ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Прогноз спроса на аренду велосипедов (Bike Sharing)

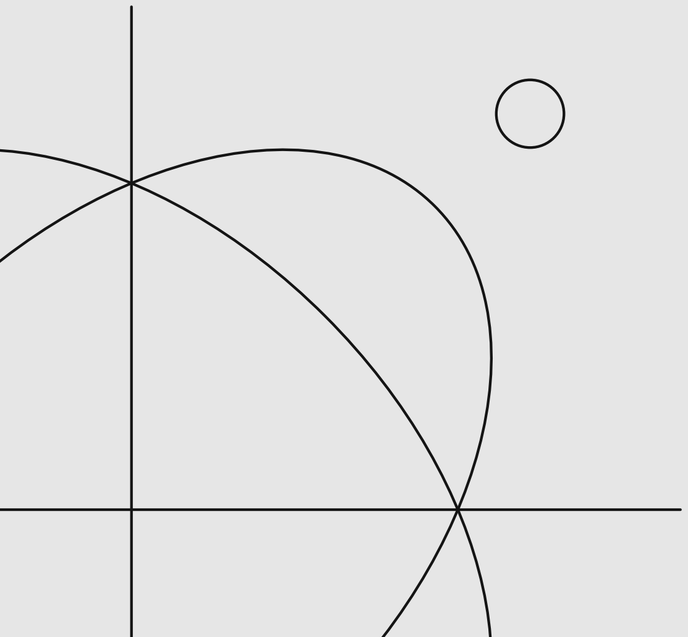
Постановка задачи и бизнес-контекст

Мы хотим построить модель, которая предсказывает количество аренд велосипедов в конкретный час на основе календарных признаков, погодных условий и характеристик дня.

Тип задачи: регрессия

Зачем это нужно в бизнес-контексте:

1. обеспечить достаточное число велосипедов в популярных локациях
2. уменьшить ситуации, когда велосипедов не хватает
3. оптимизировать работу машин, которые перевозят велосипеды
4. снизить затраты
5. улучшить сервис и пользовательский опыт



Описание датасета

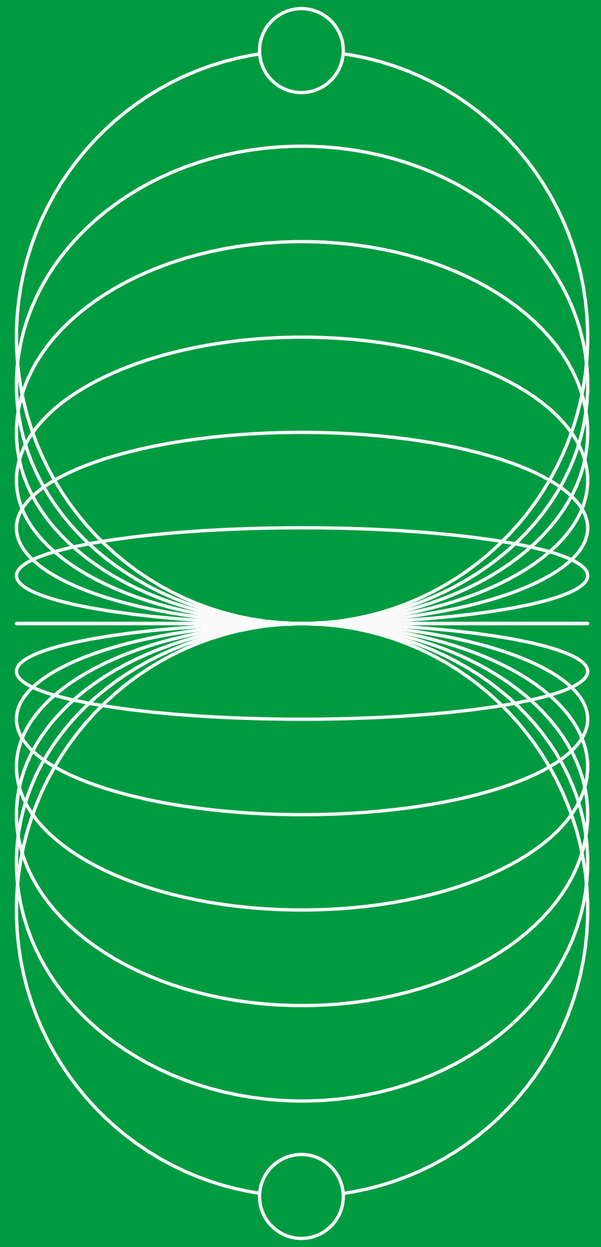
Мы используем Bike Sharing Dataset - это реальные данные из системы Capital Bikeshare (Вашингтон, США) за 2011–2012 годы.

Файлы:

hour.csv — данные по часам (17 379 строк)

day.csv— данные по дням (731 строк)

Всего 17 признаков, включая таргет



Основные признаки

01

Идентификатор и дата

instant (integer) - порядковый номер записи
dteday (date) - календарная дата

02

Календарные признаки

season (categorical) - сезон года
yr (categorical) - год наблюдения
mnth (categorical) - месяц
hr (categorical) - час суток
weekday (categorical) - день недели

03

holiday (binary) - официальный праздник или нет
workingday (binary) - рабочий день или нет

04

Погодные признаки

temp (continuous) - нормализованная температура
atemp (continuous) - нормализованная температура «по ощущению»

05

hum (continuous) - нормализованная влажность
windspeed (continuous) - нормализованная скорость ветра

Пользовательские признаки

06

casual (integer) - количество незарегистрированных пользователей
registered (integer) - количество зарегистрированных пользователей

EDA

- Проверили данные на мусор и наличие пропусков и выяснили, что пропуски мусор и пропуски отсутсвуют
- Распределение целевой переменной cnt сильно смещено влево. такая форма распределения это база для данных с неравномерной нагрузкой в течение суток и указывает на присутствие выраженной сезонности и редких пиковых значений.
- Признаки casual и registered практически полностью повторяют целевую переменную cnt, поэтому их нельзя использовать в модели, а остальные признаки имеют слабую корреляцию с количеством аренд.

Построение Baseline

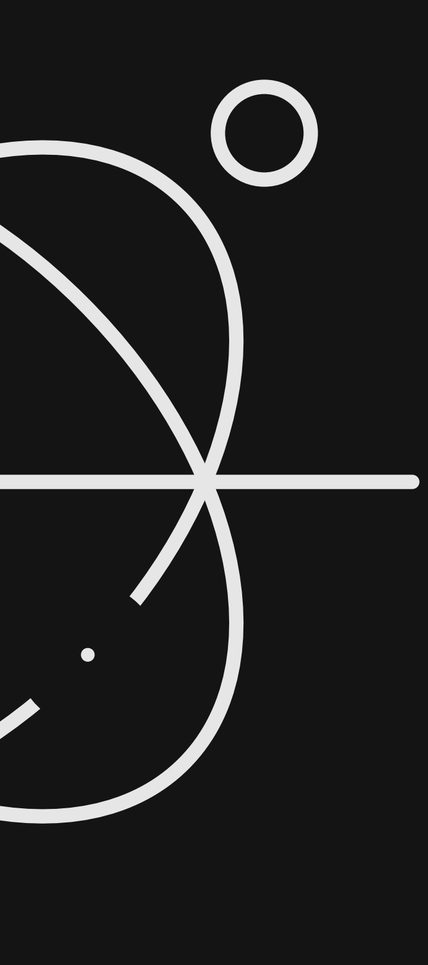
В качестве Baseline мы выбрали: LinearRegression, RandomForest и DummyRegressor и получили такие результаты на следующих метриках:

	Model	RMSE	MAE	R2
0	Dummy	178.034917	140.079835	-0.000980
1	Linear Regression	139.201123	104.795881	0.388072
2	Random Forest	40.574918	24.090368	0.948009

Работа с аномалиями и выбросами

Для поиска сложных выбросов были использованы алгоритмы Isolation Forest и LOF и оба метода выделили около 3% наблюдений как аномальные, что соответствует редким условиям в данных.

Мы выяснили, что найденные аномалии не являются ошибками данных, поэтому удаление их приведёт к потере информации о пиковых и экстремальных условиях, что ухудшит качество модели в ключевых ситуациях



Генерация пзнаков

Добавлены флаги выходных и часов пик, отражающие типичное поведение пользователей и различия в спросе в разное время

погодные условия:

- **is_good_weather, is_bad_weather** - хорошая погода увеличивает желание кататься, а плохая снижает.
- **is_cold, is_comfortable_temp, is_hot** - комфортная температура оптимальна для велопрогулок, слишком высокие или низкие температуры снижают спрос.
- **is_high_humidity, is_high_wind** - высокая влажность и сильный ветер делают поездку менее комфортной.
- **is_ideal_conditions** - сочетание всех благоприятных факторов должно максимально увеличивать спрос.
- **comfort_index** - комплексный показатель комфорта, объединяющий все погодные факторы.

взаимодействия:

- **good_weather_weekend** - хорошая погода в выходной создаёт идеальные условия для прогулочных поездок.
- **ideal_conditions_weekend** - идеальные условия в выходной должны давать максимальный спрос.
- good_weather_work_peak - хорошая погода в рабочий день в час пик влияет на использование велосипедов для поездок на работу.
- **summer_weekend, warm_month_weekend** - тёплое время года + выходной = пик использования как досуг.
- **bad_weather_workday** - из-за плохой погоды в будний день люди могут быть вынуждены использовать велосипеды.

время до пикового часа:

- **hours_to_morning_peak, hours_to_evening_peak, min_hours_to_peak** - спрос может меняться в зависимости от близости к часам пик, когда люди едут на работу или с работы. чем ближе к пику, тем выше спрос.

время до праздника:

- **days_to_holiday, days_after_holiday** - в дни перед праздником люди могут больше использовать велосипеды для подготовки или отдыха, а после праздника спрос может падать из-за усталости или возвращения к обычному ритму.
- **near_holiday** - дни около праздника могут иметь особый паттерн спроса, отличающийся от обычных дней.

сезонные признаки:

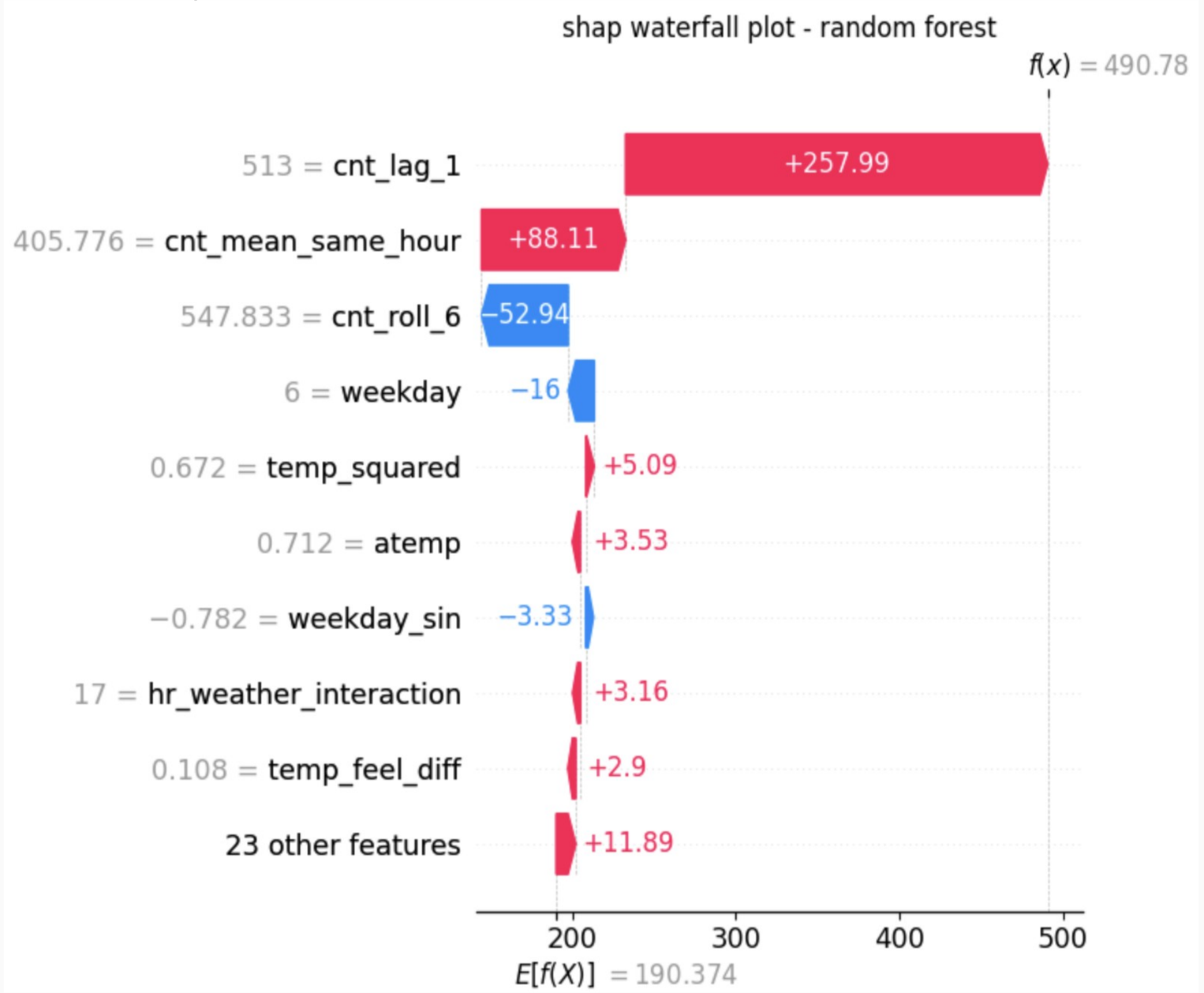
- **is_summer, is_winter** - летом спрос максимальный из-за теплой погоды, зимой минимальный.
- **is_warm_month, is_cold_month** - в тёплые месяцы (май-сентябрь) высокий спрос, в холодные (ноябрь-февраль) низкий.

время суток:

- **is_morning, is_afternoon, is_evening, is_night** - разные периоды суток имеют разную привлекательность для велопроката в зависимости от целей использования (работа, досуг).

Важность признаков

Наиболее важными оказались признаки, связанные с историческими значениями спроса (cnt_lag_1, cnt_mean_same_hour) и скользящими средними.



Вывод

По итогу оказалось, что лучшая модель была выбрана в самом начале для нашего датасета и это RandomForest, которая дает такие метрики: RMSE = 139.2, MAE = 104.8 и R2 = 0.39

Команда

Егор Шукурлаев

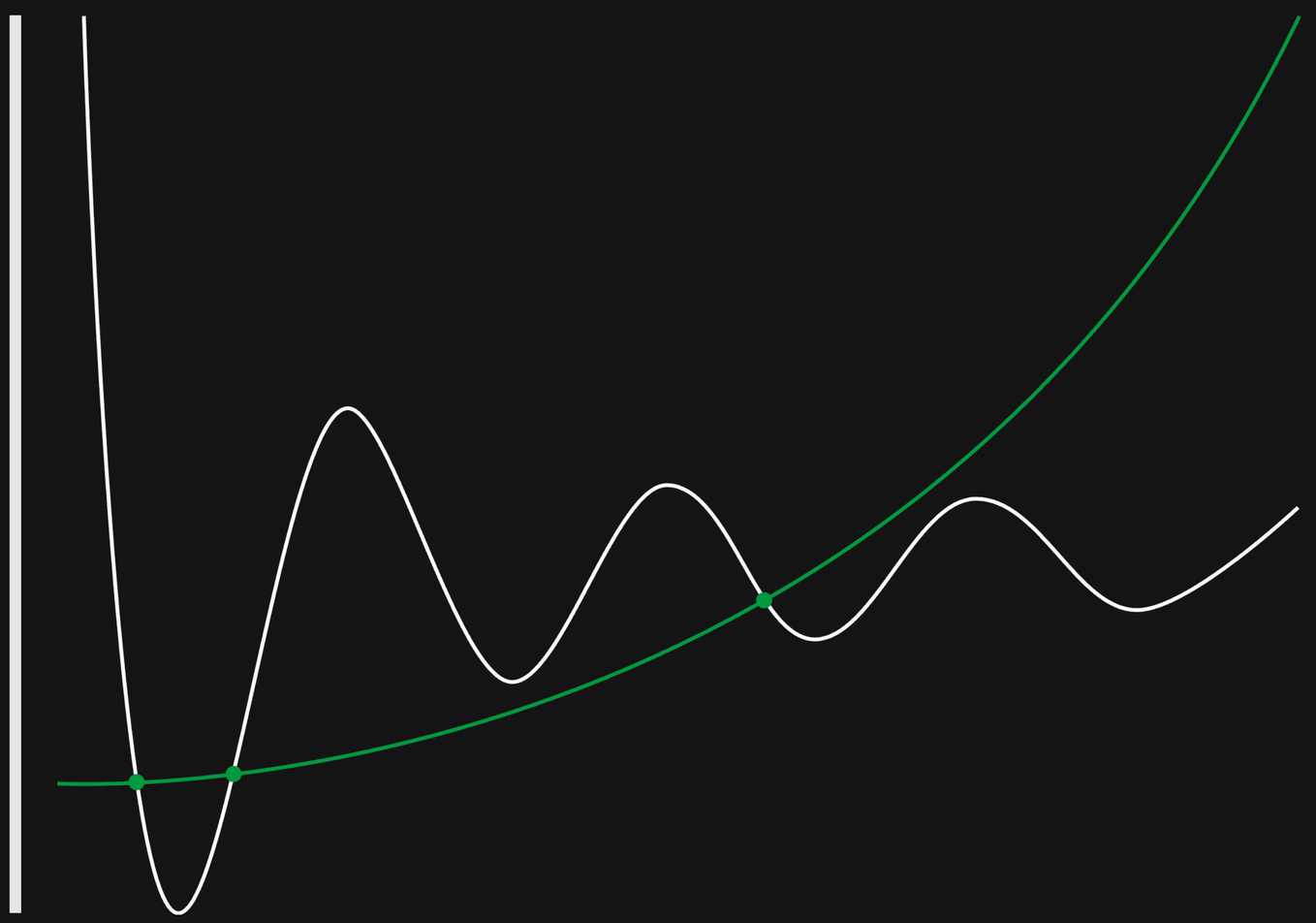
Мария Чурилова

Дмитрий Зейтц

Матвей Старых

ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Спасибо за внимание



Вопросы