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)»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

по курсу

«Data Science»

Слушатель Воронин Вадим Николаевич

Москва, 2024

# Содержание

[Содержание 2](#_Toc175566188)

[Введение 3](#_Toc175566189)

[1 Аналитическая часть 4](#_Toc175566190)

[1.1 Постановка задачи 4](#_Toc175566191)

[1.2 Описание используемых методов 6](#_Toc175566192)

[1.3 Разведочный анализ данных 7](#_Toc175566193)

[2.1 Предобработка данных 11](#_Toc175566194)

[2 Практическая часть 11](#_Toc175566195)

[2.2 Тестирование модели 11](#_Toc175566196)

[2.3 Разработка и обучение моделей 11](#_Toc175566197)

[2.4 Нейронная сеть 27](#_Toc175566198)

[Заключение 31](#_Toc175566199)

[Библиографический список 32](#_Toc175566200)

# Введение

Данная работа выполнена в рамках курса Data Science.

В качестве анализируемой задачи принята тема «Модель прогнозирования скорости транспортного потока»

# 1 Аналитическая часть

## 1.1 Постановка задачи

В настоящее время транспортные заторы являются одной из основных и глобальных проблем любого крупного города, в частности людей, проживающих в Москве. Решение этого вопроса оказывается актуальным как для отдельного человека, вследствие непроизводительных потерь времени от простоев транспортных средств (ТС), так и для экономики государства в целом. С каждым годом в Москве и Московской области увеличивается число жителей и автомобилей, в свою очередь, это ведет к тому, что нагрузка на транспортную инфраструктуру будет только возрастать.

Одной из причин транспортных заторов является динамичность дорожной обстановки, которая в свою очередь зависит от дня недели, времени суток, погодных условий и других факторов. Однако на сегодняшний день утренние и вечерние «час пики» на дорогах Москвы и примыкающих к ней федеральных трассах являются самыми проблематичными, так как огромный поток транспорта движется утром из области в город, а вечером – в обратном направлении.

Следует отметить и то, что продолжительное нахождение водителей в транспортных заторах отрицательно отражается на состоянии их здоровья из-за возрастающего психоэмоционального напряжения, а также вследствие увеличивающихся автомобильных выбросов в воздухе. К отрицательным последствиям транспортных заторов относится и экономический ущерб, связанный с увеличением расхода топлива, потерями времени в пути. Заторы наносят вред и экологии, так как ведут к возрастанию вредных веществ в атмосфере и повышению уровня шума. Из-за больших скоплений транспортных средств повышается вероятность дорожно-транспортных происшествий (ДТП), оказывается неблагоприятное влияние на комфорт и уровень жизни людей.

Одной из проблем, которая способствует образованию транспортных заторов является и то, что дорожно-транспортной инфраструктуре отведён ограниченный отвод земли (стандартное количество – 15%, а в Москве – меньше 10%). Так, историческая планировка города Москвы создаёт определенные ограничения для развития дорожной сети, поэтому решение вопроса, связанного с образованием транспортных заторов, путём строительства новых дорог невозможно. Тем не менее, новое дорожное строительство не только потребует крупных инвестиций, но и в качестве борьбы с транспортными заторами данная мера не совсем эффективна. Объясняется это тем, что новые дороги в загруженных местах лишь на время сглаживают транспортную обстановку, довольно быстро заполняясь транспортными средствами. Этот эффект носит название «единичная эластичность спроса на дорожную сеть по дорожному строительству». К тому же Москва имеет радиально-кольцевую структуру, что создаёт определенные ограничения для развития дорожной сети и является не совсем удачной для организации движения растущих транспортных потоков в периоды «час пик» на подъездных магистралях.

Уменьшить перегрузку транспортной сети, предотвратив тем самым образование транспортных заторов и их последствий на дорогах федерального значения при въезде в крупные города возможно только при помощи комплекса мероприятий, связанных с совершенствованием управления транспортными потоками – внедрением компьютерных автоматизированных систем управления дорожным движением (АСУДД) на улично-дорожной сети (УДС). Поэтому в условиях современного быстрорастущего трафика автомобильного транспорта применение АСУДД является необходимым условием, поскольку служит инструментом повышения эффективности транспортной системы.

Одним из эффективных методов борьбы с транспортными заторами, является внедрение адаптированного координированного светофорного регулирования.

Входные данные для построения графика координации:

* Геометрические параметры УДС
* Светофорные циклы транспортных узлов
* Средняя скорость транспортного потока (изменяемая переменная)

В рамках данного проекта был определен участок СЗХ – Большая Академическая улица, протяженностью 4,2 км.

На участке находится 7 транспортных узлов со светофорным регулированием. Исходные данные скорости были получены с 8 комплексов фотовидеофиксации за период с 01.07.24 – 21.07.24.

Общее количество параметров для анализа – 4:

* Скорость
* Интенсивность
* День недели фиксации
* Статистика Яндекс пробок

На выходе необходимо спрогнозировать среднюю скорость транспортного потока для Большой Академической улицы, в распределение по часам.

Актуальность: Созданные прогнозные модели на базе собранной статистки помогут получать данные о средней скорости в короткие сроки без использования дорогостоящих детекторов скорости. Так же благодаря точному прогнозу с шагом в 1 км/ч, появляется возможность использования в АСУДД более гибкой адаптации под изменение транспортного потока.

## Описание используемых методов

Для решения задачи регрессии использовались: линейная регрессия, множественная линейная регрессия, опорно-векторная регрессия, регрессия k-ближайших соседей, дерево решений, случайный лес, полиномиальная регрессия, нейросеть с одним входом и нейросеть с несколькими входами.

Оценка качества моделей указана на рисунке 1.

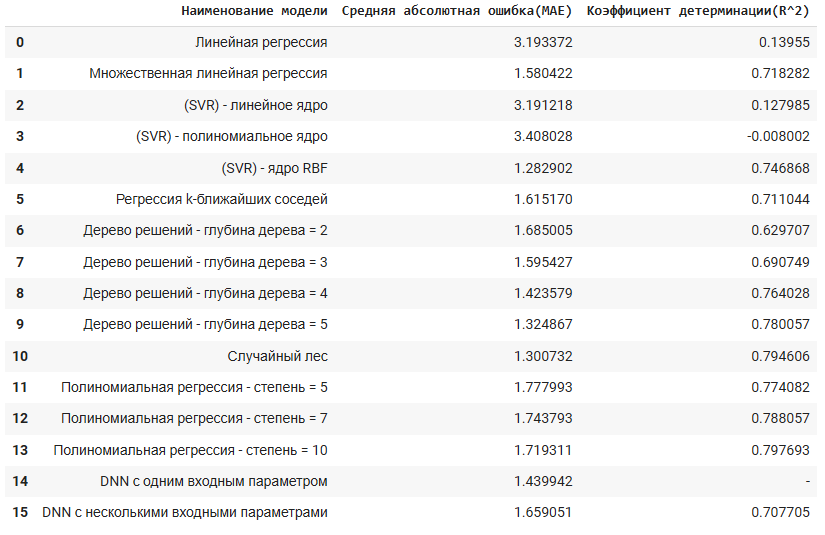


Рисунок 1 - Оценка качества моделей

## 1.3 Разведочный анализ данных

Для разведочного анализа данных использованы методы описательной статистики.

С помощью построения гистограммы было выявлено положительное распределение – распределение со смещением вправо. Гистограмма скорости представлена на рисунке 2.

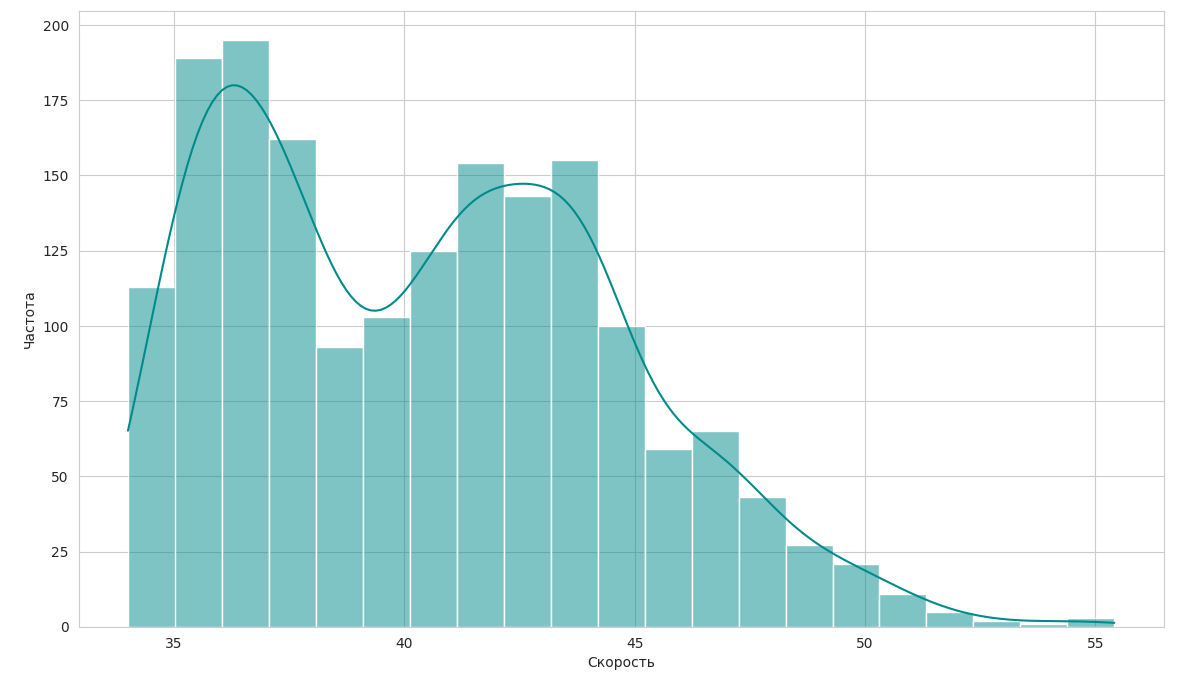


Рисунок 2 – Гистограмма скорости

С помощью диаграммы «ящик с усами» были выявлены выбросы, представленные на рисунке 3.

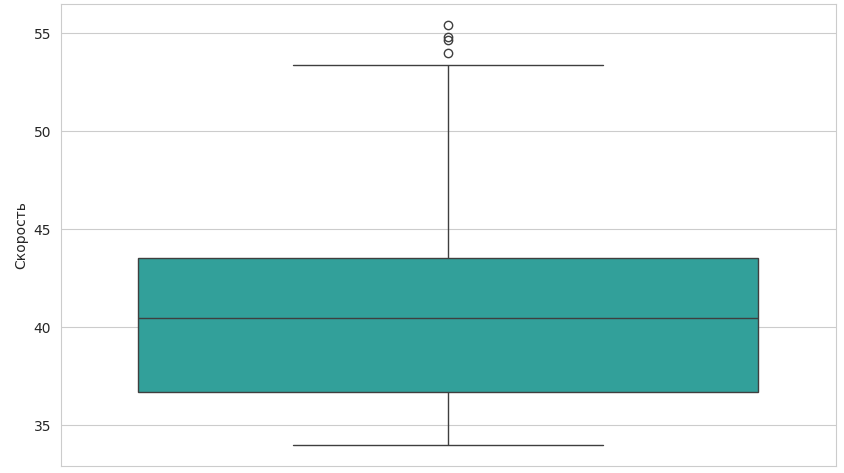


Рисунок 3 – Диаграмма скорости

Для разведочного анализа было использовано также построение графиков рассеяния точек. На рисунке 4 представлены графики рассеяния точек.

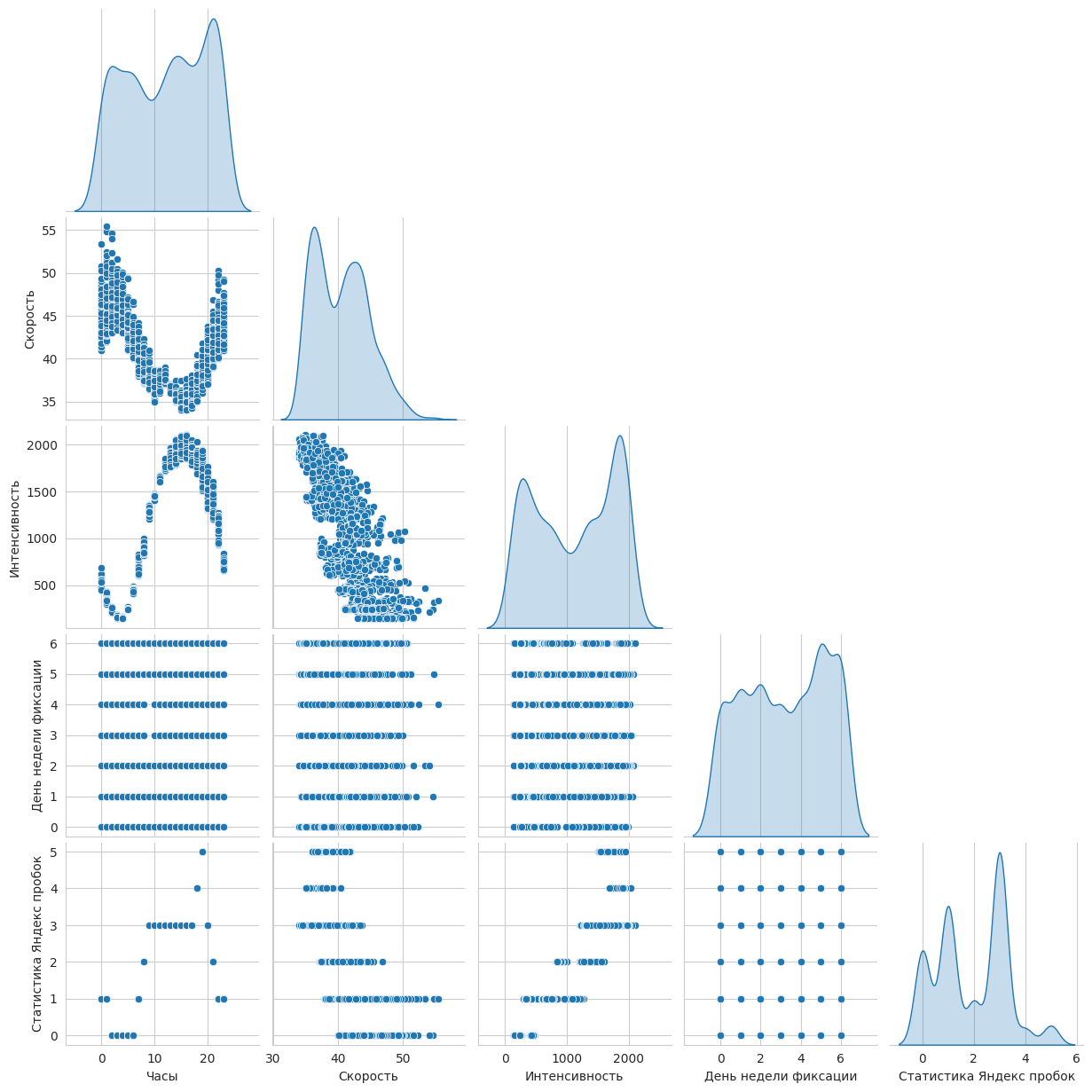


Рисунок 4 – Графики рассеяния точек

На рисунке 5 приведена треугольная тепловая карта корреляции, значения которой показывают, что все полученные коэффициенты корреляции для скорости находятся в промежутке значений от - 0,84 до -0,07.

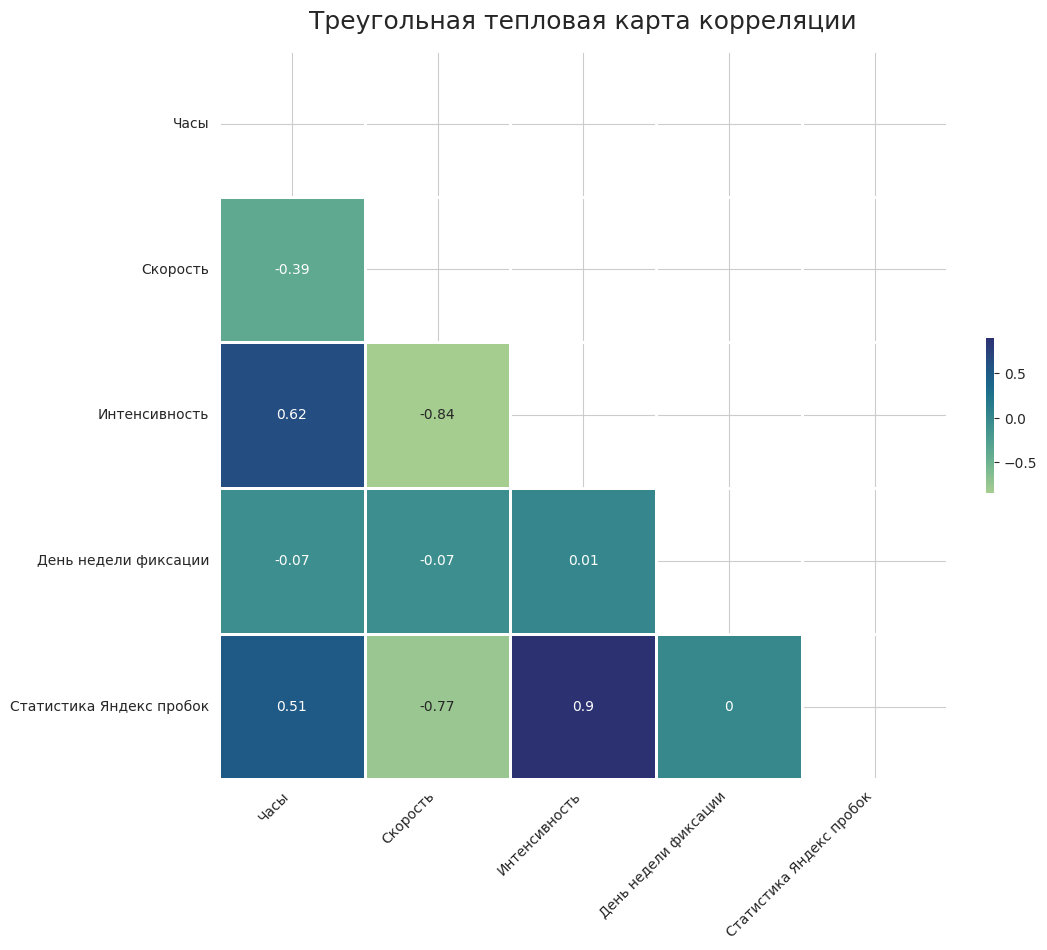


Рисунок 5 – Треугольная тепловая карта корреляции

## 2.1 Предобработка данных

Предобработка данных осуществлялась на основании разведочного анализа данных, который показал наличие незначительного количества выбросов. Для удаления выбросов был произведен расчет количества выбросов для каждого параметра, данные приведены на рисунке 6.

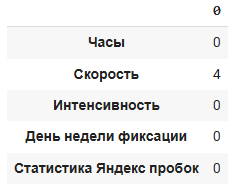


Рисунок 6 – Расчет количества выбросов

Для дальнейшей разработки и обучения модели была выполнена нормализация данных.

# 2 Практическая часть

## 2.2 Тестирование модели

Из датасета была выделена прогнозируемая (зависимая) переменная. Далее, выборка была разделена на обучающую и тестовую выборки, в соответствии с условием задачи 80% (на обучение) /20% (на тестирование).

Размер обучающей выборки: 1412.

Размер тестовой выборки: 353.

## 2.3 Разработка и обучение моделей

Линейная регрессия

График линейной регрессии представлен на рисунке 7.

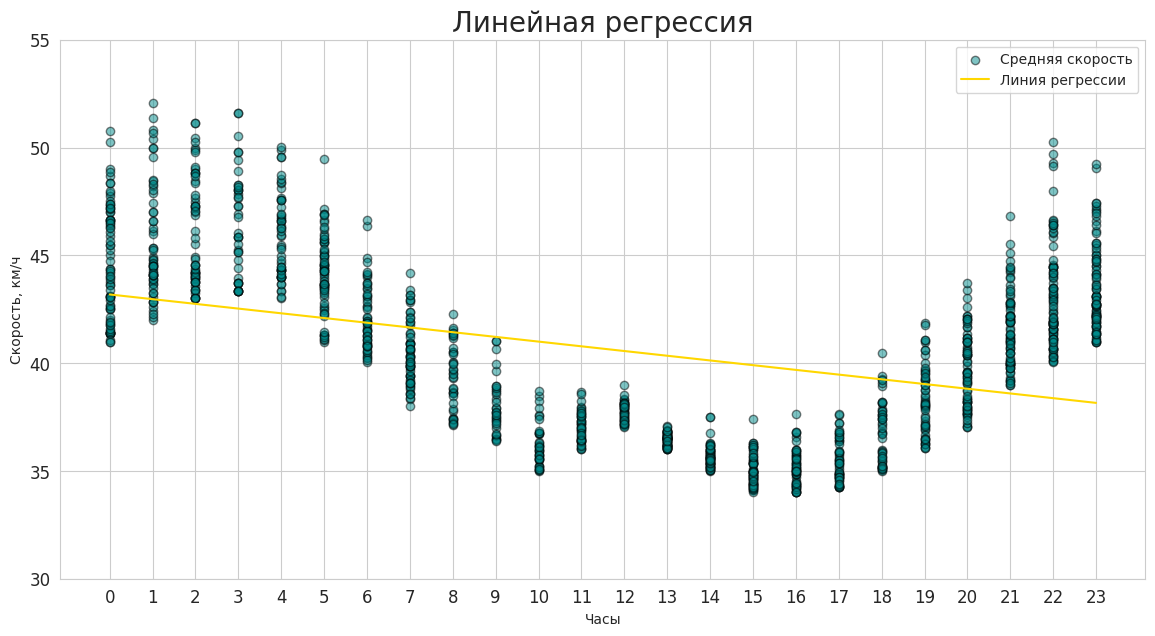


Рисунок 7 – Модель линейной регрессии

Из графика можно сделать вывод, что линейная регрессия не совсем подходит для решения, так как исходные данные имеют нелинейное распределение.

Средняя абсолютная ошибка = 3,19.

Коэффициент детерминации = 0,13.

Множественная линейная регрессия

Для обучения модели множественной линейной регрессии использовался датасет с четырьмя параметрами:

* Часы
* Интенсивность
* День недели фиксации
* Статистика Яндекс пробок

График множественной линейной регрессии представлен на рисунке 8.

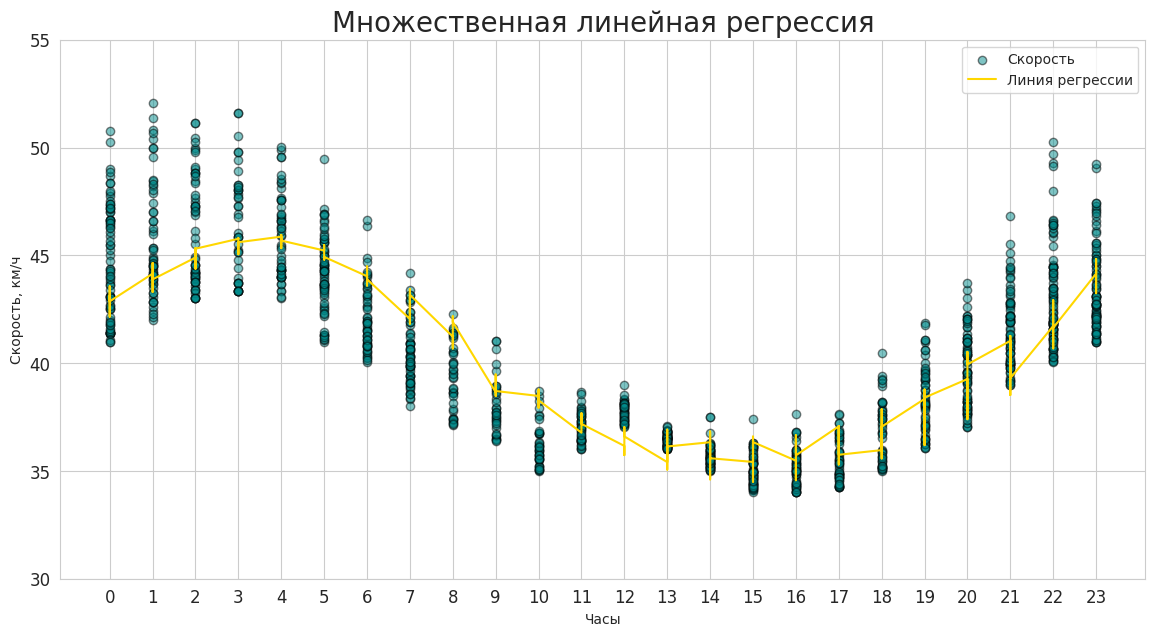


Рисунок 8 – Модель множественной линейной регрессии

На графике видно, что у скорости появился разбег значений. Это можно объяснить тем, что модель пытается подстроить выходные данные, принимая на вход признаки с разными параметрами.

Средняя абсолютная ошибка = 1,58.

Коэффициент детерминации = 0,71.

По сравнению с предыдущей, данная модель показала хороший результат.

Опорно-векторная регрессия (SVR) с использованием линейных и нелинейных ядер

График опорно-векторной регрессии с использованием линейного ядра представлен на рисунке 9.

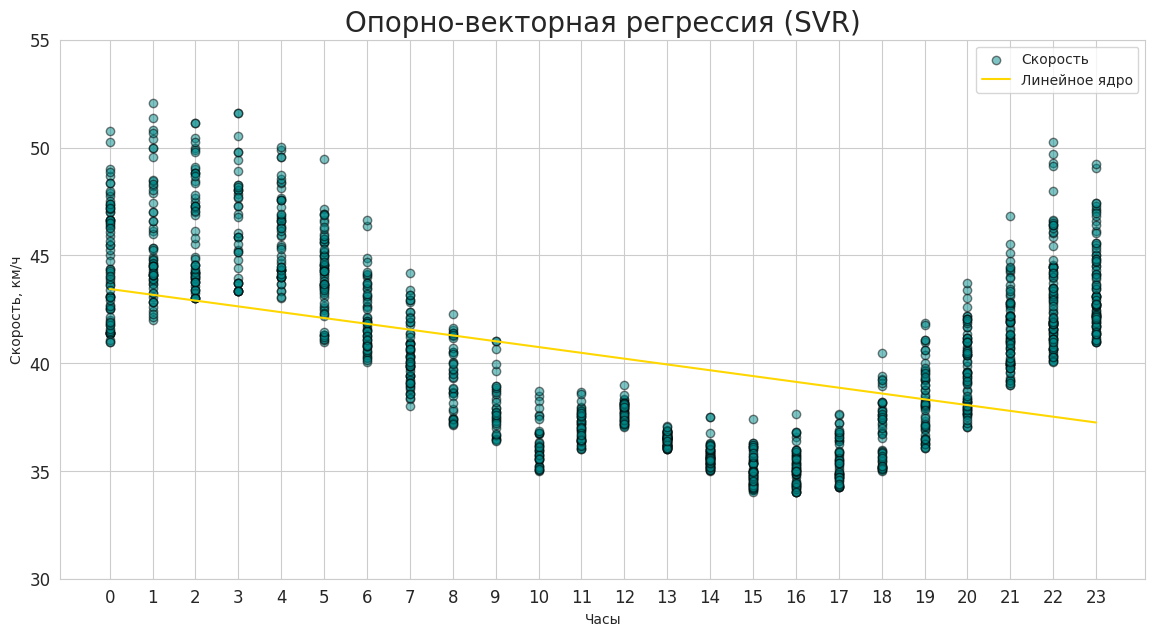


Рисунок 9 – Модель опорно-векторная регрессия с использованием линейного ядра

Опорно-векторная регрессия с использованием линейного ядра показала идентичный результат, как и при обычной линейной регрессии.

Средняя абсолютная ошибка = 3,19.

Коэффициент детерминации = 0,12.

График опорно-векторной регрессии с использованием полиномиального ядра представлен на рисунке 10.

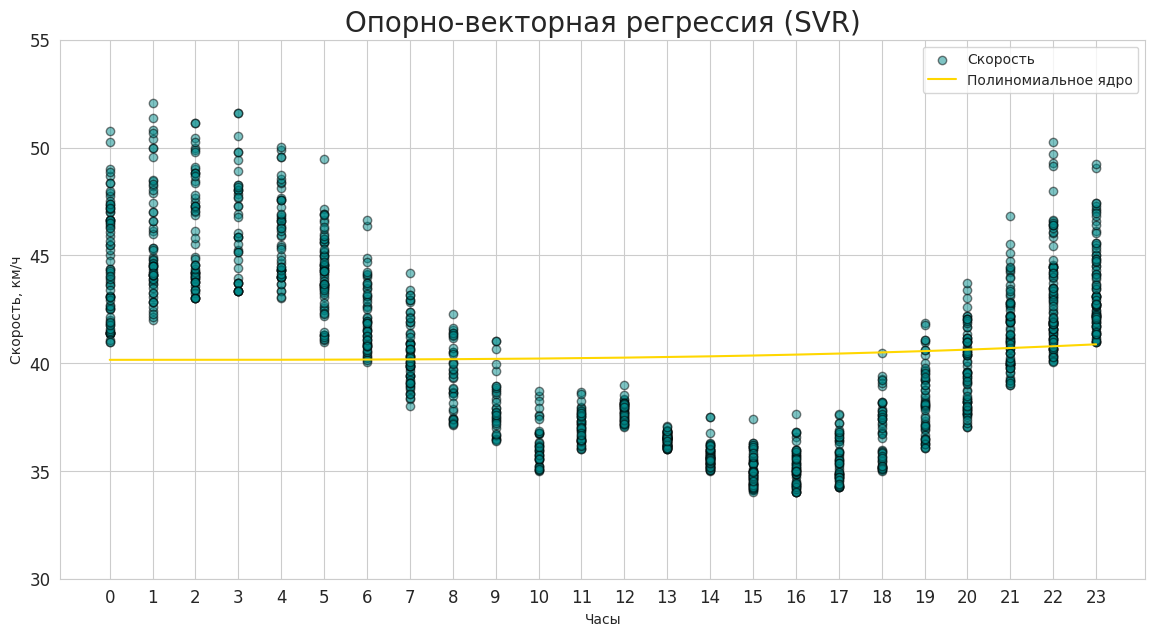


Рисунок 10 – Модель опорно-векторная регрессия с использованием полиномиального ядра

Опорно-векторная регрессия с использованием полиномиального ядра показала результат хуже, чем линейной регрессии.

Средняя абсолютная ошибка = 3,4.

Коэффициент детерминации = 0.

Прогноз с использованием данной модели является неэффективным.

График опорно-векторной регрессии с использованием ядра RBF представлен на рисунке 11.

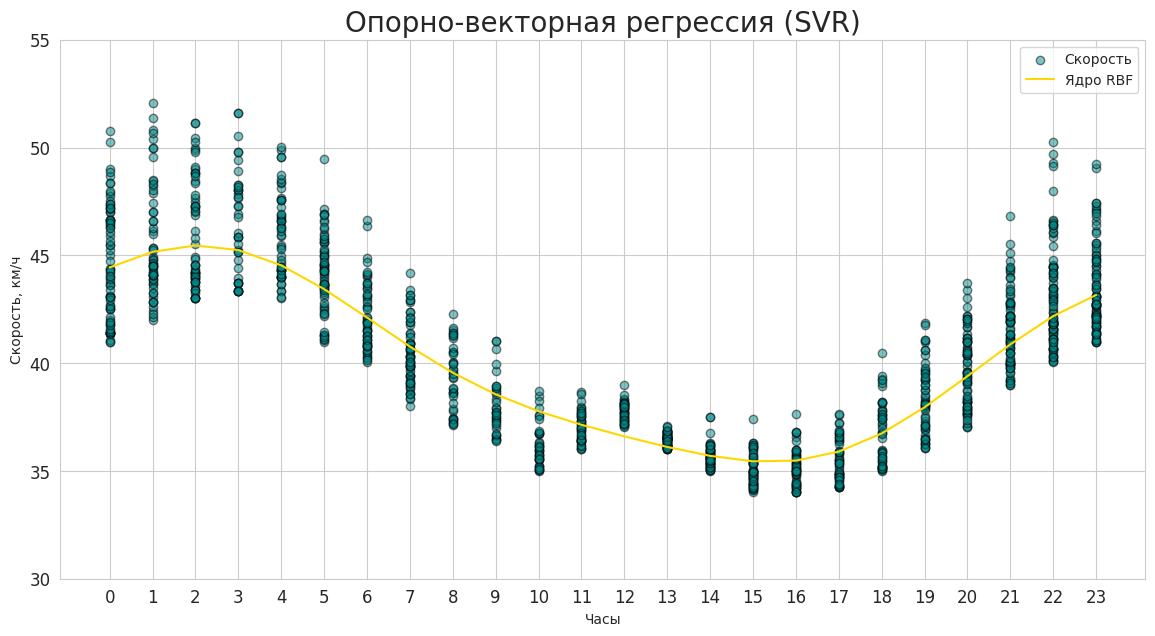


Рисунок 11 – Модель опорно-векторная регрессия с использованием ядра RBF

Опорно-векторная регрессия с использованием ядра RBF показала хороший результат.

Средняя абсолютная ошибка = 1,28.

Коэффициент детерминации = 0,74.

Регрессия k-ближайших соседей

График регрессия k-ближайших соседей представлен на рисунке 12.

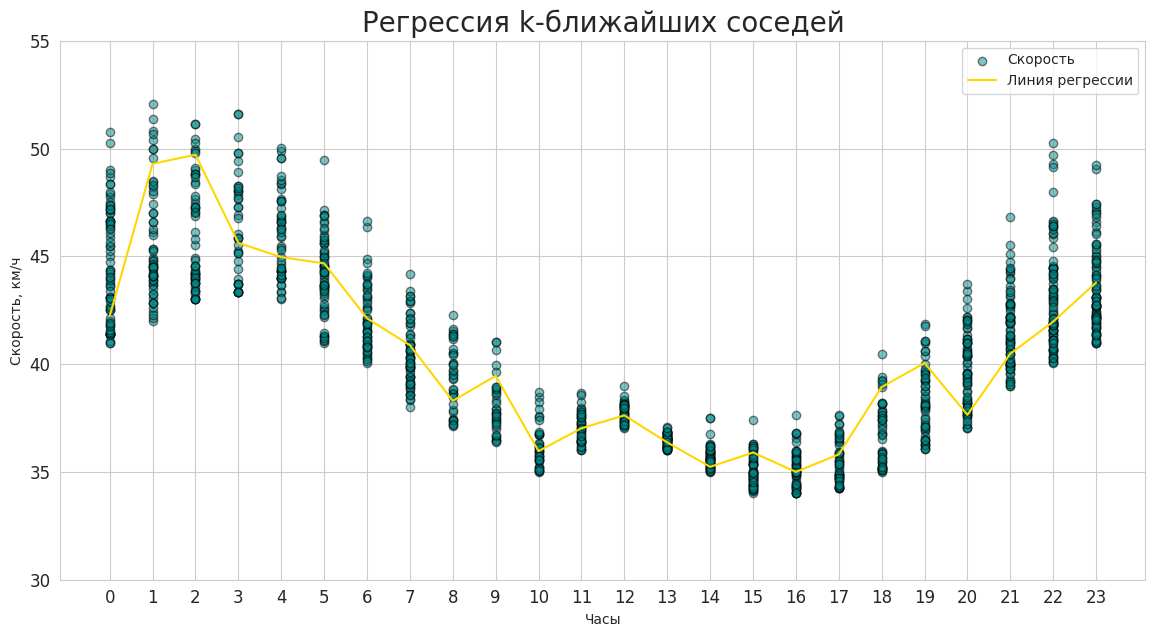


Рисунок 12 – Модель k-ближайших соседей

Регрессия k-ближайших соседей показала так же хороший результат.

Средняя абсолютная ошибка = 1,61.

Коэффициент детерминации = 0,71.

Дерево решений

Модель дерева решений с глубиной равной 2, представлена на рисунке 13.

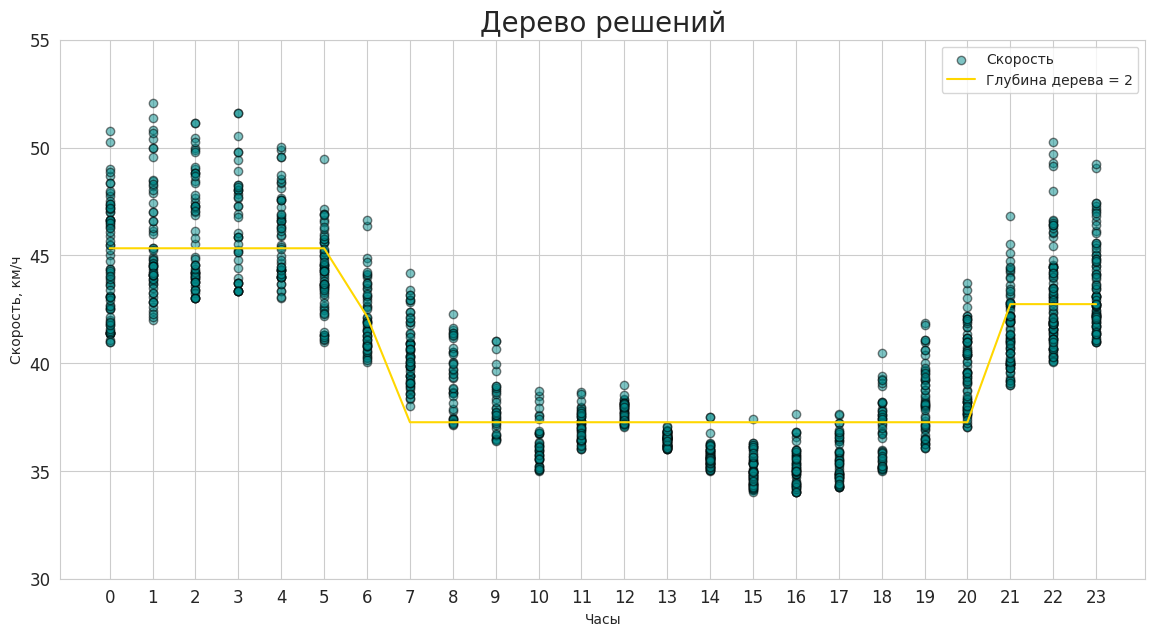


Рисунок 13 – Модель дерева решений с глубиной 2

Средняя абсолютная ошибка = 1,68.

Коэффициент детерминации = 0,62.

На рисунке 14 представлена визуализация дерева решений.

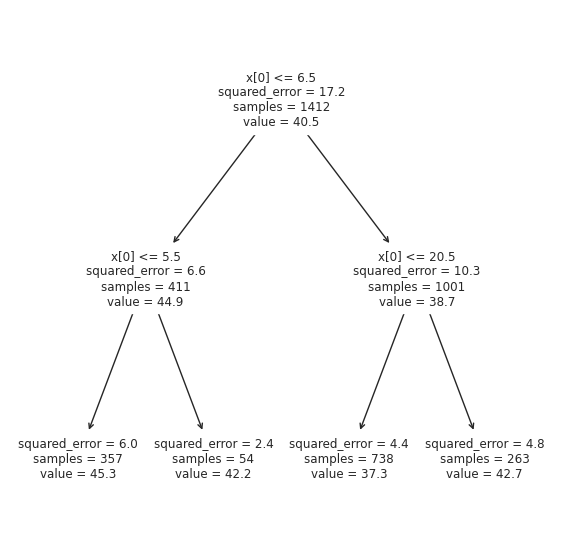


Рисунок 14 – Визуализация решения модели случайного леса

Модель дерева решений с глубиной равной 3, представлена на рисунке 15.

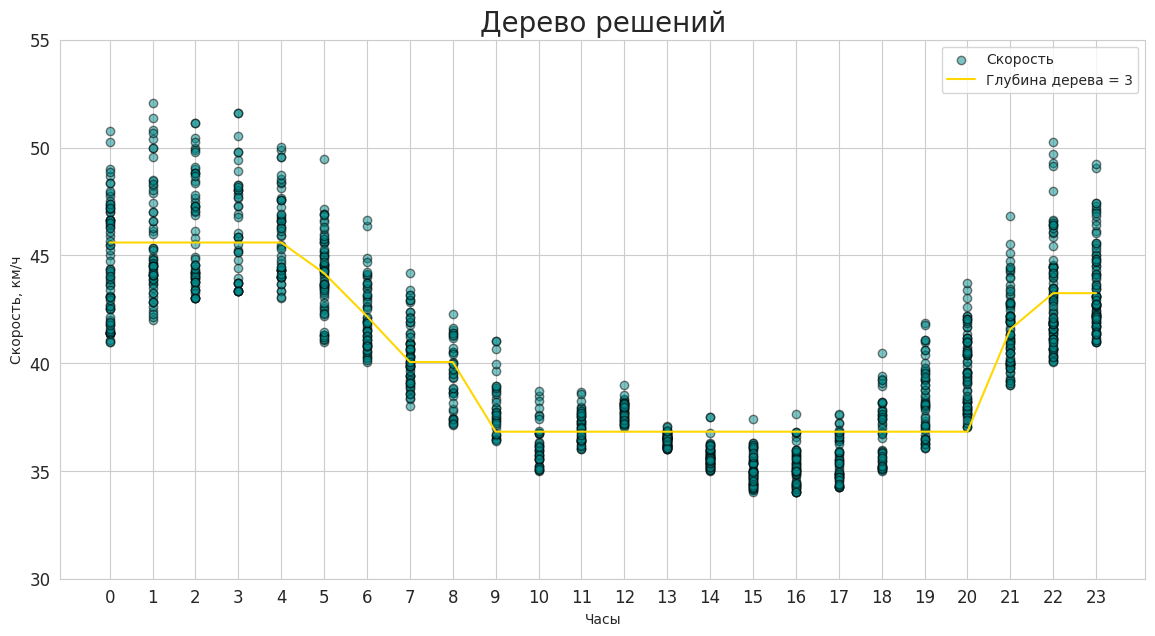


Рисунок 15 – Модель дерева решений с глубиной 3

Средняя абсолютная ошибка = 1,59.

Коэффициент детерминации = 0,69.

На рисунке 16 представлена визуализация дерева решений.

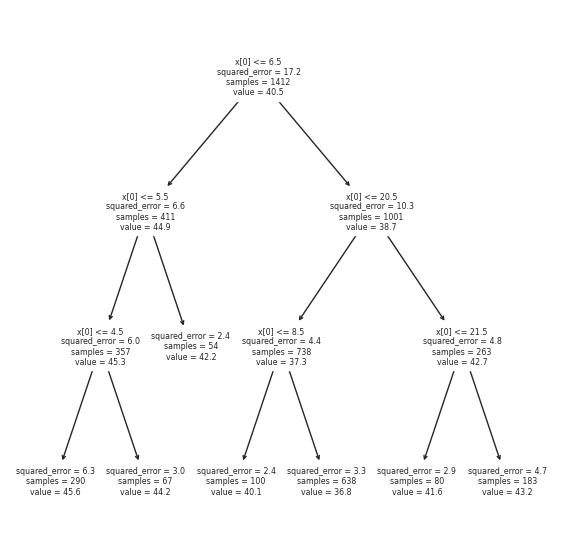


Рисунок 16 – Визуализация решения модели случайного леса

Модель дерева решений с глубиной равной 4, представлена на рисунке 17.

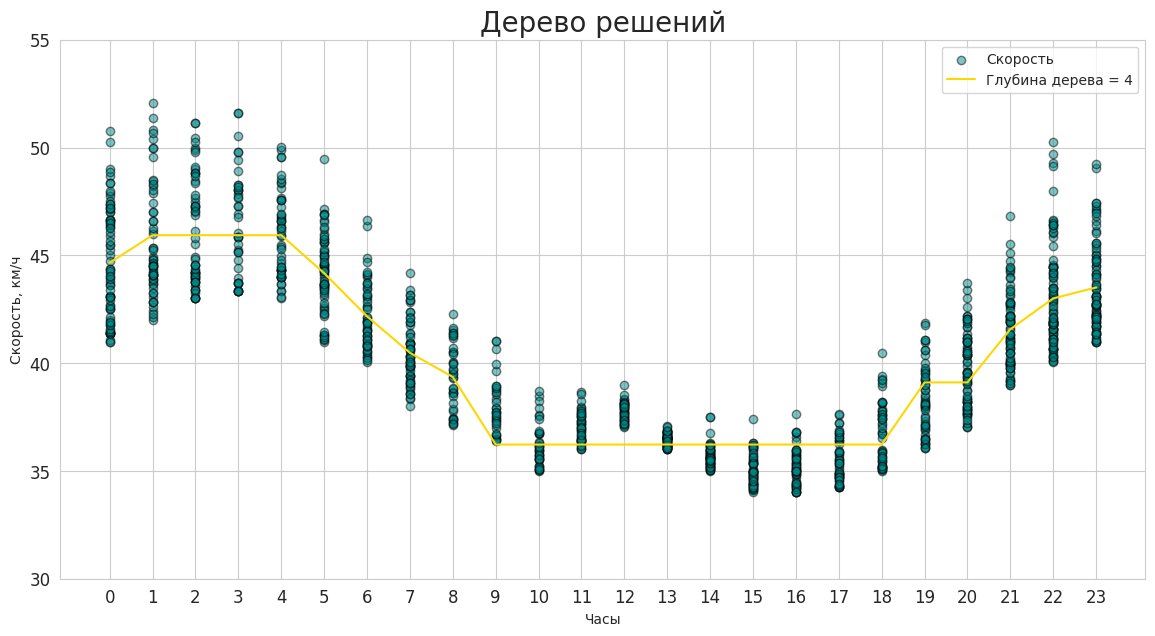


Рисунок 17 – Модель дерева решений с глубиной 4

Средняя абсолютная ошибка = 1,42.

Коэффициент детерминации = 0,76.

На рисунке 18 представлена визуализация дерева решений

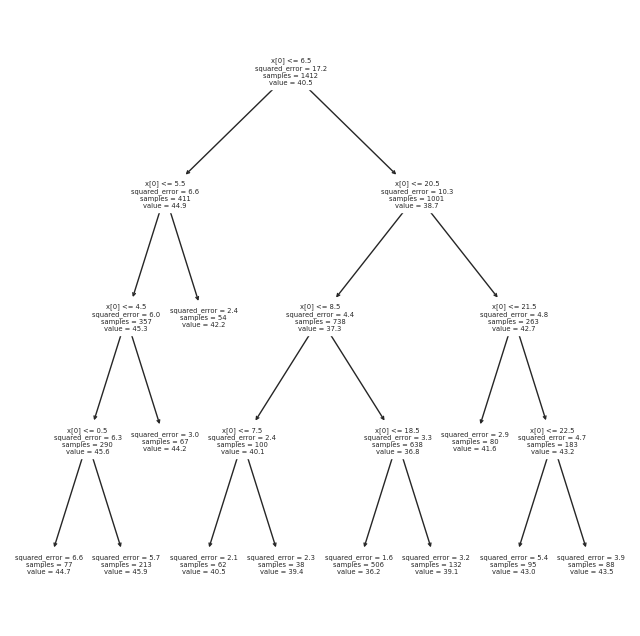


Рисунок 18 – Визуализация решения модели случайного леса

Модель дерева решений с глубиной равной 5, представлена на рисунке 19.

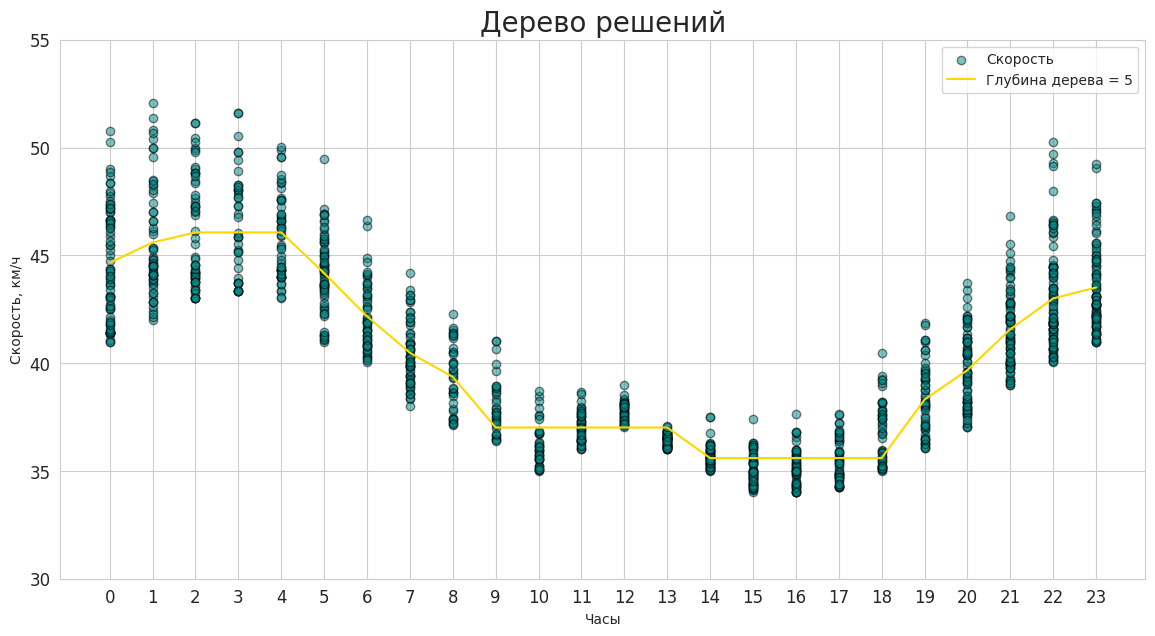


Рисунок 19 – Модель дерева решений с глубиной 5

Средняя абсолютная ошибка = 1,32.

Коэффициент детерминации = 0,78.

На рисунке 20 представлена визуализация дерева решений.

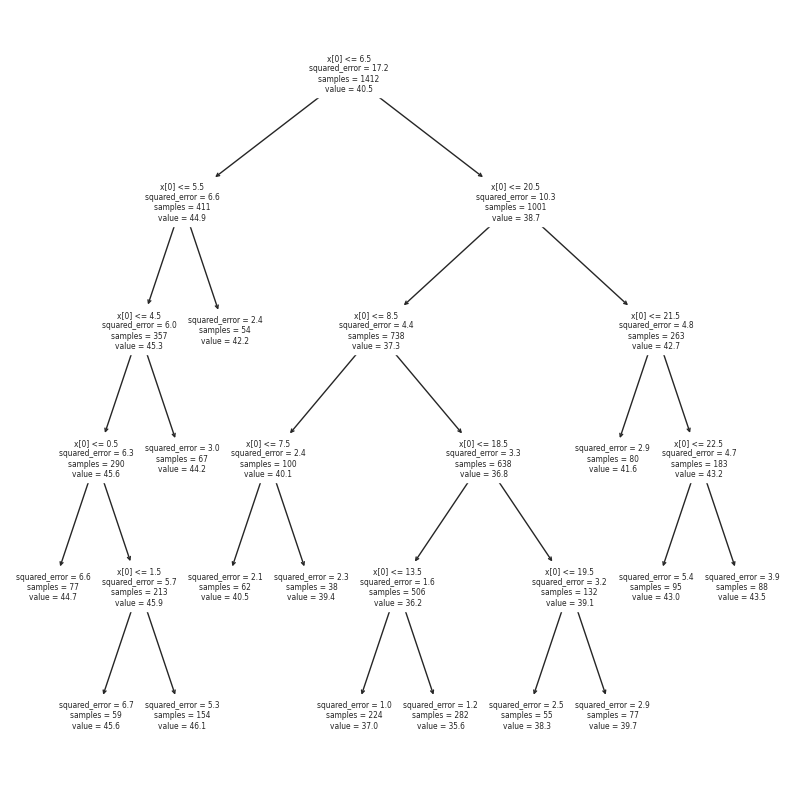


Рисунок 20 – Визуализация решения модели случайного леса

Случайный лес

Модель случайного леса с глубиной деревьев равной 2, представлена на рисунке 21.

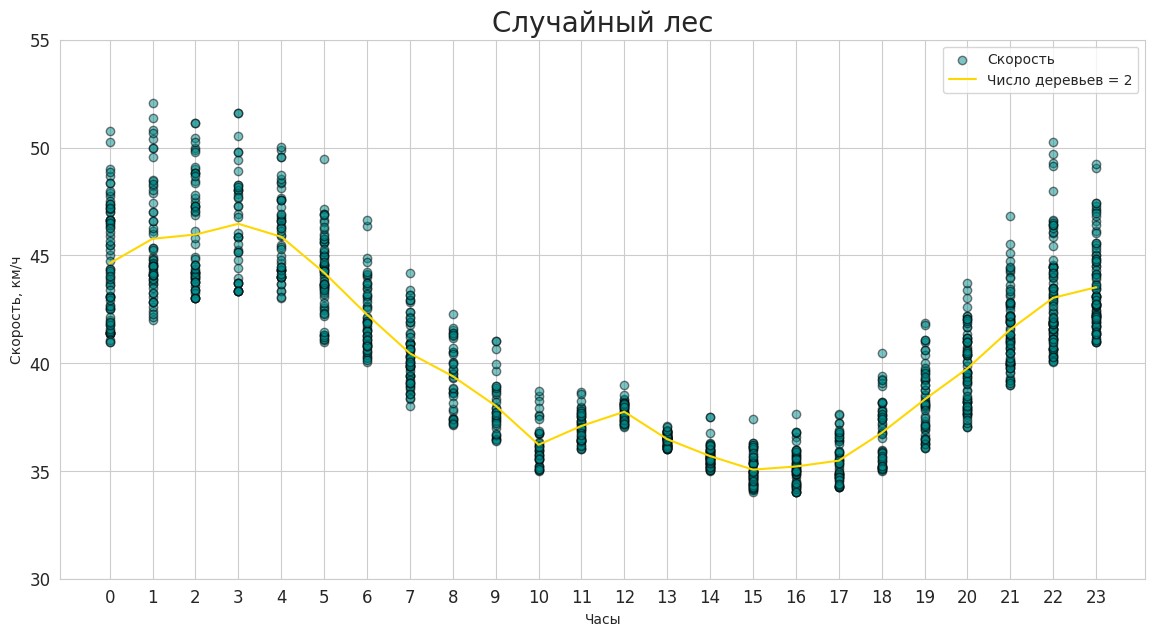


Рисунок 21 – Случайный лес

Модель случайного леса показала хороший результат и смогла отследить изменение зависимости от 10 до 15 часов.

Средняя абсолютная ошибка = 1,27.

Коэффициент детерминации = 0,79.

На рисунке 22 представлена визуализация дерева решений.

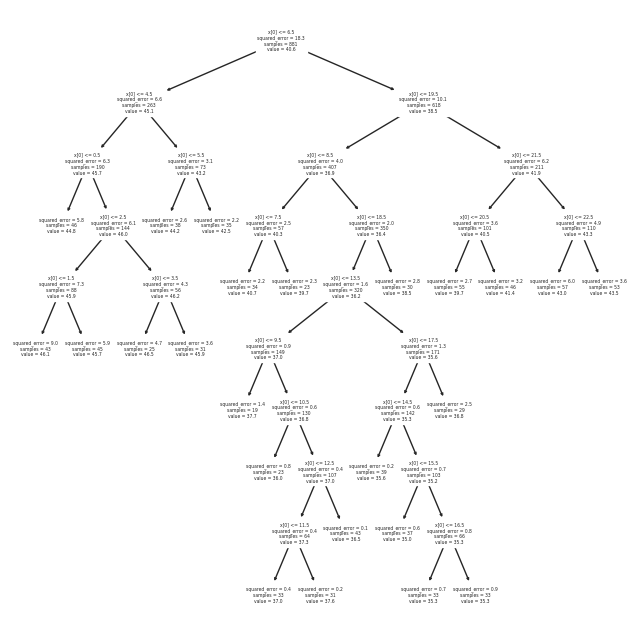


Рисунок 22 – Визуализация решения модели случайного леса

По сравнению с моделью дерева решений, удалось получить более точную зависимость данных, так как результирующая дисперсия существенно уменьшилась.

Полиномиальная регрессия

Полиномиальная регрессия позволяет моделировать более сложные зависимости, но важно следить за переобучением, особенно при использовании высоких степеней полинома.

Модель полиномиальной регрессии со степенью полинома равной 5, представлена на рисунке 23.

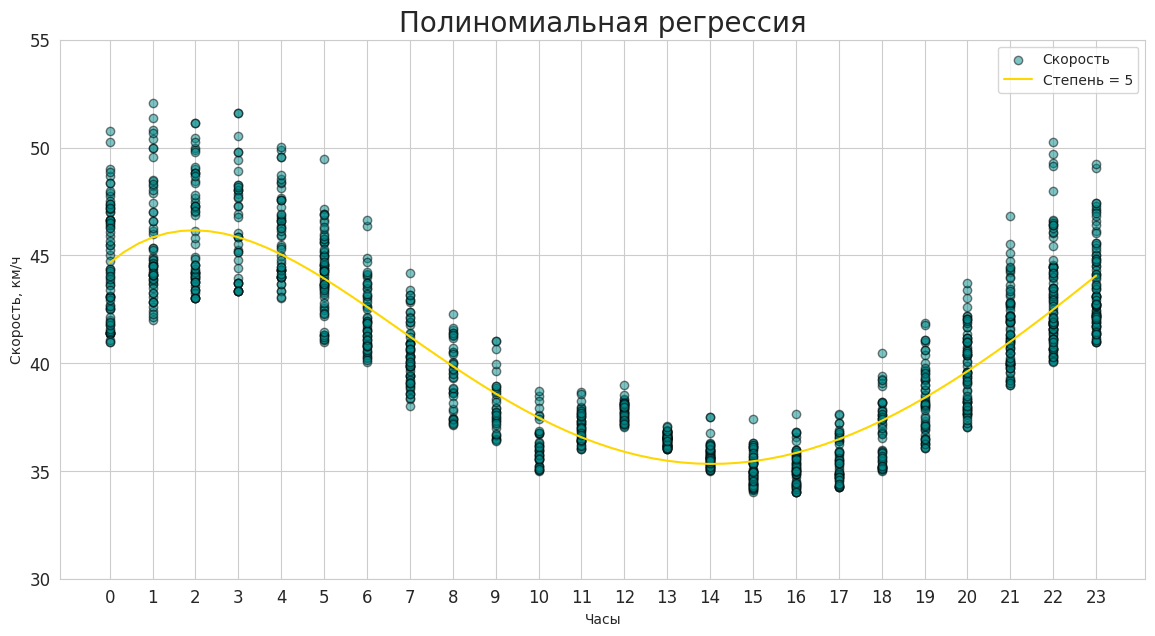


Рисунок 23 – Полиномиальная регрессия со степенью полинома 5

Средняя абсолютная ошибка = 1,77.

Коэффициент детерминации = 0,77.

Подобранное уравнение полиномиальной регрессии:

Модель полиномиальной регрессии со степенью полинома равной 7, представлена на рисунке 24.

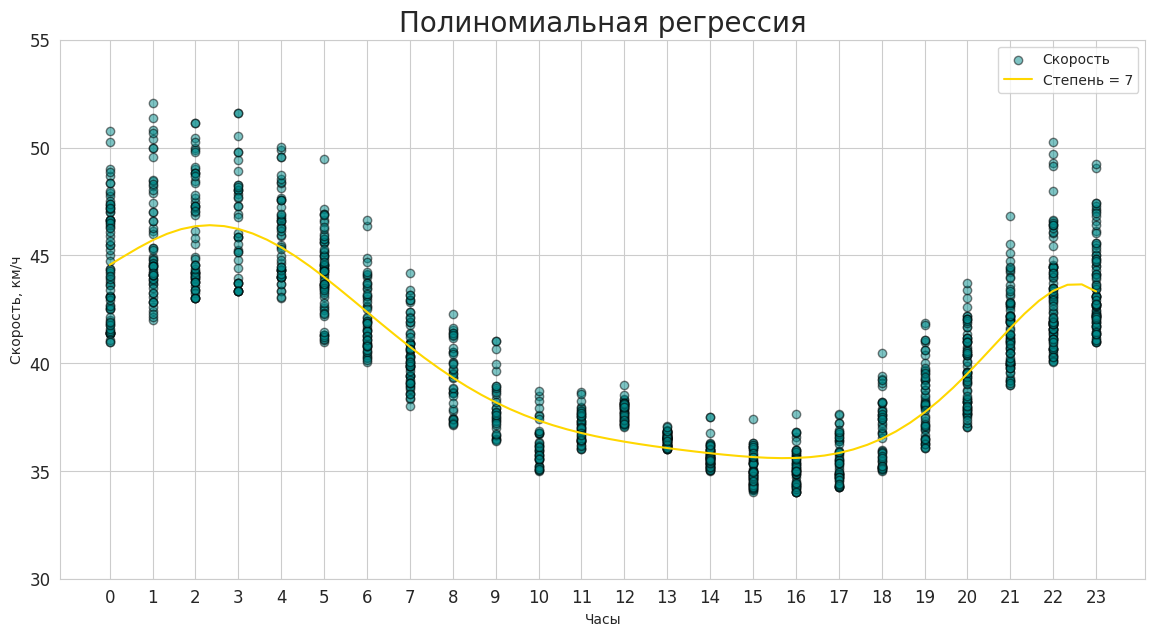


Рисунок 24 – Полиномиальная регрессия со степенью полинома 7

Средняя абсолютная ошибка = 1,74.

Коэффициент детерминации = 0,78.

Подобранное уравнение полиномиальной регрессии:

Модель полиномиальной регрессии со степенью полинома равной 10, представлена на рисунке 25.

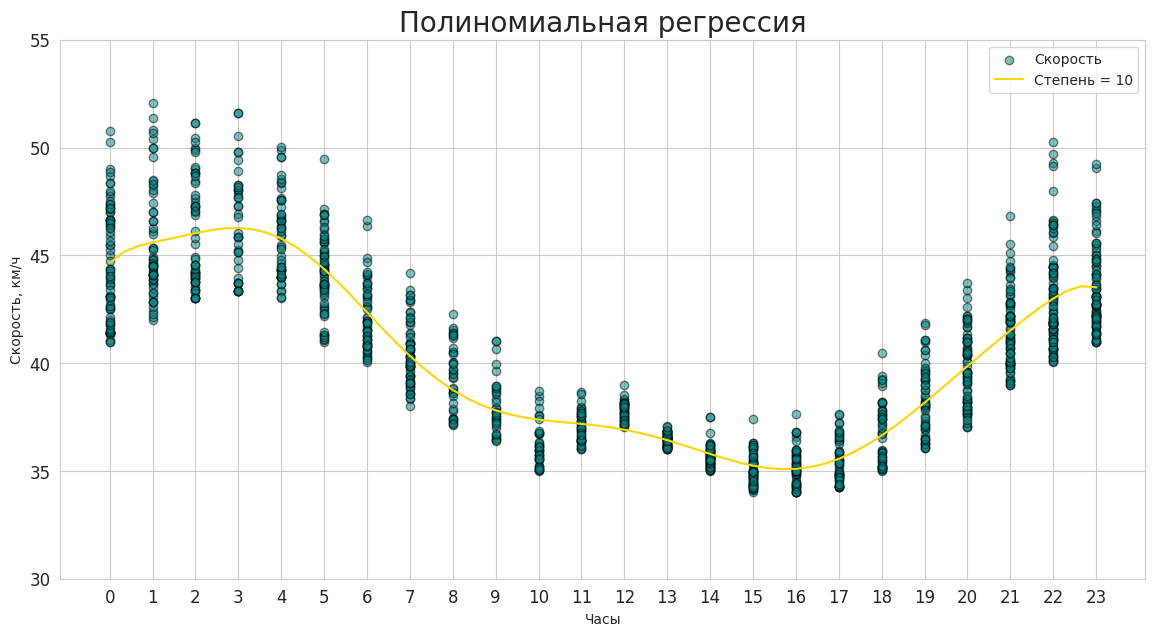


Рисунок 25 – Полиномиальная регрессия со степенью полинома 10

Средняя абсолютная ошибка = 1,71.

Коэффициент детерминации = 0,79.

Подобранное уравнение полиномиальной регрессии:

Эмпирическим путем было выявлено, что модели со степенью полинома больше 10 подвержены переобучению, так как слишком точно подгоняет обучающие данные. На тренировочных данных коэффициент детерминации показывал значения больше 0,80. Однако при использовании тестовых данных, этот показатель существенно падал.

## 2.4 Нейронная сеть

В качестве нейронной сети был принят многослойный персептрон.

Гиперпараметры модели:

- количество скрытых слоев = 3;

- количество нейронов на слое = 120 и 30;

- активационная функция «relu»;

- количество нейронов на выходном слое = 1;

- оптимизатор «Adam»;

Нейронная сеть с одним входным сигналом

Обучение модели происходило за 100 эпох.

Изменение ошибки обучения и валидации, представлена на рисунке 26.

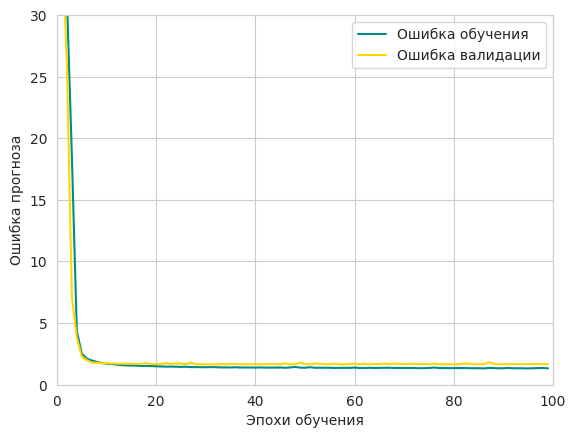


Рисунок 26 – Изменение ошибки обучения и валидации за время обучения модели

Модель прогноза нейросети, представлена на рисунке 27.

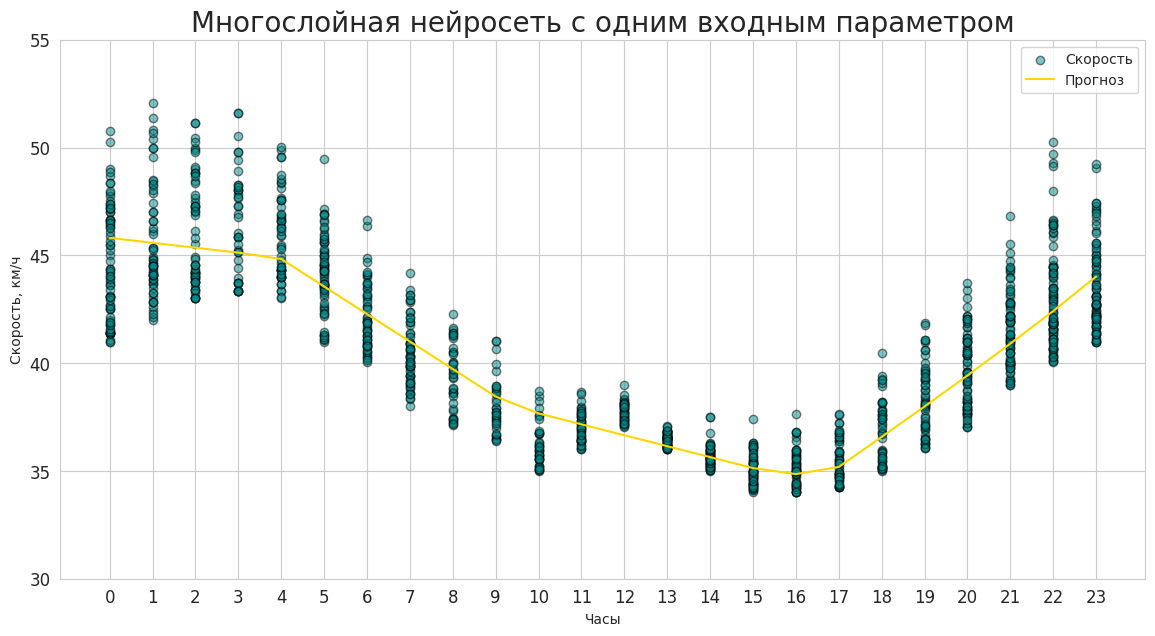


Рисунок 27 – Результат прогноза нейросети

Нейронная сеть с несколькими входными параметрами

Изменение ошибки обучения и валидации, представлена на рисунке 28.

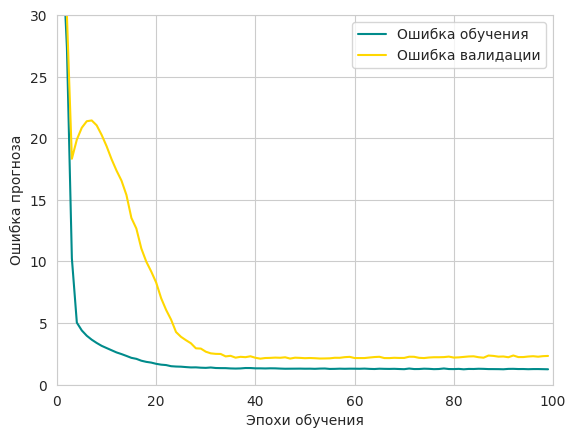


Рисунок 28 – Изменение ошибки обучения и валидации за время обучения модели

Предсказанные значения скорости, представлены на рисунке 29.

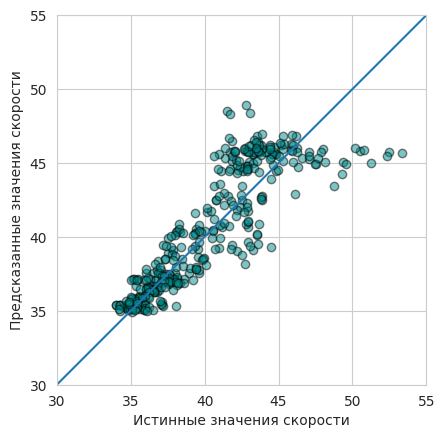


Рисунок 29 – Предсказанные значения скорости

Распределение ошибки, представлено на рисунке 30.

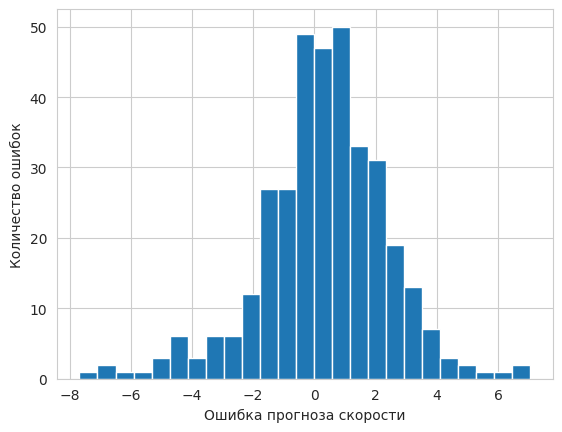


Рисунок 30 – Распределение ошибки

# Заключение

В ходе выполнения ВКР были изучены способы анализа и предобработки данных. Так же было обучено 8 моделей машинного обучения с различными параметрами. Для решения задачи прогнозирования скорости транспортного потока были выявлены наиболее эффективные модели, а именно: случайный лес и полиномиальная регрессия со степенью полинома 10.

Полученные модели нейронных сетей так же позволяют прогнозировать скорость транспортного потока, но абсолютная ошибка не много больше, чем у вышеописанных моделей.

# Библиографический список

1. Бобков, А.Ю. Прогнозирование транспортного потока в условиях динамического спроса / А.Ю. Бобков. – М.: Транспорт, 2020. – 200 с.
2. Дорогов, Н.П. Координированное управление светофорами в системе АСУДД / Н.П. Дорогов. – Казань: Казанский университет, 2021. – 220 с.
3. Иванов, П.А. Оптимизация транспортных потоков на основе прогнозирования / П.А. Иванов. – Новосибирск: Сибирское научное издательство, 2018. – 150 с.
4. Петров, В.Д. Анализ и прогнозирование транспортных потоков / В.Д. Петров. – Владивосток: ВГГУ, 2020. – 160 с.
5. Тихонов, Г.И. Устойчивое управление транспортными потоками / Г.И. Тихонов. – Ярославль: ЯГТУ, 2019. – 195 с.
6. Ульянов, К.П. Основы модели прогнозирования транспортных потоков / К.П. Ульянов. – Омск: ОГУ, 2020. – 210 с.