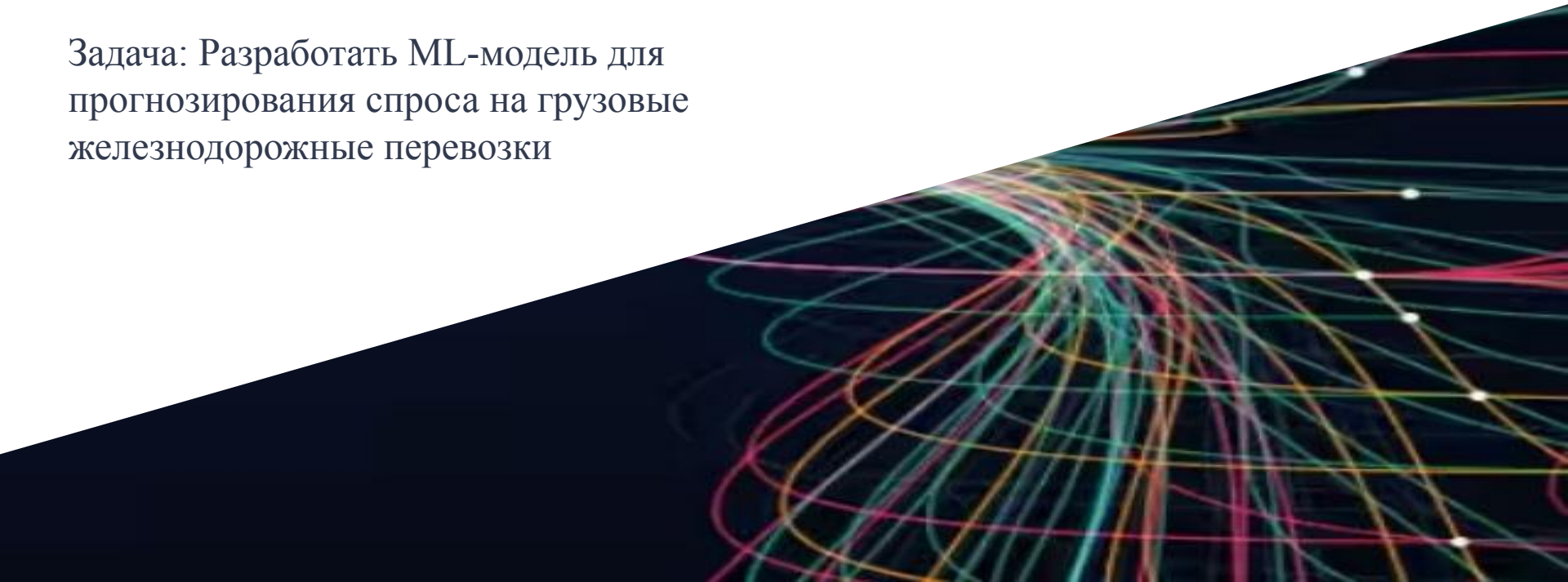
Трек 1 - ПГК Оракул / Прогнозирование спроса на грузовые ЖД перевозки

Задача: Разработать ML-модель для прогнозирования спроса на грузовые железнодорожные перевозки



Команда Cargo Data Explorers

1. Михаил Грибанов - @gribanov_m
2. Меркулов Артем - @Tesla2060
3. Ирина Балычева - @IrinaBalycheva
4. Алла Мишра - @allasr
5. Ислом Алиев - @islomchick

Основная проблема и ее решение.

- Основная проблема заключается в необходимости прогнозирования спроса на грузовые железнодорожные перевозки для эффективного распределения ресурсов компании.
- Ее решение заключается в построении модели с применением библиотеки ARIMA, CatBoost и Orbit которые решают задачи прогнозирования временных рядов.

- Решение, стэк
- Интересные фичи, наблюдения

Стэк: Python, Numpy, Pandas, Matplotlib, Cat Boots, ARIMA.

Решение: для решения задачи мы делали скоринг данных и проводили эксперименты на 3-х моделях: CatBoost, ARIMA, Orbit. В качестве целевых данных мы взяли real_weight и real_wagon_count.

Наблюдения:

- количество вагонов по времени
- вес вагона по определенным периодам времени
- влияние типа вагона на его вес
- направление путей отправки и прибытия состава.
- Лучшая модель получилась при проверке 3 гипотезы - обучении данных на временных периодах, при помощи CatBoost
- Хорошие результаты дал подход на разбиение на 4 части датасета по загруженности пунктов отправления: хорошо показала себя модель ARIMA, но к сожалению, не успели досчитать 4 датасет в 2 миллиона, нехватило ресурсов.
- Orbit мы не успели прогнать из-за нехватки времени.

Решение и выводы

- **Гипотезы:**
  - на старте было 3 гипотезы:
  - 1 - разделение датасета по загруженности станций отправления разделили на 4 части и обучить 4 модели по этим датасетам, затем делать предсказание с помощью моделей соответствующих групп.
  - 2 - разделение по направлениям поездов с точки зрения популярноти и обучить модели по этим датасетам, затем делать предсказание с помощью моделей соответствующих групп.
  - 3 - провести обучение по всему датасету по периодам.
- **Результаты:** По итогу проверки гипотез получили следующее:
  - **Лучшая модель 3 гипотезы** - обучении данных на временных периодах, при помощи **CatBoost**

Внедрение и масштабирование

- Насколько решение внедряемо в реальные бизнес-процессы?

Решение по внедрению ML-модели для прогнозирования спроса на грузовые железнодорожные перевозки является очень внедряемым в реальные бизнес-процессы.

- Насколько решение масштабируемое?

Внедрение ML-модели для прогнозирования спроса на грузовые железнодорожные перевозки обладает хорошей масштабируемостью, что позволяет справляться с большими объемами данных и адаптироваться к изменяющимся требованиям и условиям бизнеса.

Проблемы в ходе написания модели и дальнейшее улучшение модели

- Что планируете изменить/заменить/улучшить в дальнейшем?

1. Доделать решение на ARIMA.
2. Проверить как ведет себя ARIMA при подходе разбиения на направления.
3. Была мысль предсказывать все в тоннах и по нему рассчитывать количество вагонов, но не хватало данных.
4. Можно поработать с аномалиями и чисткой данных.

- Чего не хватает для реализации продукта и дальнейшего масштабирования?

1. доступа к данным типов вагонов, незашифрованных, это бы ускорило результаты анализа и группировки
2. формул расчета грузоподъемности по типам вагонов, чтобы проверить наш подход предсказаний по весу грузов. По нему высчитывать количество вагонов.
3. ресурсов;
4. времени, что-то мы могли бы улучшить сами, было бы чуть больше времени,
5. технических ресурсов - Арима на 2млн вылетает, но было бы время можно было бы подумать что сделать.

Код github: https://github.com/islomchickk/CargoDataWagon.git