BISONS

Поиск аномалий :
Автоматизация процесса определения сбоя в поставках данных

Cup IT 2024

Для оценки валидности датасета была использована легко внедряемая модель Логистической регрессии.
Для оценки модели использовались метрики ROC AUC и F1 score.

EDA

Найденные инсайты:

- Существует id, для которого не определен benchmark_name
- Сезонность в количестве уникальных выгружаемых значений отсутствует
- медианное количество бенчмарков за одну поставку - 458

ML МОДЕЛИ

- Выделили топ-10 наиболее важных признаков
- точность выбранной модели:

  ROC AUC = 0.999
  F1 score = 0.973

- Провели сравнение 5 ML моделей
- Датасет от 15 марта валиден

Описание базы данных

Количество строк: 995 004
База данных содержит передачи данных с котировками бенчмарков от ДЗО с 09-04-2018 по 15-03-2024.
То есть, каждый день подгружается отдельный пакет с котировками нескольких бенчмарков

*ДЗО— дочернее зависимое общество, организация, в уставном капитале которой Сбер имеет контрольную долю

Задача кейса

Сделать ml-модель, которая будет своевременно выявлять ошибки в поставках данных, то есть находить в них выбросы

Количество столбов : 6

benchmark_id - ID бенчмарка

benchmark_name-Наименование бенчмарка

date_of - Дата котировки

quote - Котировка

ctl_loading - Идентификатор (id) загрузки

ctl_loading_date - Дата загрузки

В базе данных есть 4 названия бенчмарков, которым соответствует несколько id бенчмарков

Пример:
бенчмарку "Technology Select Sector Index"
соответствуют индексы

"8D8AADAB-A754-4C02-9C81-0E0B4890856B"
"51D93A6F-358F-4AD9-9C70-C00BAE62B98C"

Существует id, для которого не определен benchmark_name

→ Сделали единые id для бенчмарков с одинаковыми названиями

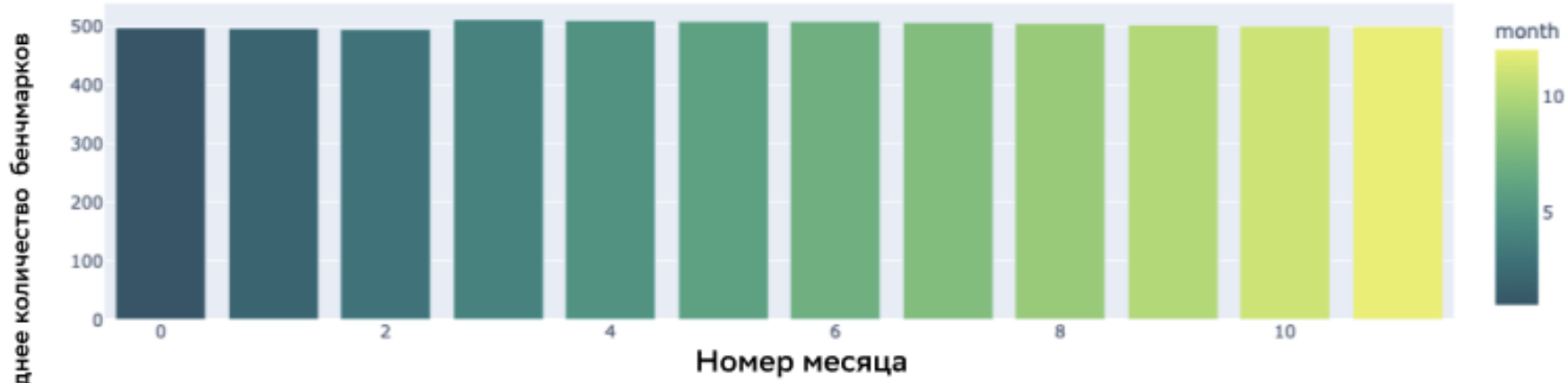
→ Для каждой поставки создали массив с id бенчмарков из этой поставки

Есть 176 дней, в которые было больше одной поставки за день из-за дополнительных поставок данных

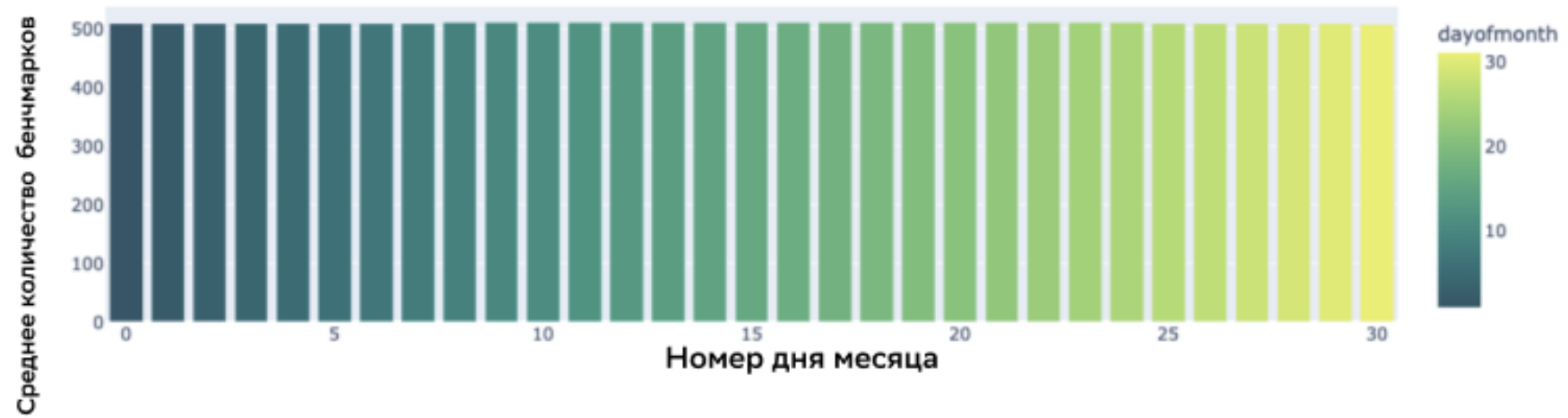
→ В эти дни были допоставки данных

→ При создании модели мы не классифицируем допоставки как ошибки

Среднее количество уникальных бенчмарков по месяцам



Среднее количество уникальных бенчмарков по дням месяца



ВЫВОД → сезонность в количестве уникальных выгружаемых значений отсутствует

Текущий механизм отгрузки котировок бенчмарков

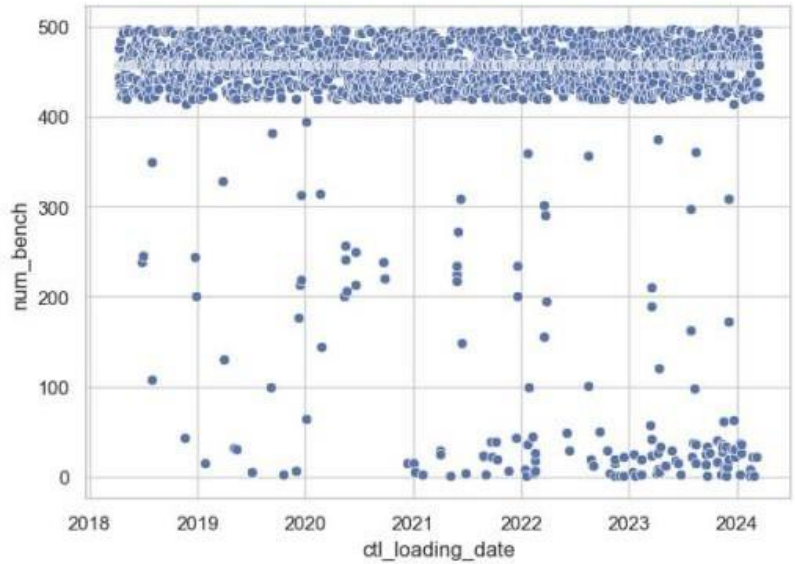
- происходит отгрузка бенчмарков
- если значение не соответствует норме (скорее всего это на текущий момент определяется вручную), отправляется запрос на доотправку пропущенных значений
- через несколько дней происходит повторная отправка пропущенных ранее данных, если выбросы заметили

ИНСАЙТ: иногда после выбросов доотправки не происходит, так как присутствует человеческий фактор)

Диапазон нормальных значений:

Мы измеряем с помощью межквантильного размаха
[403, 503]
Заметим четкую линию на уровне 458
Это медиана в нашей выборке!

Со временем количество выбросов увеличивается
Особенно начиная с 2021 года.
Это может быть связано с ростом нагрузки на систему



МЕДИАНА

458

МОДА

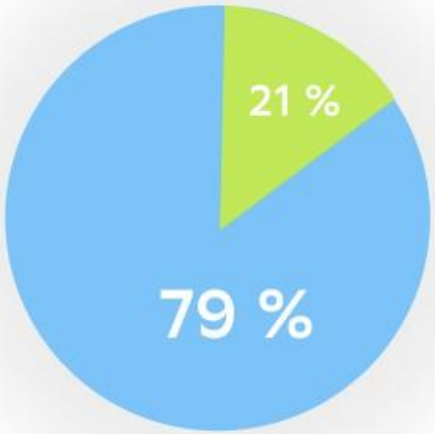
458

СРЕДНЕЕ

433

ЧТО МЫ СЧИТАЕМ ВЫБРОСОМ?

Выброс - это отклонение , которое которое не входит в размах



21% выбросов
компенсировалиотдельной поставкой

79% выбросов
которые не компенсировали отдельной поставкой

- добавили новый признак, чтобы модель отличала выброс от допоставки
- мы заметили, что медиана количества бенчмарков в рамках одной поставки на протяжении всех 6 лет - это 458.
- можно предположить, что существует всего 458 бенчмарков, по каждому из которых каждый день как минимум один раз должна выгружаться котировка

- если значение бенчмарков в рамках одной поставки совсем маленькое через несколько дней происходит допоставка --> в сумме получается 458 (+- 10)
- при этом наша команда нашла случаи когда есть некритичные недопоставки, и в таком случае допоставок не происходит, что может негативно отразиться на бизнес-процессах.

ML модели для предсказывания кол строк в поставке

Постановка задачи:
бинарная классификация пакетов

Пайплайн модели

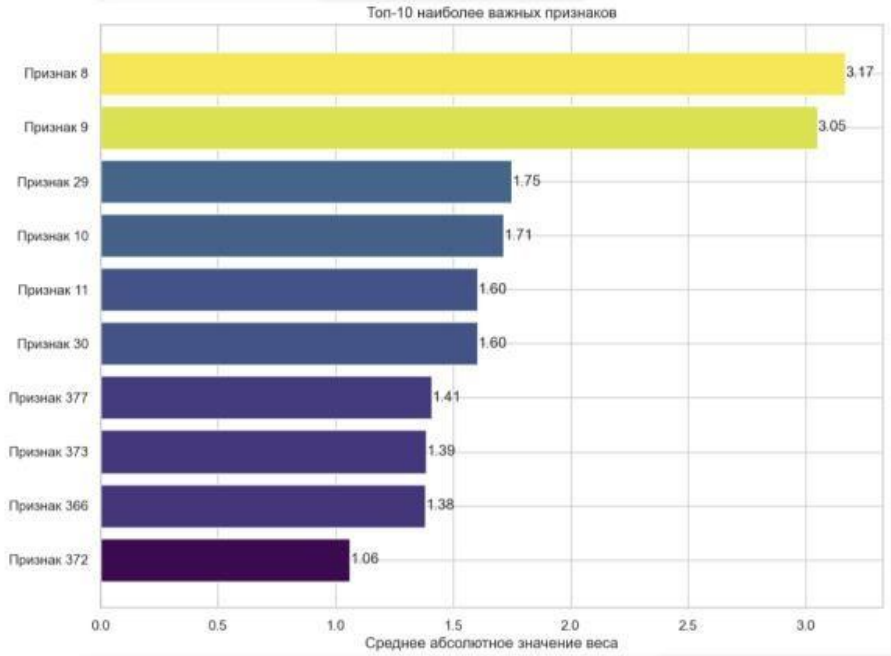
1. Разметка данных
   таргет:
   1 - Выброс
   0 - Нормальная поставка
   или Допоставка
2. Выбор метрики
   Метрика для выбора наилучшей модели:
   - ROC AUC
   - F1 score
3. Разбиение на Train/Test
4. Выделение признаков для модели
5. Кодируем категориальные признаки (OneHotEncoder)
6. Масштабирование численных признаков (StandartScaler)
8. Проверка модели на тестовой выборке
9. Оценка

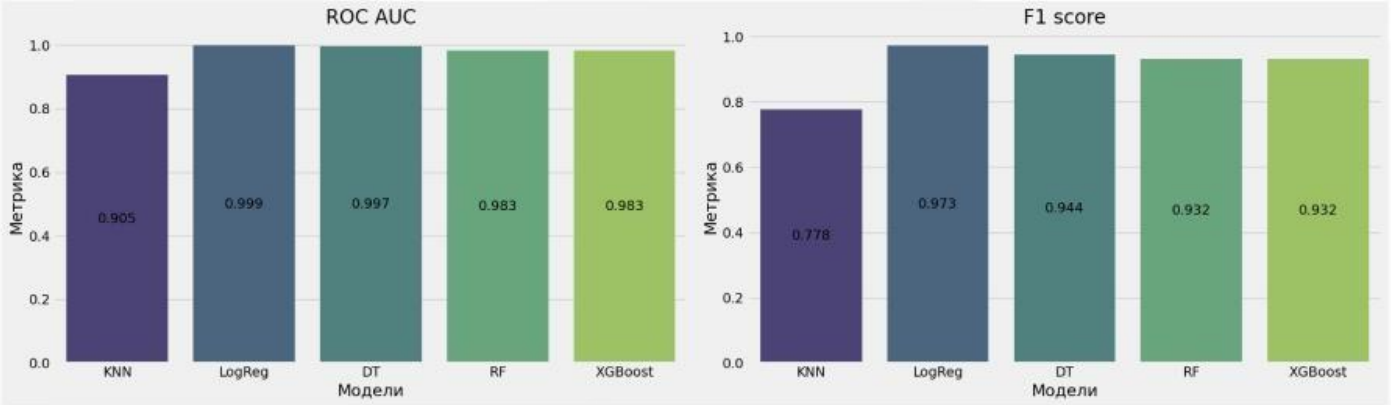
Выбранная модель

- Логистическая Регрессия

Оптимальный гиперпараметр C = 2,78

Плюсы модели:
- Скорость работы
- Простота реализации
- Устойчивость к переобучению





$$p_+ = \frac{1}{1 + exp(-x^T w)} = \sigma(x^T w)$$

Вероятность выброса

Вектор Весов
(значения см. на графике выше)

Валидность датасета от 15 марта

Да

colab.research.g…

Google Colaboratory

https://colab.research.google.com/drive/1pRhzQ1Hy3Pc3QOi_k_zM9FDiH7m5T_zF?usp=sharing

Команда

Алиев Руслан

Аналитик
НИУ ВШЭ
Бизнес-Информатика

Ванин Виктор

Аналитик
НИУ ВШЭ
Бизнес-Информатика

Ковалева Дарья

Аналитик
НИУ ВШЭ
Бизнес-Информатика

Панкова Злата

Аналитик
НИУ ВШЭ
Бизнес-Информатика

Кислова Алена

Дизайнер
B&D
Цифровой дизайн