«Прогнозирование отправления вагонов в ремонт»

Содержание

1 Введение

Прогнозирования даты отправления вагона в плановый ремонт было одной из задач хакатона «Data Wagon 2023» от Первой Грузовой Компании (АО «ПГК»). Работу курировали и оценивали эксперты ПГК Диджитал.

Отправка вагона в плановый ремонт может происходить по разным причинам - как по регламенту (срок/пробег), так и из-за того, что накопились мелкие дефекты(было много текущих ремонтов), не было вариантов на погрузку и т.д. Этих причин много, и все они влияют на возможность осуществления ремонта.

2 Постановка задачи

Требуется создать ML-модель прогнозирования даты отправления вагона в плановый ремонт. Авторы хакатона предоставили несколько типов данных о вагонах:

- Данные по характеристикам вагона
- Данные по текущим ремонтам вагона
- Информация по дислокации
- Данные по плановым ремонтам
- Справочник грузов
- Справочник станций

Для удобства, исходные данные и тагрет представлены в виде блок схемы: Рис. 1

Такое количество призников будет неэффективно использовать в "сыром"виде, поэтому основная часть работы заключается в их анализе и построении "полезных"признаков. Дальнейшее исследование заключается в построении ML-модели для решения задачи.

Таргета в задаче два: отправление вагона в плановый ремонт "в текущем месяце"и "в течении 10 дней". Если вагон отправляется в плановый ремонт в течении 10 дней, это значит, что он также отправляется в плановый ремонт и в текущем месяце. Таргеты принимают значения 1 – отправление в ремонт, 0 – ремонт вагону не требуется. Таким образом, имеем задачу бинарной классификации для двух таргетов. Критерием успешного

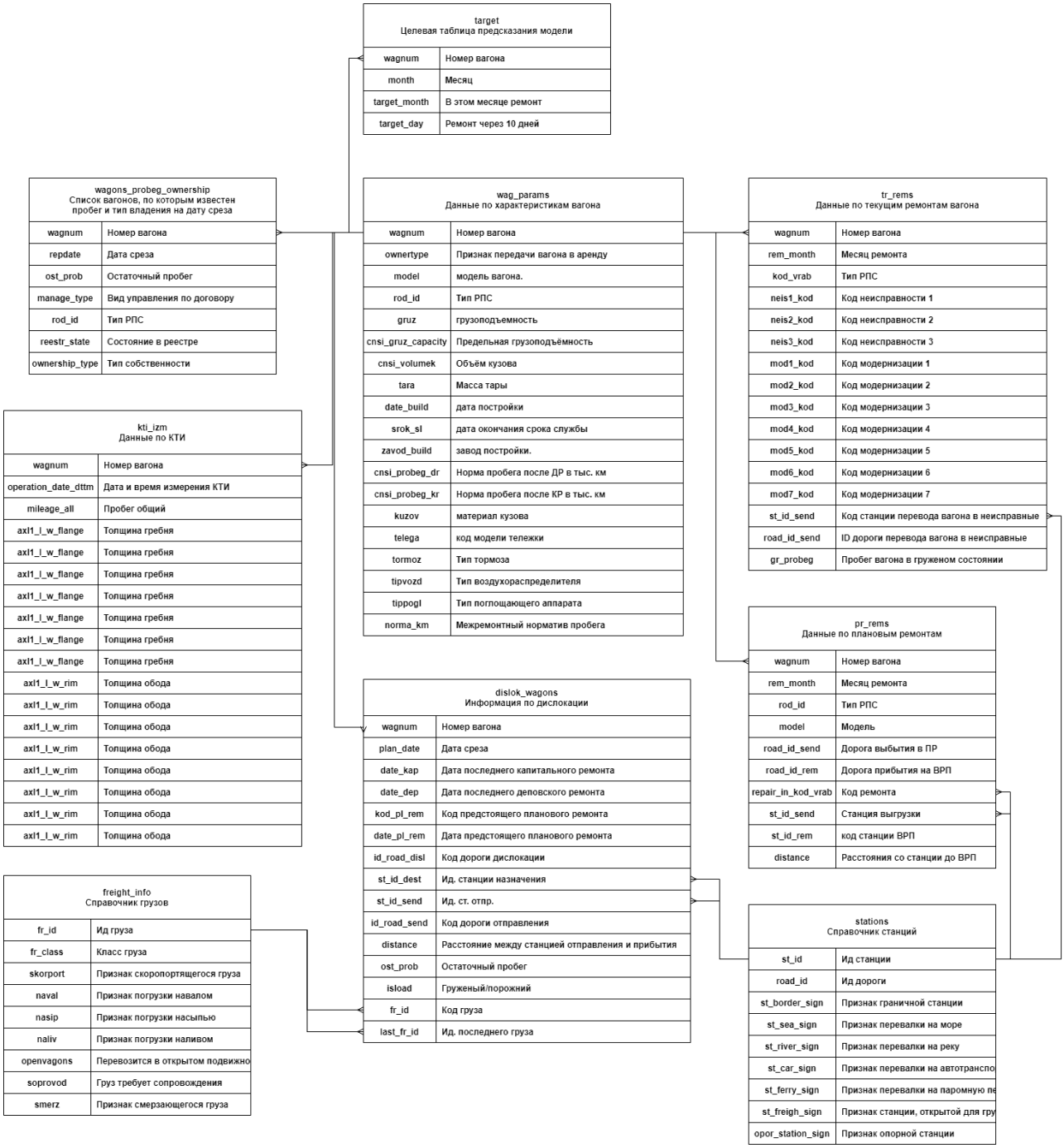

Рис. 1: Набор данных о вагонах и target предсказания модели

решения задачи является достижение наилучшего качества модели по средней среди таргетов F1-мере в сравнении с другими участниками соревнования и преодоления базового решения с качеством 0.23 по средней F1-мере.

3 Построение признаков

В данных имеется информация о вагонах за промежуток времени с августа 2022 года по февраль 2023 года. Также имеется информация о характеристиках вагонов, перевозимых ими грузов и станциях, в которых останавливались вагоны. Всего имеется 34000 вагонов, для которых необходимо предсказать тагргет для марта 2023 года. Признаки строятся и добавляются друг за другом, что позволяет нам оценить важность каждого признака (влияние на результат). Каждый признак либо увеличивал, либо не менял качество модели.

3.1 Дислокация

В данном источнике данных (dislok_wagons) отображена информация о дислокации вагонов за каждый день с августа по февраль. Признаки представляют из себя временные ряды. По ним считаются определённые статистики за последний месяц:

1. суммарный путь вагона; среднее и дисперсия пройденного пути вагона за день.
2. количество дней, когда вагон перевозил груз; путь, который вагон прошёл нагруженным; доля пути, пройденный с грузом от общего пути.

Вагон в предыдущие месяце мог испытывать экстраординарную нагрузку. Это может сыграть свою роль в принятие решения отправить вагон на ремонт заранее. Статистики помогают находить такие случаи. Также они помогают находить случаи, когда обновление пробега и постановка метки об отправке на ремонт происходят в разное время. С такими случаями помогает бороться время до предыдущего ремонта.

Кроме того, у каждого вагона есть своя "официальная"дата планового ремонта. Дело в том, что для планового ремонта регламенты периодичности проведения восстановительных работ введены на законодательном уровне. При погрузке вагона специалисты проверяют его техническое состояние, остаточный пробег, срок планового ремонта. Если остаточный пробег вагона составляет, например, меньше 500 км, то владелец обязан отправить его в плановый ремонт. Таким образом, этот признак является некоторым базовым решением задачи, использовавшийся ранее в грузоперевозках. На основе этой информации, строится признак ближайшего планового ремонта вагона в будущем. Также среди признаков

брался остаточный пробег, который выставлялся вагонам после последнего ремонта, код предстоящего ремонта.

Использовался признак станции отправления (последняя посещённая станция). Он является географическим признаком. Владельцы склонны отправлять вагоны в ТО, когда запас хода все ещё достаточно велик: 20-50 тысяч км. Это может быть связано с тем, что где-то ремонтироваться дешевле. И даже находясь рядом с какой-нибудь ремонтной станцией, при достаточном запасе хода вагон выгоднее отправить туда, где дешевле или гарантировано произведётся более качественный ремонт.

Информация по дислокации в итоге принесла наибольший прирост качества.

3.2 Характеристики вагонов

Характеристик вагона очень разнообразны: от характеристик отдельных деталей, до возраста и количества ремонтов. Есть характеристики присуще вагону с создания, а есть характеристики зависящие от времени. Среди них выделим следующие признаки:

1. Возраст вагона – вычисляется по дате постройки вагона.
2. Время до истечения срока службы (присвается вагону при построении)
3. Количество ремонтов вагона
4. Вид управления по договору
5. Время до крайнего ремонта

Самыми важными признаками оказались возраст вагона, время до истечения срока службы и время до его крайнего ремонта. Срок службы оказался самым важным признаком. Время до крайнего ремонта совместно с пройденным путём дают понимание модели, как долго вагон используется без ремонта.

Также построены признаки отвечающие за типы перевозимых грузов вагонами: для каждого типа груза вычислялась доля пути, пройденная вагоном нагруженным этим грузом от общего пути, который вагон проехал нагруженным. Признаки грузов, однако оказались не столь важными для результата.

4 Модель

Выпишем явно метрику:

$$F1 = \frac{precision\_month * recall\_month}{precision\_month + recall\_month} + \frac{precision\_day * recall\_day}{precision\_day + recall\_day}$$

Для построения решения использовались градиентные бустинги LightGBM и Catboost. Всего исследовалось несколько вариантов построения модели:

1. LightGBM и Catboost с параметрами по умолчанию
2. LightGBM с подбором оптимальных параметров
3. Ансамбль семи LightGBM на разных сидах
4. Ансамбль семи Catboost на разных сидах

Подбор параметров происходит с помощью библиотеки optuna с количеством итераций равным 500. Следующим шаг – увеличение количество итераций в 10 раз и уменьшением темпа обучения в 10 раз – этот приём принёс прирост в качестве.

5 Результаты

Самыми значимыми признаками оказались:

- Накопленные статистики вагона за предыдущий месяц.
- Станции отправления.
- Время с последнего ремонта.
- "Официальная"дата планового ремонта.

Данные признаки давали вплоть до 94% вносимого результата. Остальные признаки давали оставшиеся 6% конечного результата.

Результаты представлены в таблице 1

Модель	F1 month	F1 day	F1 mean
LightGBM	0.622	0.521	0.572
Tuned LightGBM	0.686	0.620	0.653
Ensemble LightGBM	0.692	0.673	0.683
Catboost	0.681	0.621	0.651
Tuned Catboost	0.693	-	0.657
Ensemble Catboost	0.686	0.625	0.656

Таблица 1: F1 мера для моделей

6 Заключение

Данное решение оказалось лидирующим, и мы заняли первое место в соревновании. Решение используется аторами хакатона в своих продуктах: [3], [2]. Дальнейшая работа – сделать более сложные ансамбли для прогнозирования планового ремонта и извлечь больше признаков из данных: например, никак не использовались id дорог и типы станций. Также возможным продолжением работы является построение метрических признаков для повышения качества решения.

Список литературы

[1] DataWagon: https://datawagon.ru

[2] Оптимизатор ремонтов: https://habr.com/ru/companies/pgk/articles/781928

[3] Результаты хакатона: link

[4] Github: https://github.com/makriot/DataWagon