Прогноз температуры почвы по метеоданным RP5

Учебный проект по анализу данных

19 декабря 2025 г.

Источник данных

- Метеоданные получены с RP5 (архивы наблюдений метеостанций)
- Температура почвы — из отдельных файлов замеров
- Данные имеют временную структуру и пропуски

Мотивация проекта

- Температура почвы важна для изучеия поведения вечной мерзлоты
- Влияет разрушение городской инфраструктуры
- Задача прогноза нетривиальна из-за инерционности процесса

Цель проекта

Построить модель машинного обучения, прогнозирующую температуру почвы на основе метеорологических параметров.

Линейная регрессия используется как базовая (baseline) модель.

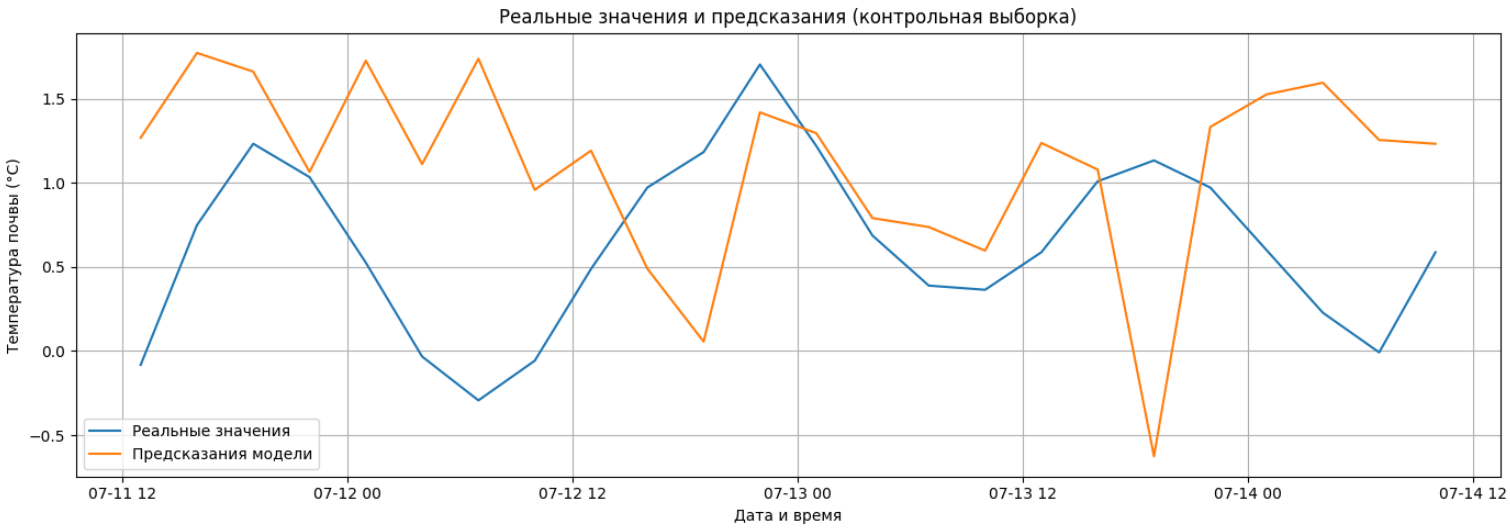
Подготовка данных

- Объединение данных RP5 и замеров по дате и времени
- Очистка пропусков
- Кодирование категориальных признаков
- Масштабирование признаков
- Добавление лагов температуры воздуха

Модель

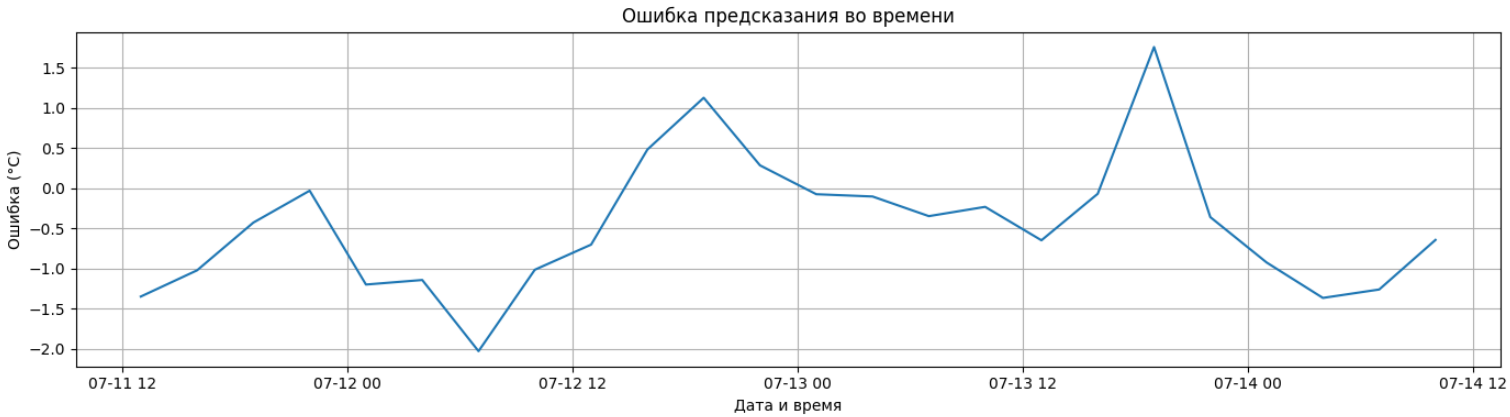
- Линейная регрессия (метод наименьших квадратов)
- Лаги температуры воздуха: $t-1, t-2, t-3$
- Разбиение данных: 70% обучение / 30% контроль
- Без перемешивания (учёт временной структуры)

Реальные значения и предсказания



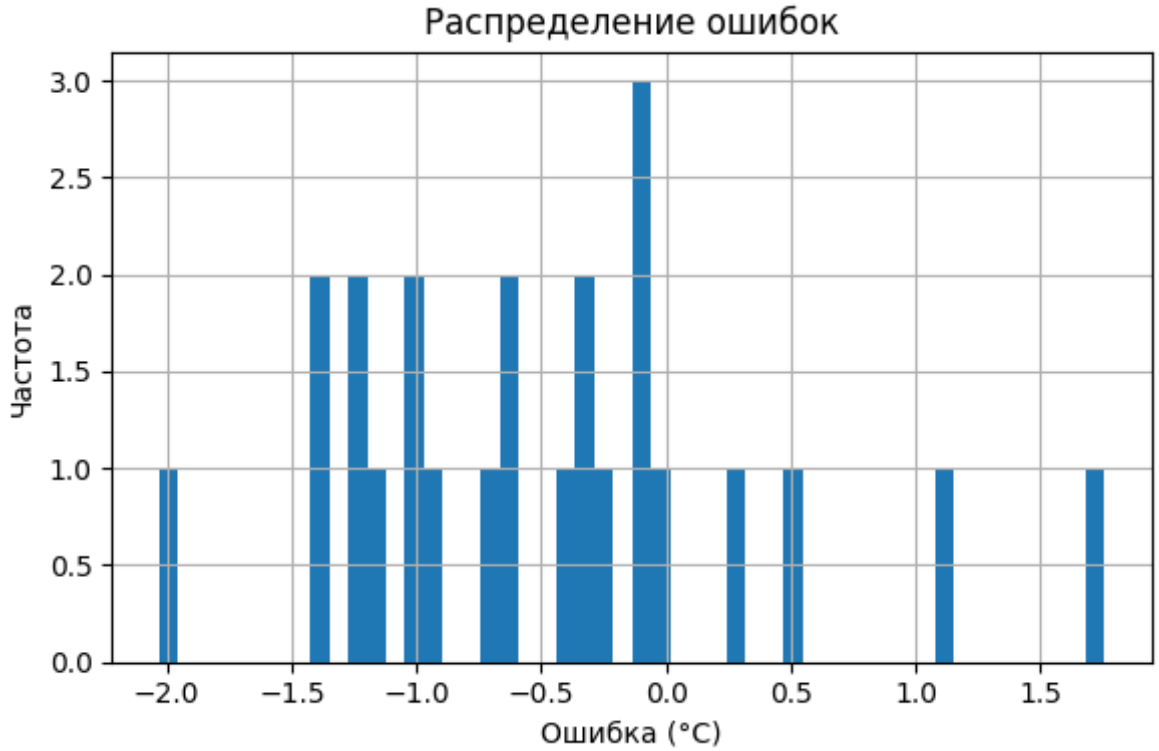
Модель хорошо воспроизводит общий тренд, но сглаживает резкие изменения.

Ошибка предсказания во времени



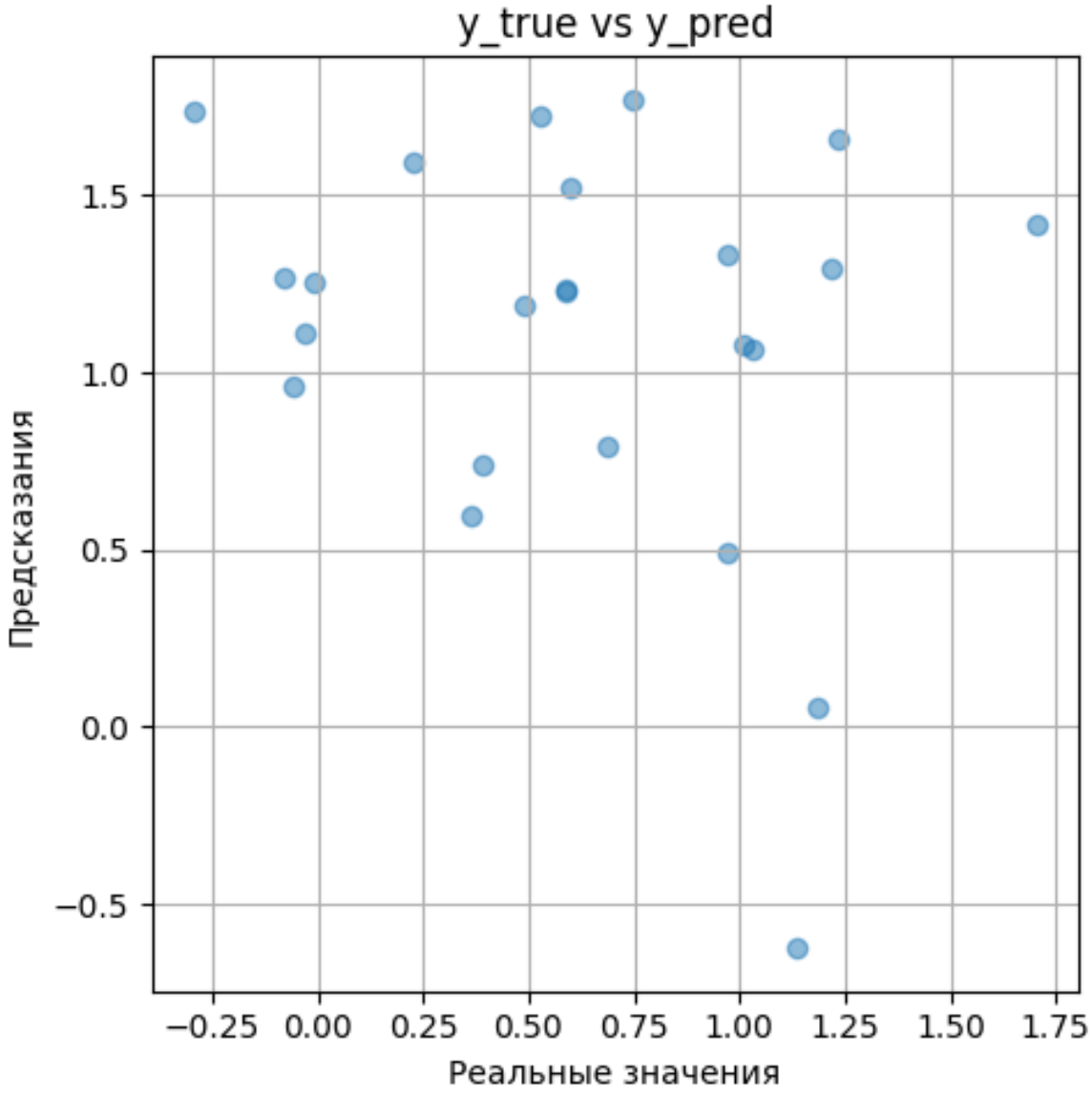
Наблюдается автокорреляция ошибок.

Распределение ошибок



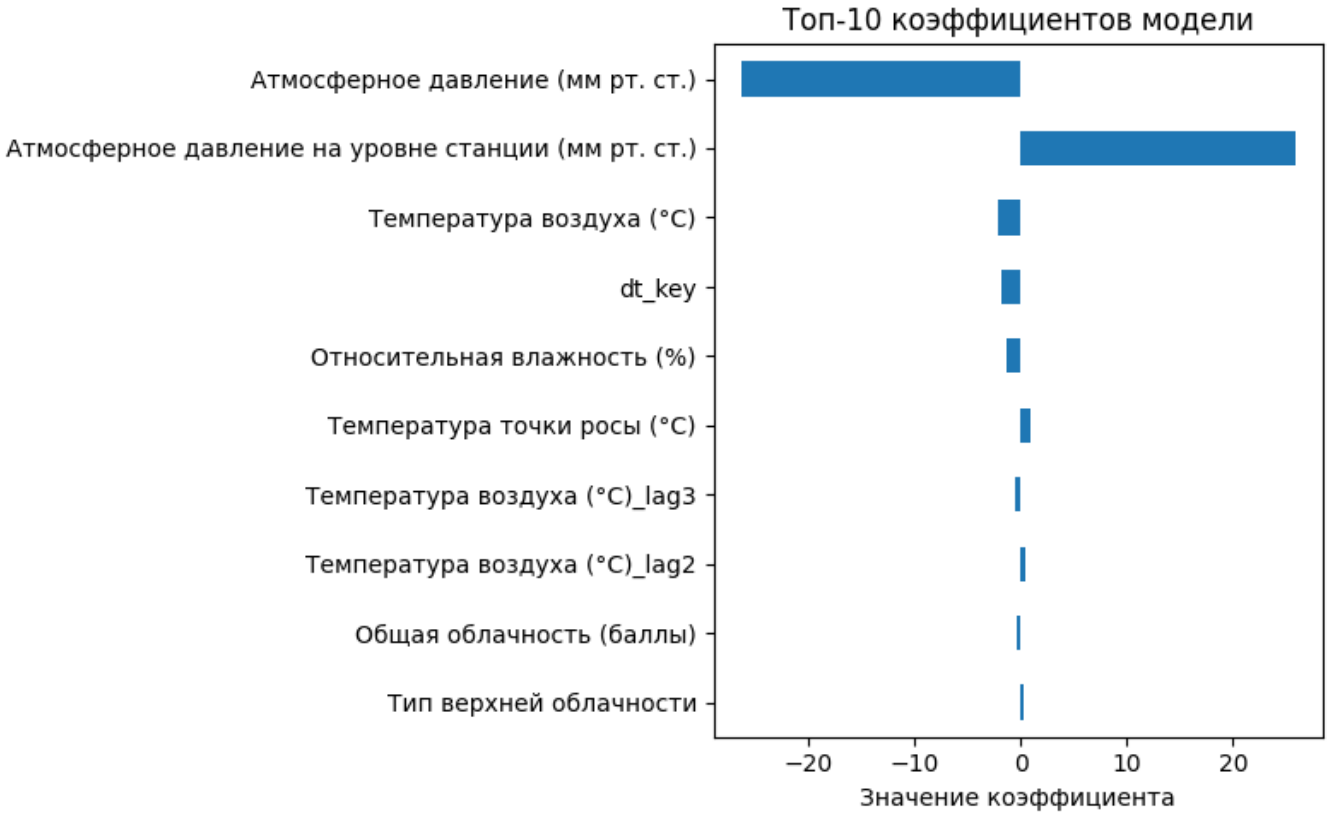
Ошибки сосредоточены около нуля, но присутствуют выбросы.

Диаграмма y_{true} vs y_{pred}



Линейная модель плохо воспроизводит экстремальные значения.

Коэффициенты модели



Наибольший вклад вносят параметры давления и лаги температуры воздуха.

Количественные результаты

- MAE $\approx 0.8^\circ$C — приемлемая абсолютная ошибка
- RMSE $< 1^\circ$C
- $R^2 < 0$ — модель плохо объясняет дисперсию

MAE и R^2 отражают разные аспекты качества модели.

Итоговые выводы

- Линейная регрессия — корректный baseline
- Добавление лагов улучшает MAE и RMSE
- Отрицательный R^2 указывает на ограничения модели
- Ошибки автокоррелированы

Перспективы развития

- Увеличение числа лагов
- Добавление сезонных признаков
- Ridge / ElasticNet
- Нелинейные модели (градиентный бустинг)