ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Проект по Machine Learning

Артем Варданян, Руслан Исламов, Никита Чен

ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Этап 1

Работа с аномалиями и генерация признаков

Цель

Проект направлен на построение модели, способной точно прогнозировать ежедневный спрос на прокат велосипедов по часам. Точный прогноз позволит оператору заранее планировать балансировку парка, оптимизировать логистику и снизить издержки.

Постановка задачи

Данные

17 379 почасовых записей за 2 года. Включают метеопараметры, праздники, выходные и фактический спрос.

Задача

Разработать модель для предсказания числа аренд в следующем часу по внешним условиям.

Метрики

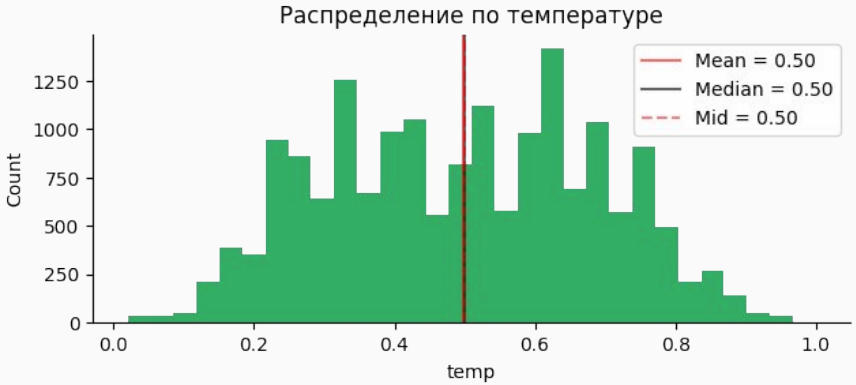
Качество оценивается по **MAE** и **RMSE** на отложенной выборке.

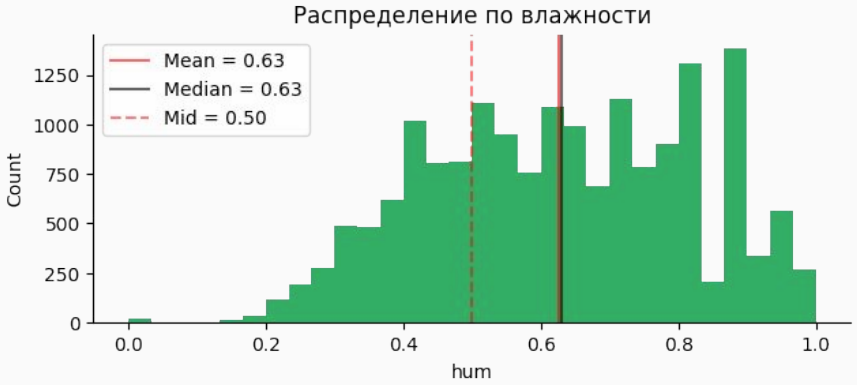
Загрузка данных

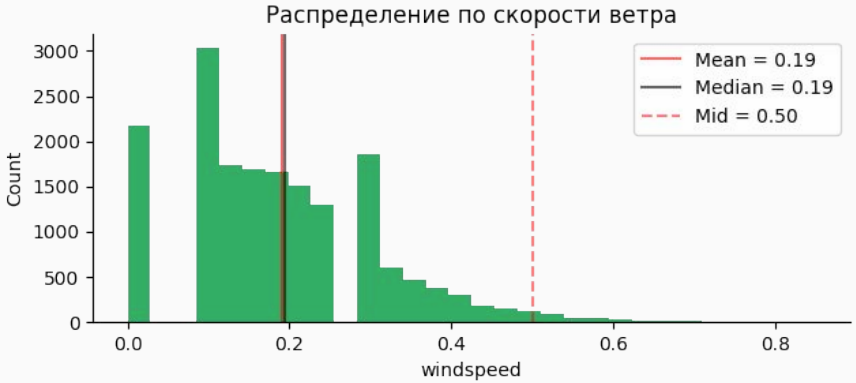
Набор данных загружен через KaggleHub. Он содержит 17 379 почасовых строк и 17 полей. Первые строки демонстрируют структуру: дата, сезон, час, погода, нормализованные температура и влажность, категории пользователей и итоговый спрос.

№	Название	Кол-во заполненных строк	Тип
0	instant	17379	int64
1	dteday	17379	object
2	season	17379	int64
3	yr	17379	int64
4	mnth	17379	int64
5	hr	17379	int64
6	holiday	17379	int64
7	weekday	17379	int64
8	workingday	17379	int64
9	weathersit	17379	int64
10	temp	17379	float64
11	atemp	17379	float64
12	hum	17379	float64
13	windspeed	17379	float64
14	casual	17379	int64
15	registered	17379	int64
16	cnt	17379	int64

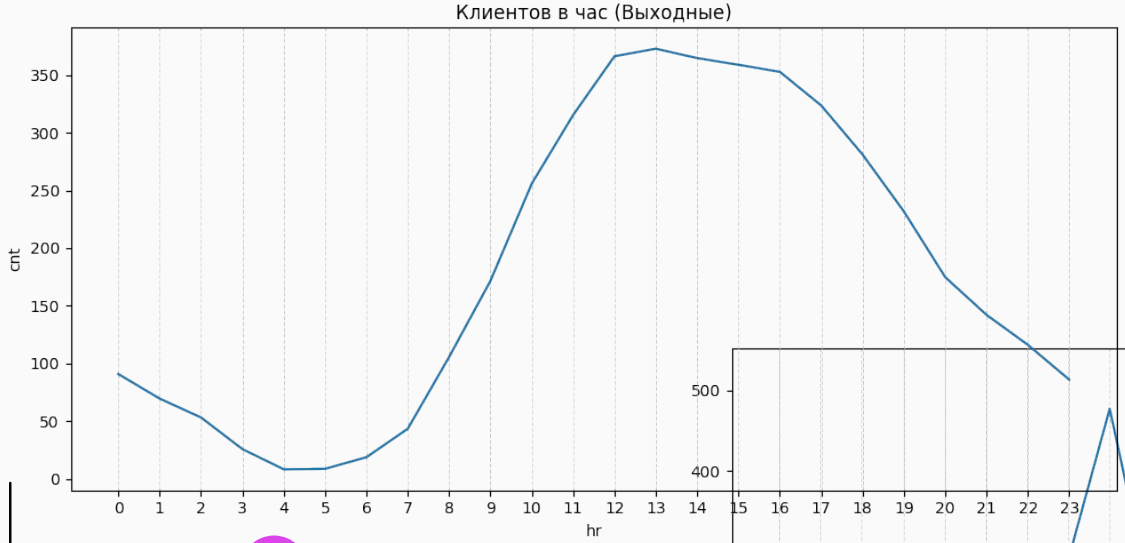
Визуализация распределений

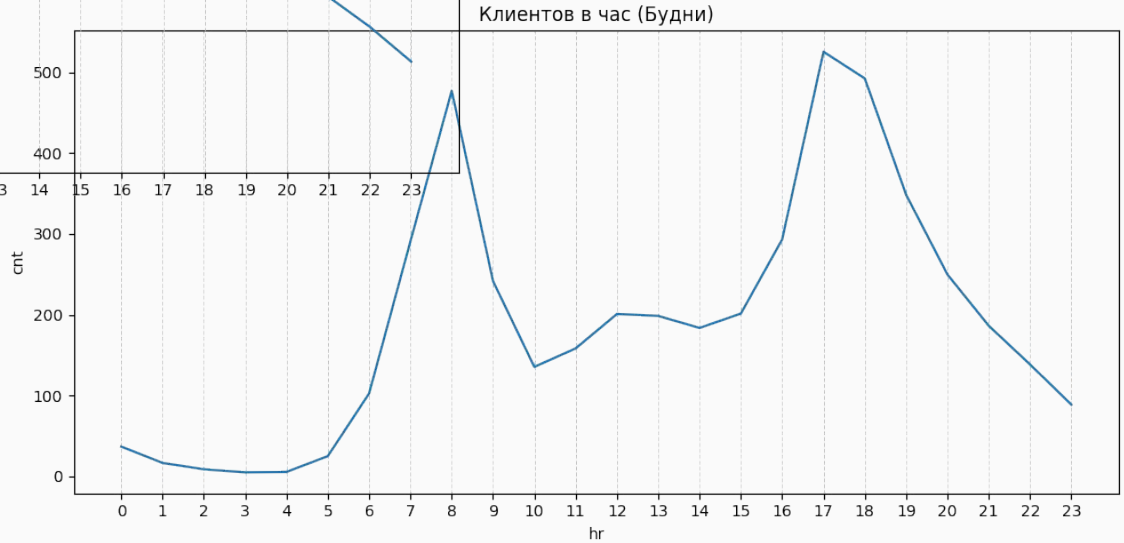






Распределение на неделе





Baseline

MSE

19376,95

RMSE

139,2

Модель

Линейная регрессия

MAE

104,8

R^2

0,3881

ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Этап 2

Работа с аномалиями и генерация признаков

Почему «выбросы» оказались пиками спроса

Стандартизация и выявление аномалий

Стандартизированная переменная cnt_z выявила записи с |z|>3, которые изначально были расценены как выбросы. Однако анализ показал, что это не шум, а естественные пики спроса. Вечерние часы 17-18 и утренний 8 в будни формируют высокий спрос у офисных работников.

Аномалии как признаки

Аномалии подтверждены распределением по датам и часам, что исключает их удаление. Вместо этого, они были учтены в создании новых признаков, чтобы улучшить модельное понимание структуры данных и спроса.

IQR vs Grubbs vs IsolationForest

Метод IQR

Для семи числовых признаков были построены границы IQR и флаги выбросов. Этот метод помог выявить потенциальные аномалии, но оказался не столь эффективным для некоторых переменных

Тест Граббса

Grubbs нашёл выбросы в hum и windspeed. Этот метод оказался полезным для выявления редких, но значимых аномалий в данных

IsolationForest

IsolationForest при contamination=0.01 выделил 1 % наблюдений. Метрика F1=0.83 подтверждает редкость истинных выбросов и целесообразность soft-очистки данных

Генерация признаков

Создание признаков для пико

Созданы is_rush_hour и is_weekday_rush для вечерних будних пиков. Эти признаки помогают моделью учесть специфические часы, когда спрос на аренду велосипедов особенно высок

Доля подписчиков

Значение registered_ratio показывает долю подписчиков. Этот признак важен для понимания структуры спроса и различий между зарегистрированными и незарегистрированными пользователями

Лаги и скользящие средние

Признаки cnt_lag_24h и скользящий hour_mean_cnt Эти признаки помогают модели учесть исторические данные и тенденции спроса

Кодирование категориальных переменных

Кодирование сезонов и погоды завершено через get_dummies с drop_first. Это позволяет модели лучше интерпретировать категориальные данные.

Корреляция и ANOVA F-score

01

Корреляция числовых признаков

Для числовых признаков построена корреляция с cnt: cnt_lag_24h и hour_mean_cnt лидируют. Это показывает их высокую значимость для прогнозирования спроса

02

ANOVA для категориальных признаков

ANOVA показала высокий F-score у season_3, is_rush_hour, workingday. Это подтверждает их важность для модели

03

Отсечение незначимых признаков

Порог $p<0.01$ отсек holiday, weathersit_4 и ряд дней недели. Это позволяет избежать шума и улучшить качество модели

Синус-косинус кодирование времени

01

Циклическое кодирование дней недели

Дни недели преобразованы в weekday_num, затем в weekday_sin/cos с периодом 7. Это позволяет сохранить порядок и близость концов цикла, что повышает качество деревьев и линейных моделей.

02

Циклическое кодирование часов и месяцев

Аналогично строятся цикличные пары для hr и месяца. Такой подход улучшает представление временных данных и способствует более точному прогнозированию.

RFECV и LassoCV: два взгляда на sparsity

RFECV

RandomForest-based RFECV выбрал 9 признаков. Этот метод помог определить основные признаки, которые вносят наибольший вклад в модель.

LassoCV

LassoCV с α, подобранным по 5-fold, оставил 8 переменных с ненулевыми коэффициентами. Пересечение методов дало ядро сильных признаков.

Базовый vs расширенный набор

Базовая модель

CatBoostRegressor с базовыми 9 признаками показал RMSE 68.17 и R^2 0.85

Расширенная модель

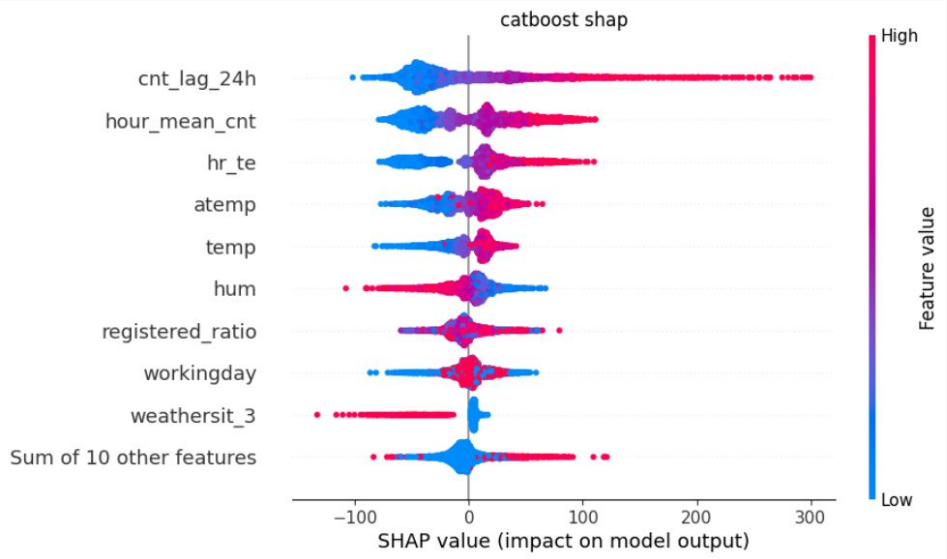
Добавление hum, windspeed, registered_ratio и части
weekday/weathersit снизило RMSE до 53.11, а R^2 выросло до 0.91.
Это подтверждает важность дополнительных признаков

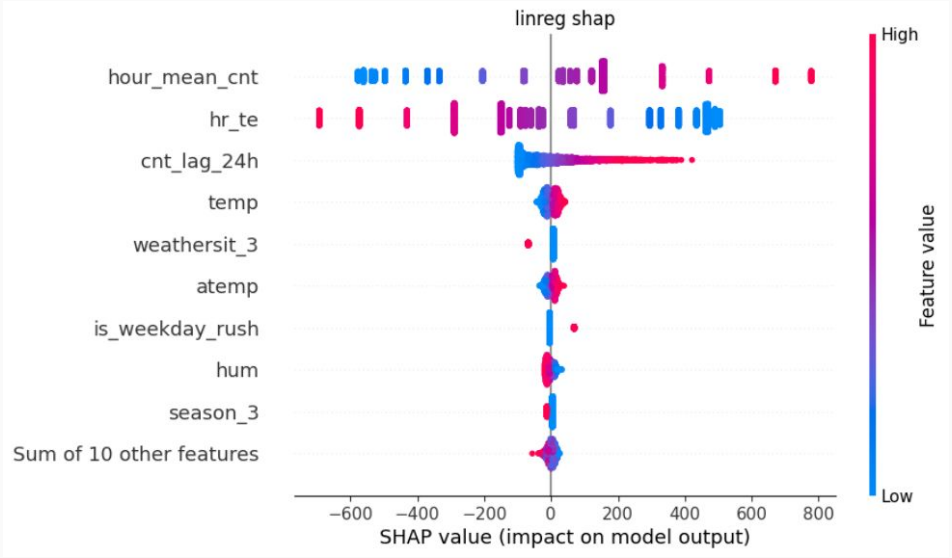
ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Этап 3

Интерпретация и диагностика моделей

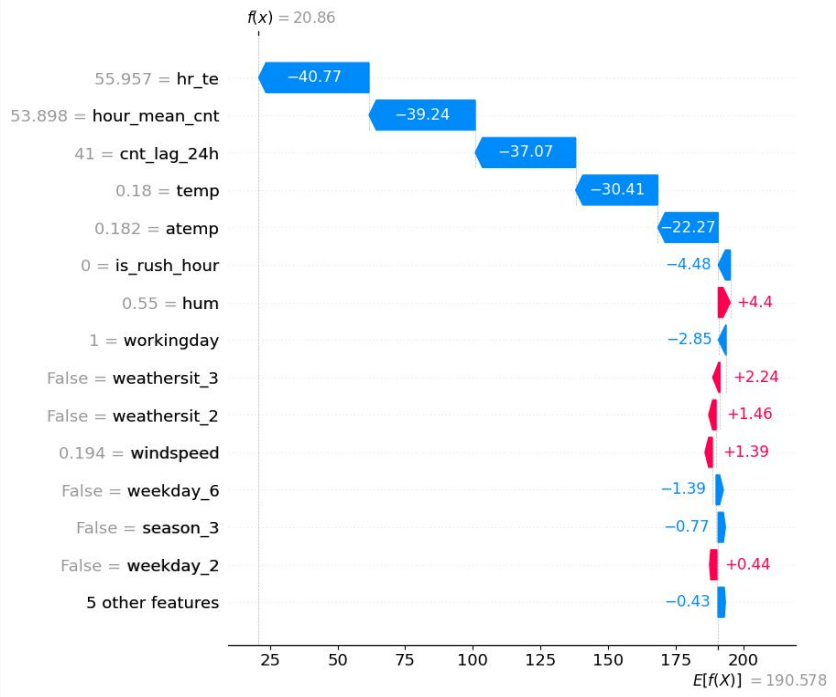
Глобальные интерпретации LIME и SHAP



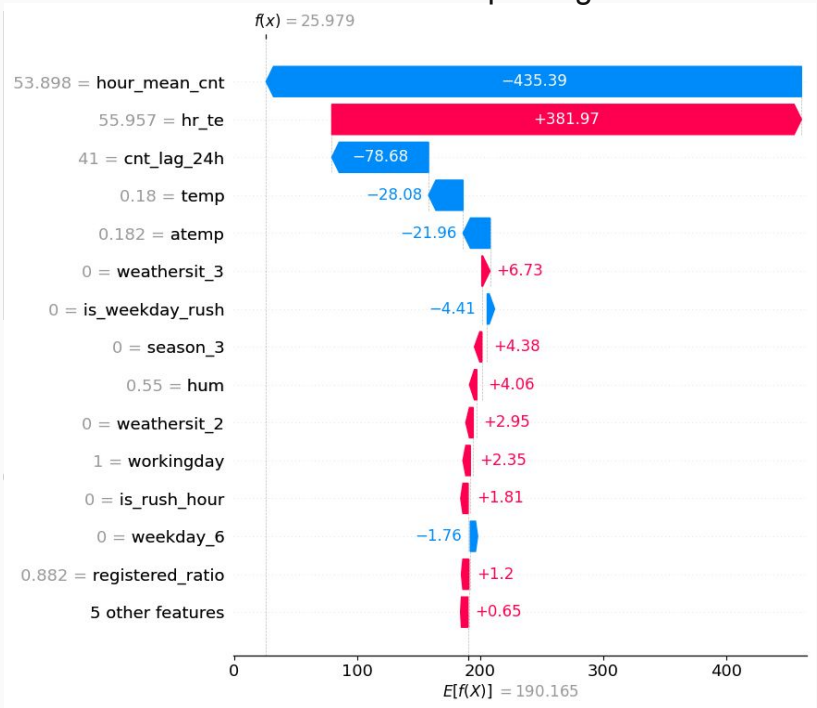


Локальные интерпретации LIME и SHAP

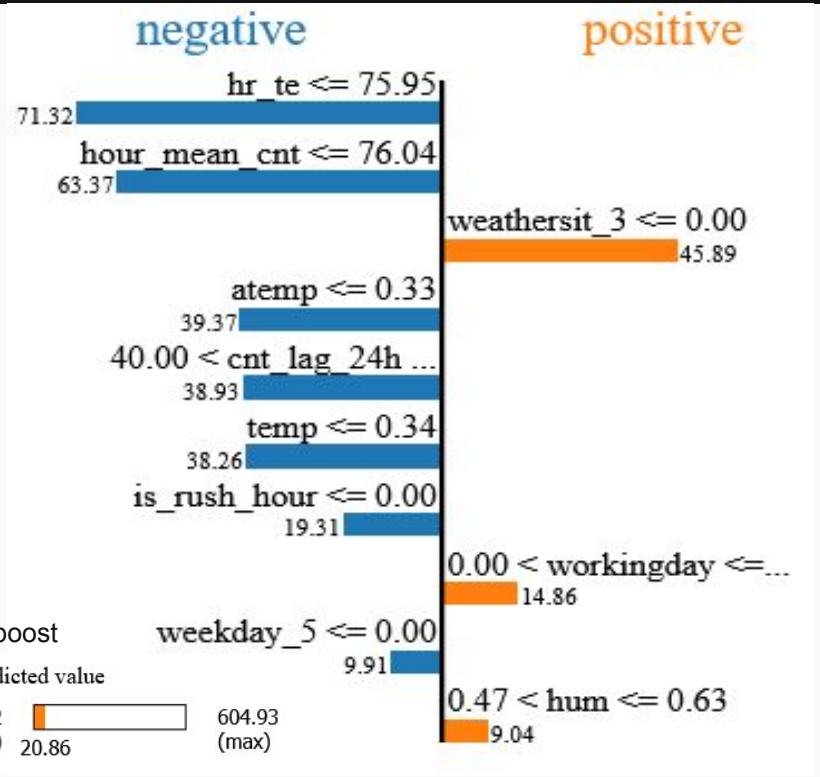
shap catboost

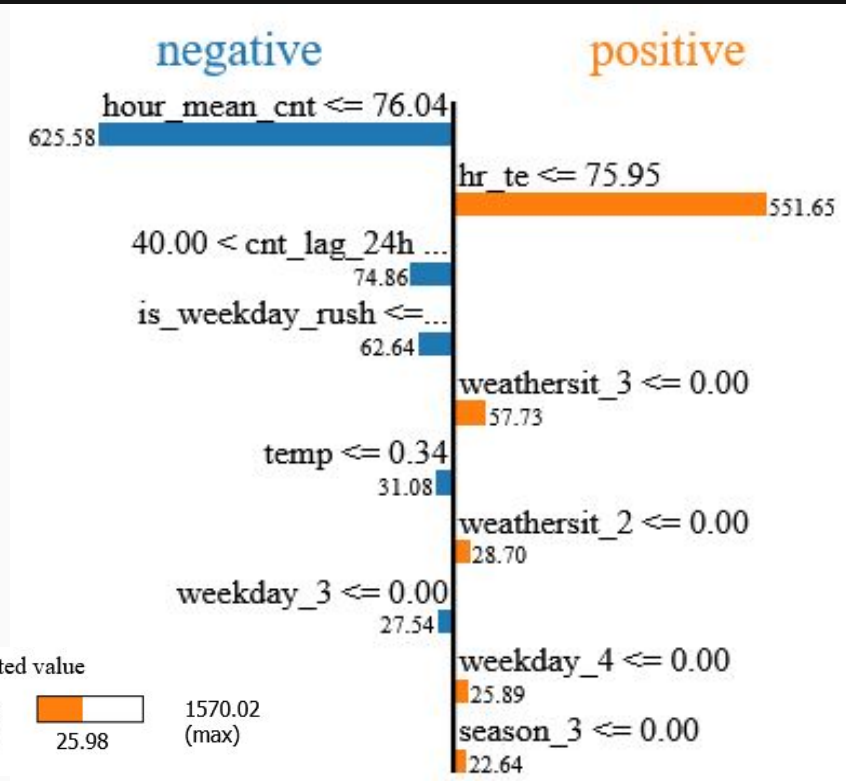


shap linreg



Локальные интерпретации LIME и SHAP

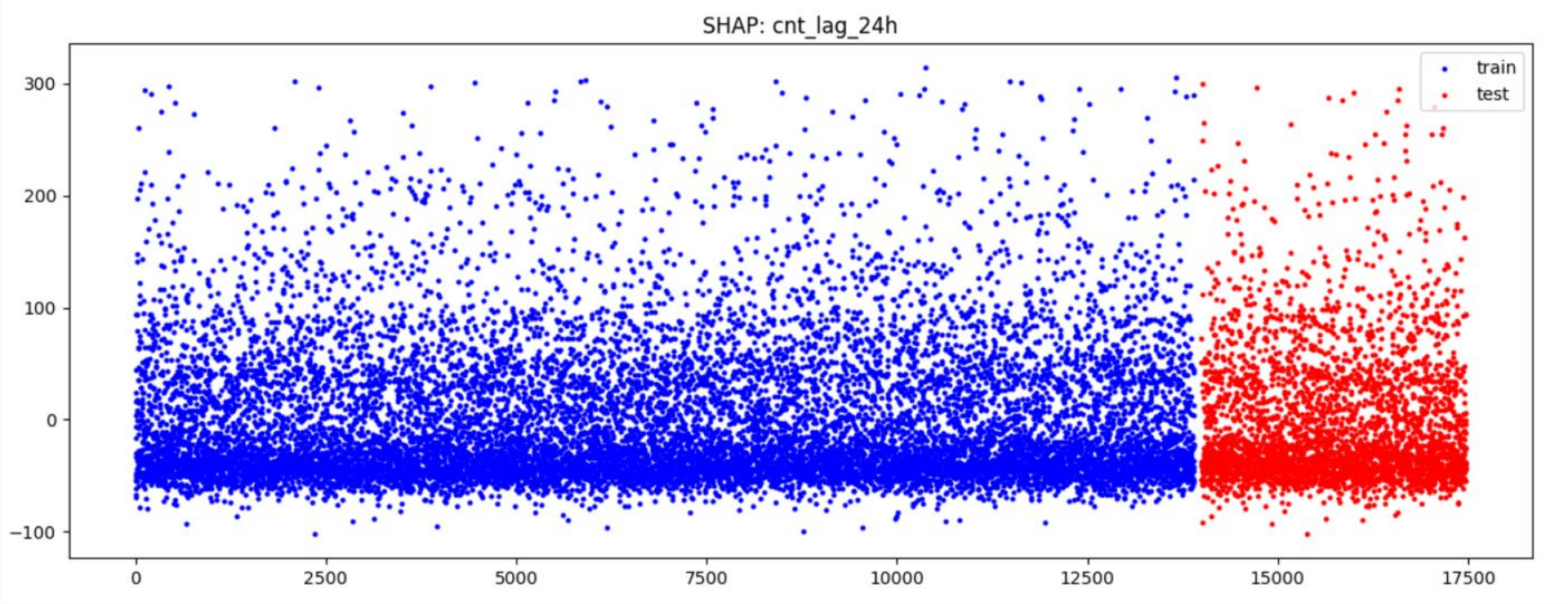




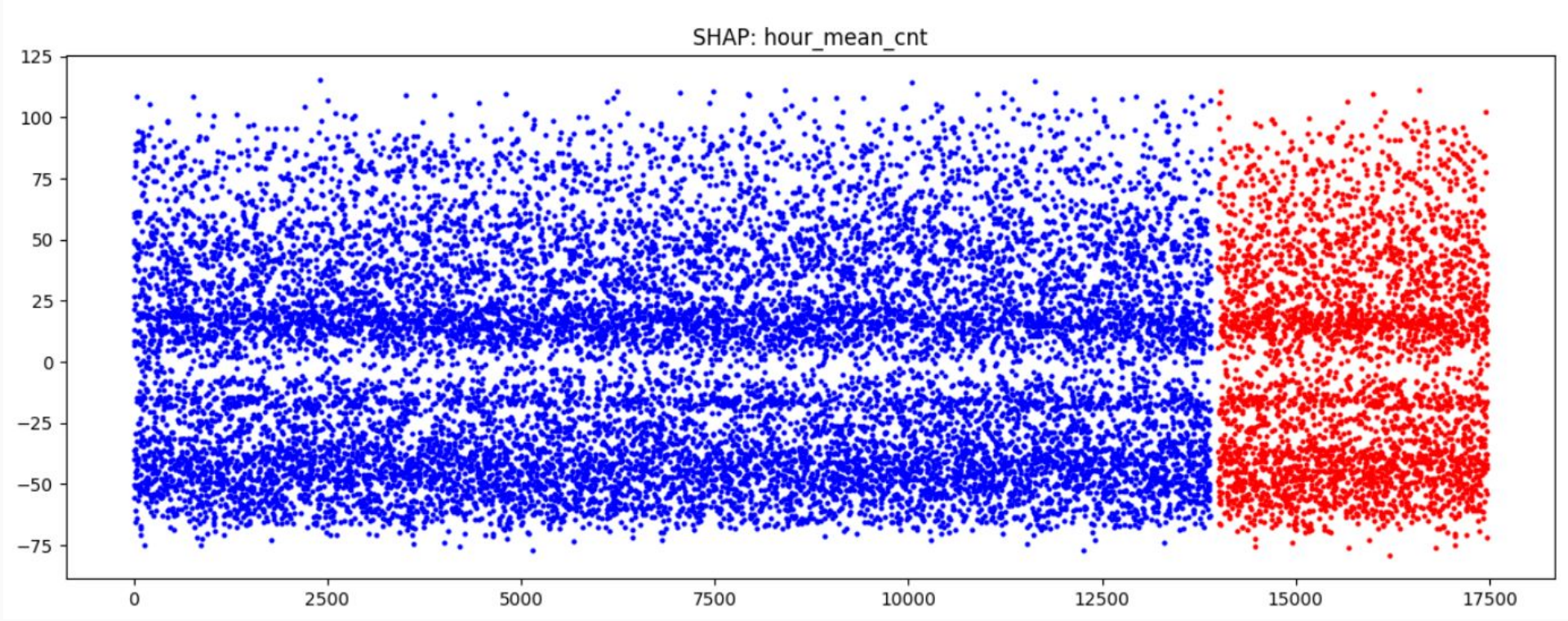
Выявление сдвигов и аномалий

	train	test	diff
cnt_lag_24h	-1.2087	-3.3847	-2.1759
hour_mean_cnt	0.1618	-0.7181	-0.8799
hr_te	0.5243	-0.1097	-0.6341
hum	0.3091	-0.0972	-0.4063
registered_ratio	-1.9673	-2.3376	-0.3703

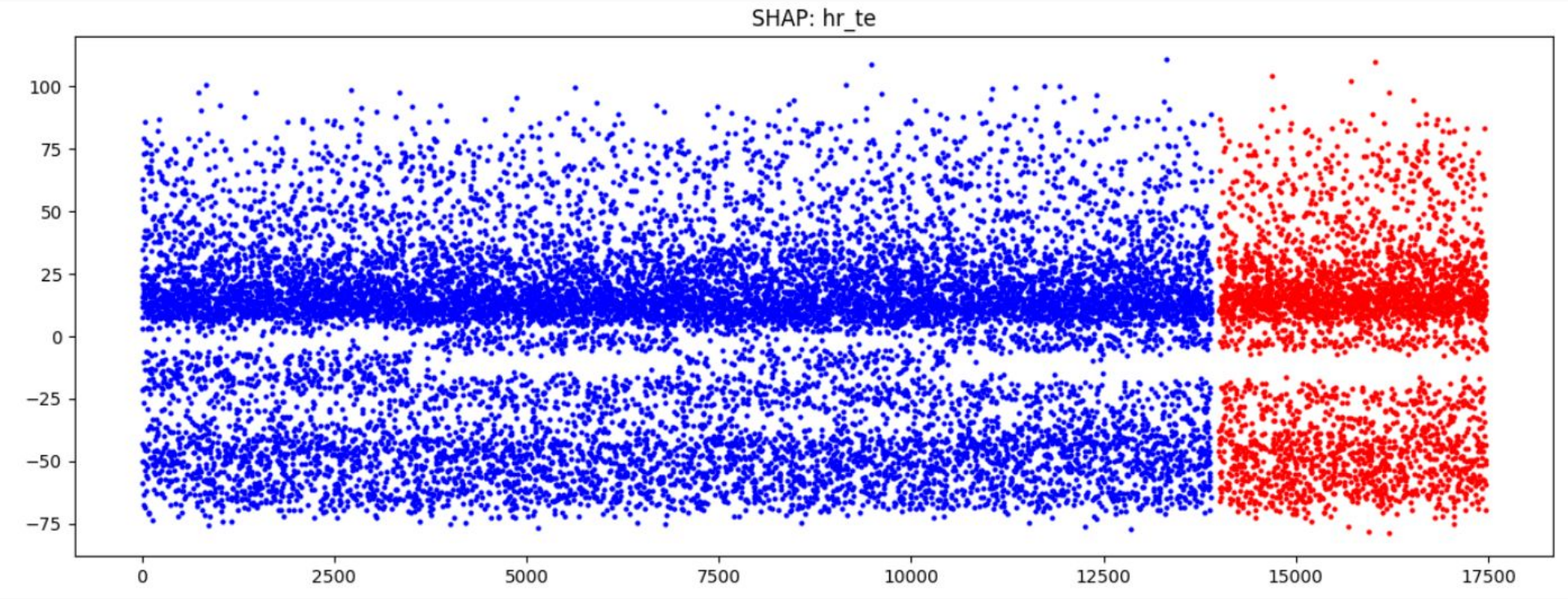
Выявление сдвигов и аномалий



Выявление сдвигов и аномалий



Выявление сдвигов и аномалий

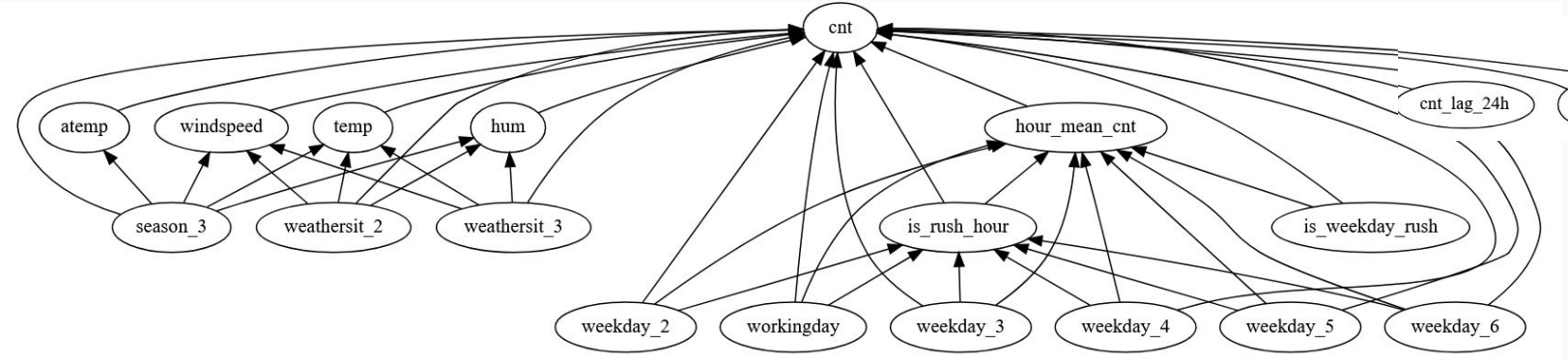


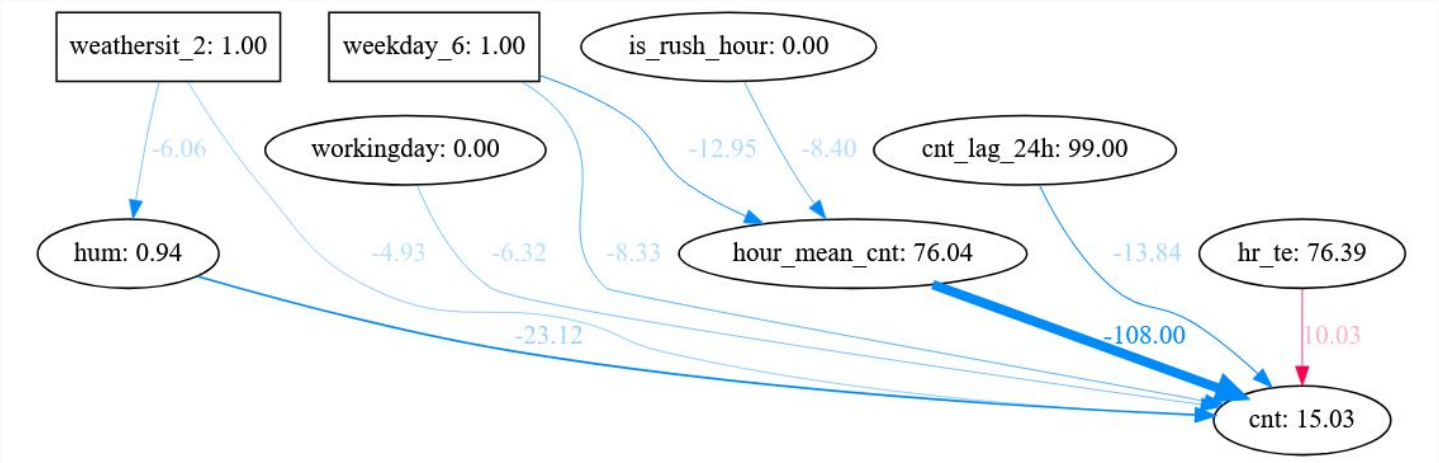
Кластеризация SHAP эмбеддингов

Кросс-валидация с SHAP эмбеддингом

```
=============== fold 1 ==========
baseline rmse: 58.3256, r2: 0.8964
embedding rmse: 55.8631, r2: 0.9050
base+embedding rmse: 55.7081, r2: 0.9055

=============== fold 2 ==========
baseline rmse: 61.5117, r2: 0.8833
embedding rmse: 58.3005, r2: 0.8952
base+embedding rmse: 57.7900, r2: 0.8970

=============== fold 3 ==========
baseline rmse: 58.5227, r2: 0.8950
embedding rmse: 56.6798, r2: 0.9015
base+embedding rmse: 56.2797, r2: 0.9029

=============== fold 4 ==========
baseline rmse: 60.3036, r2: 0.8910
embedding rmse: 55.8627, r2: 0.9064
base+embedding rmse: 56.0006, r2: 0.9060

=============== fold 5 ==========
baseline rmse: 61.3318, r2: 0.8868
embedding rmse: 58.3061, r2: 0.8977
base+embedding rmse: 57.7449, r2: 0.8997
```

Shapley Flow





Расширенный набор vs расширенный + SHAP

Расширенный набор

CatBoostRegressor с расширенной линейкой признаков показал RMSE 53.11 и R^2 0.9108

Расширенная набор + SHAP

Добавление SHAP эмбеддингов немного снизило RMSE до 52.73, а R^2 выросло до 0.9122. Это подтверждает важность дополнительных признаков.

Итоги нашего проекта

Провели EDA

Сгенерировали новые признаки и отобрали лучшие из них

Проинтерпретировали полученную модель и добавили в нее новые признаки - SHAP эмбеддинги

Улучшили изначальную модель в 3 раза.
Научились объяснять 91% дисперсии целевой переменной

С помощью нашего решения Заказчик сможет точно прогнозировать спрос на велосипеды, заранее планировать балансировку парка, оптимизировать логистику, и снизить издержки.

Вопросы