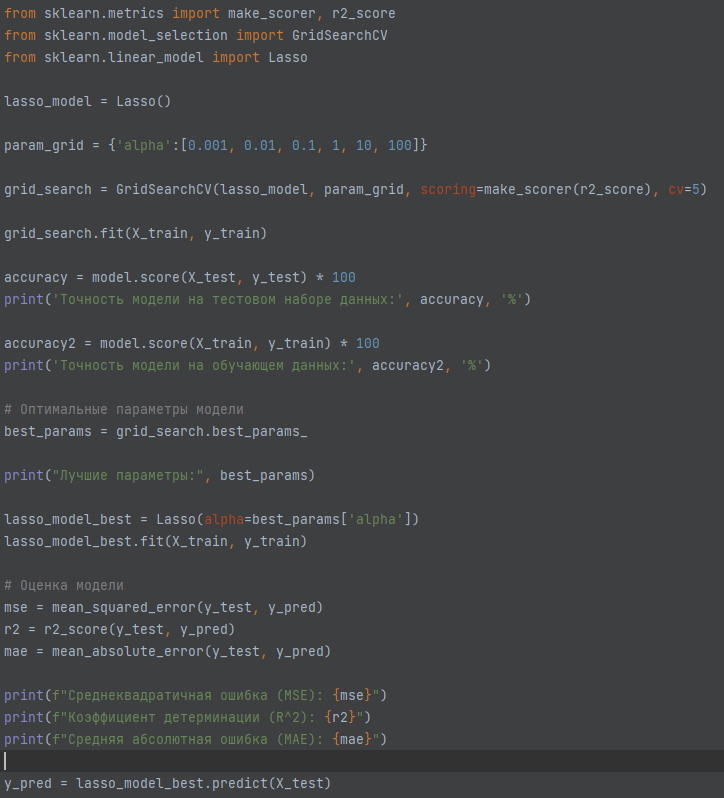
Средняя абсолютная ошибка (MAE): 0.874 — Это среднее абсолютное значение ошибки между предсказанными и фактическими значениями. Меньшее значение MAE также указывает на более точные предсказания модели.

Сходство параметров – *Точность на тестовом наборе*, Т*очность на обучающем наборе* и *Коэффициент детерминации* указывает на то, что модель хорошо обобщает данные и демонстрирует неплохую предсказательную способность.

Модель имеет довольно высокую точность как на обучающем, так и на тестовом наборе данных.

# **LASSO**

Код метода



Результаты:

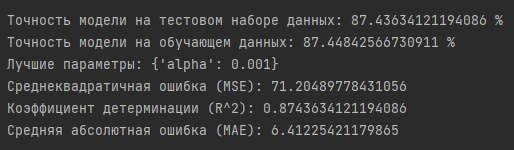
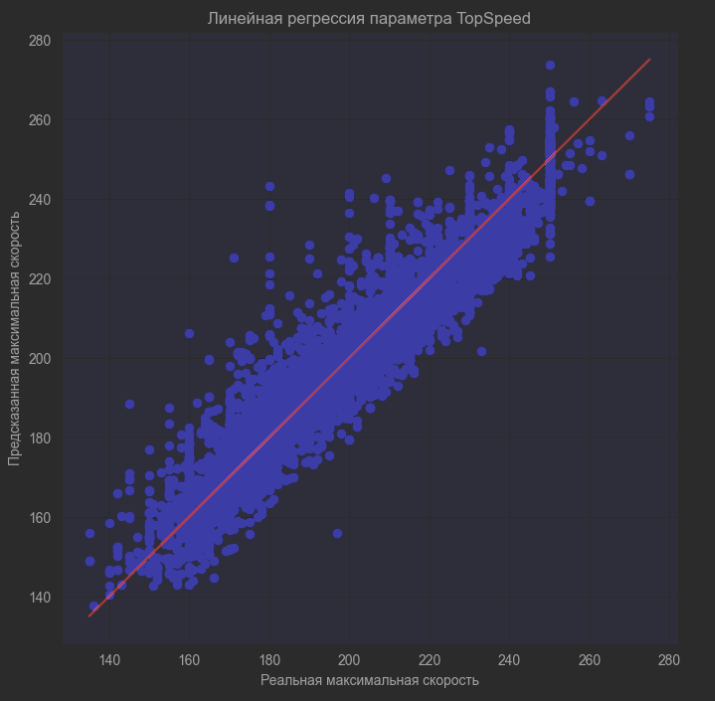


График:



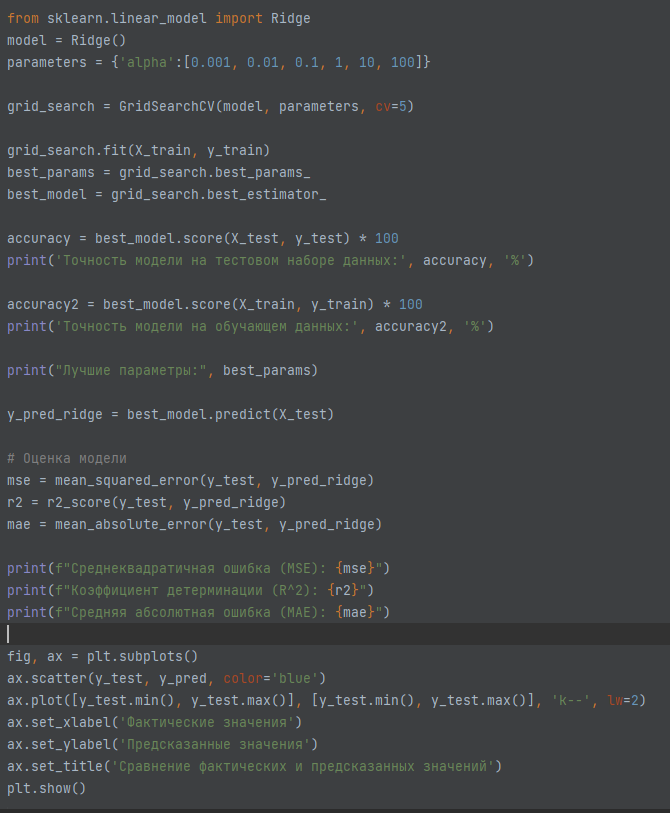
При подборе параметров методом GridSearch коэффициент альфа (α), играющий ключевую роль в контроле степени регуляризации модели и использующийся для управления сложностью модели и предотвращения переобучения, был выбран 0.001. При таком подборе данного параметра точность модели достигается максимальной и схожа с линейной регрессий из первого пункта. Коэффициент альфа (α) в методе Lasso определяет силу регуляризации и влияет на количество признаков, которые будут отобраны моделью. Выбор оптимального значения альфа требует баланса между сложностью модели и ее предсказательной способностьюКогда альфа (α) установлено в ноль, Lasso регрессия стремится минимизировать только сумму квадратов ошибок модели, без применения штрафа на величину коэффициентов. В результате, Lasso регрессия в этом случае выбирает только наиболее важные и информативные признаки, назначая им ненулевые коэффициенты, а остальные признаки получают нулевые коэффициенты.

Если установить значение альфа равным нулю в методе Lasso регрессии, мы получим модель, которая будет выбирать все доступные признаки без ограничений. Это может быть наилучший выбор, когда мы хотим сохранить все признаки и обеспечить высокую предсказательную способность модели.

Модель имеет довольно высокую точность как на обучающем, так и на тестовом наборе данных.

**Ridge регрессия**

Код метода:



Результаты:

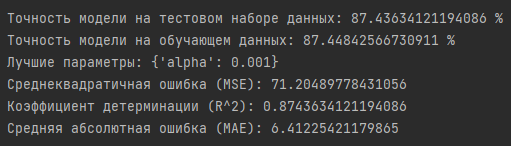
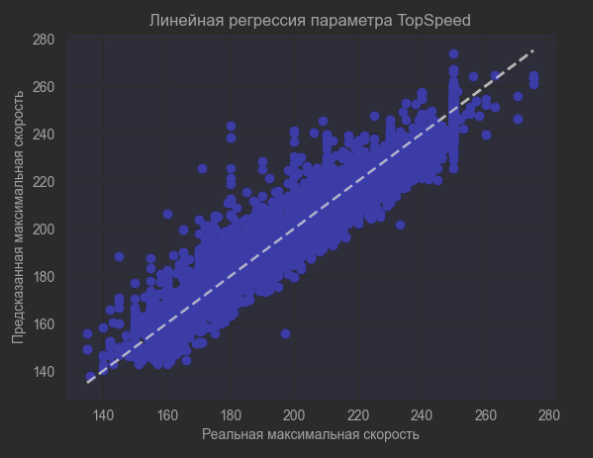


График:



Ridge регрессия - это метод линейной регрессии, который включает регуляризацию с помощью L2 нормы коэффициентов. Регуляризация служит для управления сложностью модели и предотвращения переобучения.

Точность модели на тестовом наборе данных: 87.43681536963264% - это показатель, который указывает на качество предсказаний модели на новых данных, которые она не видела во время обучения.

Точность модели на обучающем наборе данных: 87.44841767959038% - это показатель, который указывает на точность модели на данных, которые она использовала для обучения. Высокая точность на обучающем наборе данных может означать, что модель хорошо соответствует этим данным.

В случае Ridge регрессии, альфа (α) - это гиперпараметр, который управляет силой регуляризации. Значение альфа (α) = 10 указывает на средний уровень регуляризации.

Среднеквадратичная ошибка (MSE): 71.20221048194536 - это средняя квадратичная ошибка между предсказанными значениями модели и фактическими значениями. Чем ниже значение MSE, тем лучше модель соответствует данным.

Коэффициент детерминации (R^2): 0.8743681536963264 - это мера, которая показывает, как хорошо модель объясняет вариацию в данных. Значение R^2 близкое к 1 означает, что модель хорошо объясняет данные.

Средняя абсолютная ошибка (MAE): 6.412155266030508 - это среднее абсолютное отклонение между предсказанными значениями модели и фактическими значениями. Чем ниже значение MAE, тем лучше модель соответствует данным.

Сходство параметров – *Точность на тестовом наборе*, Т*очность на обучающем наборе* и *Коэффициент детерминации* указывает на то, что модель хорошо обобщает данные и демонстрирует неплохую предсказательную способность.

# **Вывод**

В процессе выполнения данной работы освоили навыки предобработки данных и применения методов машинного обучения для задач регрессии.

Результаты исследования показали, что все рассмотренные методы регрессии демонстрируют примерно одинаковую эффективность. Важно отметить, что применение регрессии на данном датасете возможно при условии учета некоторого количества ошибок.

Таким образом, можно сделать вывод, что линейная регрессия, Lasso-регрессия и Ridge-регрессия показывают схожую точность в районе 87%. Это означает, что все эти методы способны примерно на одном уровне соответствовать данным и объяснять около 87% вариации зависимой переменной. Данный результат указывает на то, что исходный набор данных близок к идеалу, имеет минимум шумов и выбросов.

Подводя итог – выбранный массив данных про машины, может быть далее использован в машинном обучении для разных целей, как учебных так и профессиональных, в связи с хорошими показателями по всем параметрам, изученным в ходе курса.