

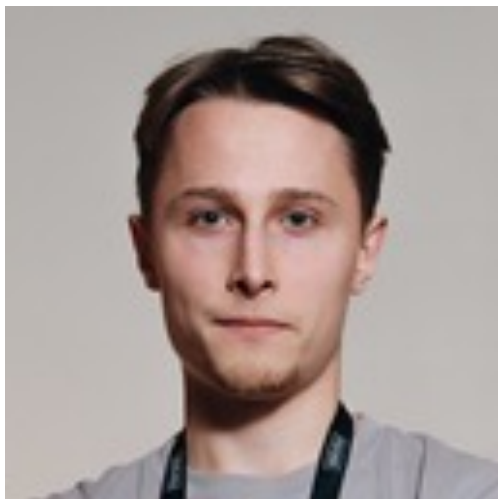
DataWagon

ПГК Оракул

Команда ETNA

Александр Чиков, Малышев Яков

Команда ETNA



Саша Чиков



Малышев Яков

Кто мы?

Мы ML специалисты, разрабатывающие open-source библиотеку для работы с временными рядами

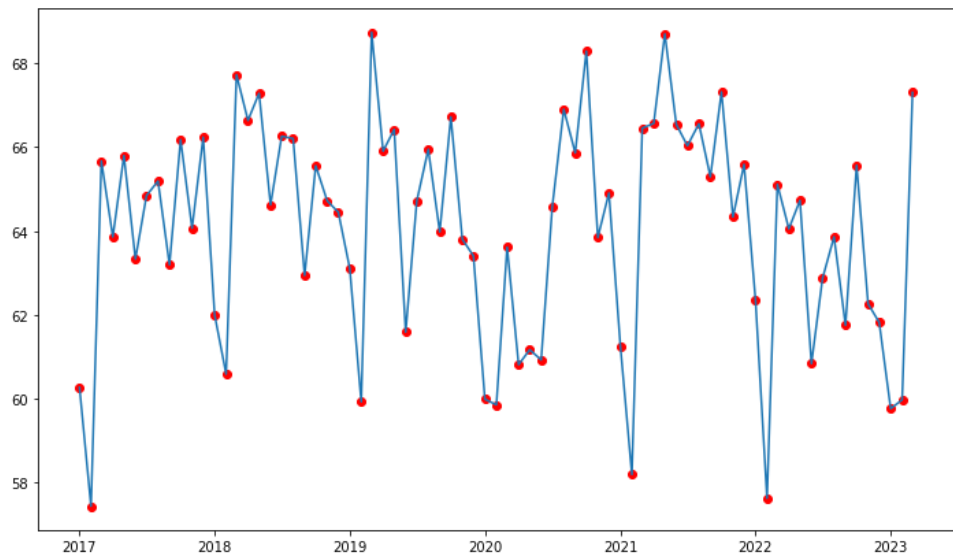
Проблема

- Для эффективной работы компании необходимо с высокой точностью прогнозировать спрос на грузоперевозки
- Сложность задачи:
 - большое количество рядов
 - необходимость оценивать вероятность появления новых рядов
 - нерегулярные ряды
 - ряды разного уровня

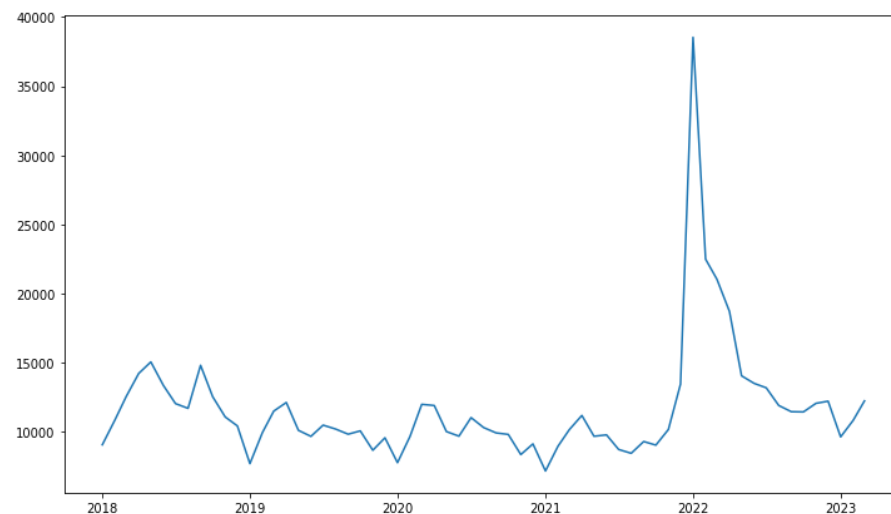
EDA

Некоторые наблюдения:

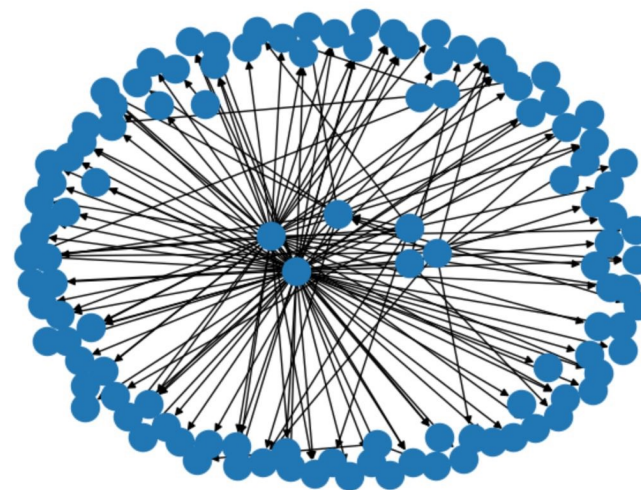
- каждый месяц появляются около 10-12 тыс новых рядов
- верхнеуровневый ряд можно прогнозировать с относительно низкой ошибкой
- Существуют станции-хабы (из них много возят)



Ряд всех грузоперевозок в млн. тон






Количество новых рядов



Граф активности конкретного клиента со станциями

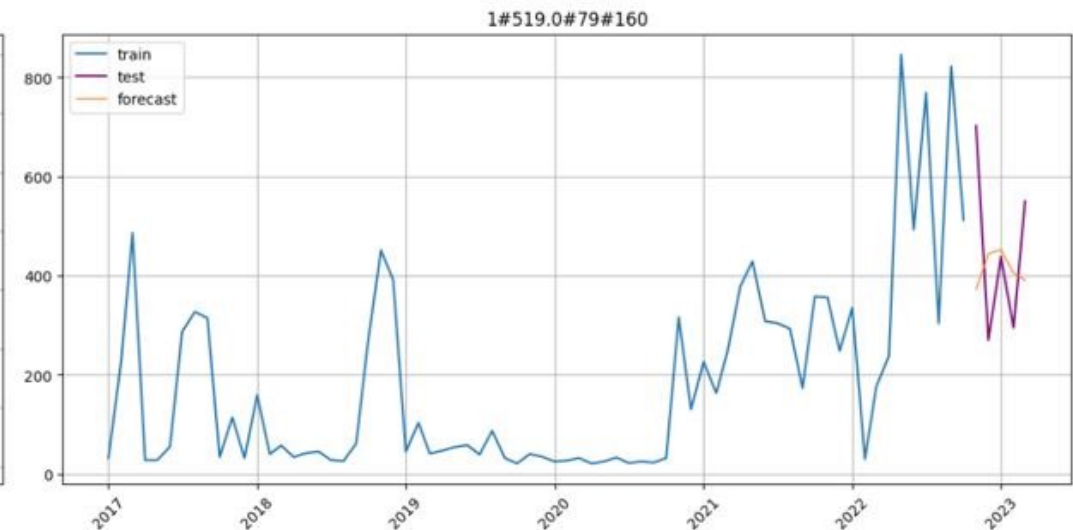
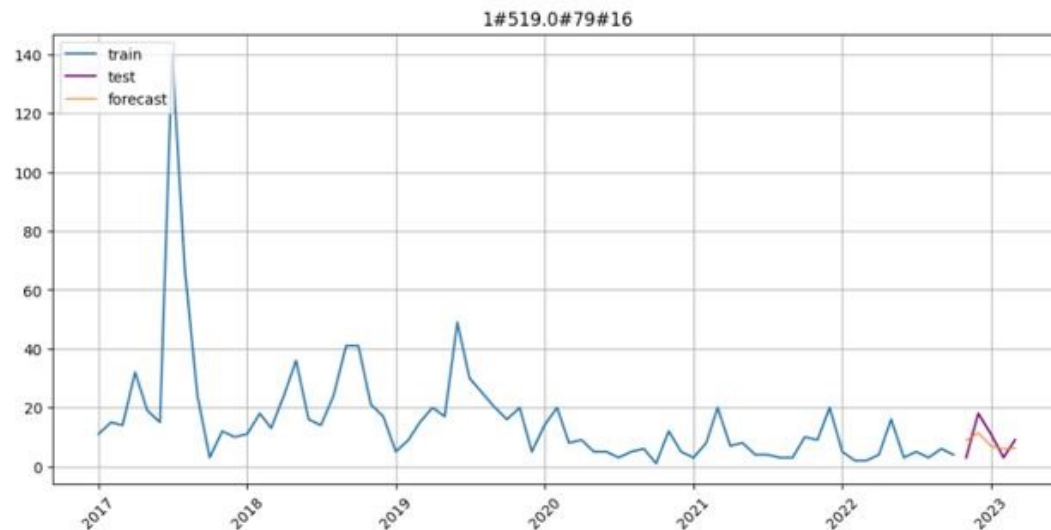
ETNA

Подходы

- Ансамбль из простых моделей SMA (для каждого ряда своя модель, с самого низкого уровня – более 1 млн. рядов) 
- Линейная регрессия (для каждого ряда своя модель, со среднего уровня ~70 тыс. рядов) 
 - Сезонные компоненты
 - Лаги
- модель CatBoost с большим количеством фичей (одна большая модель, со среднего уровня ~70 тыс. рядов) 
 - Фичи на основе графов
 - Сезонные компоненты
 - Лаги
 - Кодирование холдинга

Плюсы нашего решения

- Решение обрабатывает более 1 млн. рядов с самого низкого уровня (расчет занимает короткий промежуток времени)
- Легко интерпретировать прогнозы
- Легко агрегировать прогнозы до нужного уровня
- Из зависимостей лишь одна библиотека: etna



Шаги по улучшению

- Доработка модели CatBoost для улучшения качества прогнозов
- Проработка алгоритма перехода от верхних уровней к нижним (матрица весов)
- Разработка модели, оценивающей вероятность появления ряда в будущем
- Перейти от прогноза вагонов к прогнозу весов грузов

Спасибо за внимание