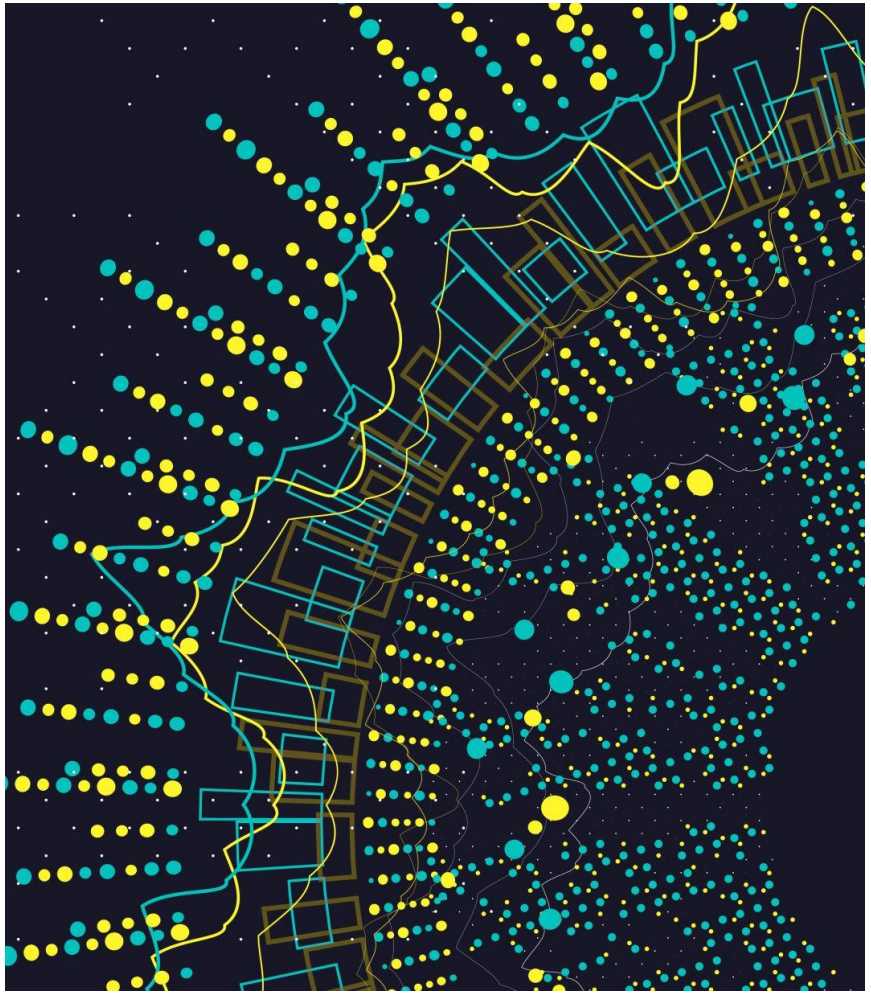
No Loss November

Трек №2. Чек-ап вагона.



Много табличек

• Соревнование содержало умопомрачительное количество табличек, которое тяжело не только обработать за 1.5 дня, но и понять их суть и влияние на конечный результат, и это команде аж из 5 человек. Мы понимали, что построить сложные ансамбли мы не успеем, поэтому мы сосредоточились на EDA и Feature engineering.

Фичи и инсайты – наш главный продукт за эти 2 дня

• Мы использовали достаточно **простые** алгоритмы машинного обучения с некоторыми стандартными приёмами спортивного ML: 2 LGBM, по одному на каждый таргет с усреднением по семи seed’ам. Подбор параметров осуществлялся через optuna. Валидация стандартная, 5 фолдов, контроль на феврале. После подбора количество деревьев увеличивалось в 10 раз, а learning rate понижался в 10 раз.

• **Основным же драйвером роста оставался feature engineering**.

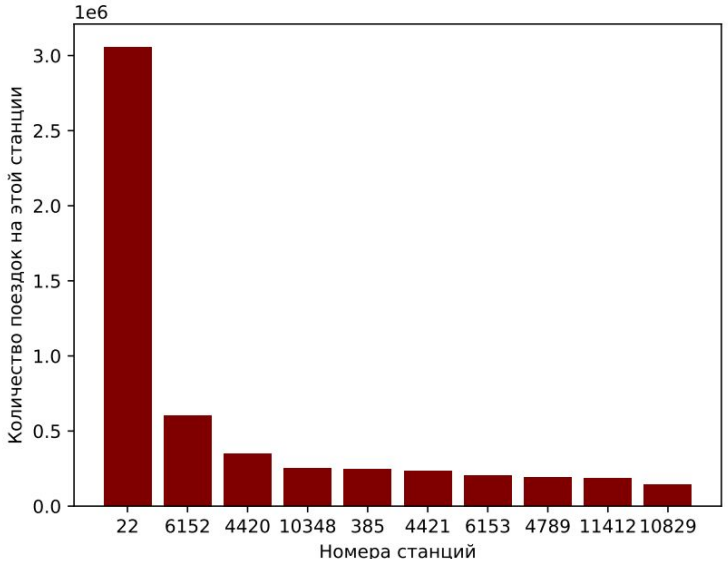
Таблица dislok принесла нам наибольший прирост качества

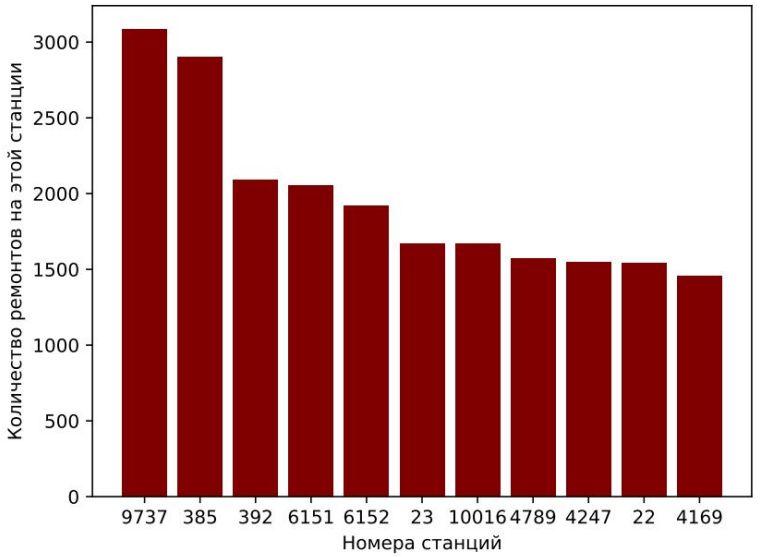
Out[10]:

	plan_date	wagnum	date_kap	date_dep	kod_vrab	date_pl_rem	id_road_disl	st_id_dest	id_road_dest	st_id_send	id_road_send	ost_prob	isload	fr
917123	2022-12-01	29930	NaT	2021-03-13	1	2024-03-13	36	8882	36	10016	38	20013	1	7
917124	2022-12-02	29930	NaT	2021-03-13	1	2024-03-13	36	8882	36	10016	38	19777	1	7
917125	2022-12-03	29930	NaT	2021-03-13	1	2024-03-13	36	22	36	8882	36	19738	0	7
917126	2022-12-04	29930	NaT	2021-03-13	1	2024-03-13	36	22	36	8882	36	19738	0	7
917127	2022-12-05	29930	NaT	2021-03-13	1	2024-03-13	36	22	36	8882	36	19738	0	7
917128	2022-12-06	29930	NaT	2021-03-13	1	2024-03-13	36	22	36	8882	36	19738	0	7
917129	2022-12-07	29930	NaT	2021-03-13	1	2024-03-13	36	9754	38	8882	36	19712	1	10
917130	2022-12-08	29930	NaT	2021-03-13	1	2024-03-13	36	9754	38	8882	36	19404	1	10
917131	2022-12-09	29930	NaT	2021-03-13	1	2024-03-13	38	22	38	8882	36	19110	1	10
917132	2022-12-10	29930	NaT	2021-03-13	1	2024-03-13	38	22	38	9754	38	18989	0	10
917133	2022-12-11	29930	NaT	2021-03-13	1	2024-03-13	38	22	38	9754	38	18989	0	10
917134	2022-12-12	29930	NaT	2021-03-13	1	2024-03-13	38	10348	38	9754	38	18989	0	22
917135	2022-12-13	29930	NaT	2021-03-13	1	2024-03-13	38	10348	38	9754	38	18672	0	22
917136	2022-12-14	29930	NaT	2021-03-13	1	2024-03-13	38	10348	38	9754	38	17839	0	22
917137	2022-12-15	29930	NaT	2021-03-13	1	2024-03-13	38	22	38	10389	38	17839	0	22
917138	2022-12-16	29930	NaT	2021-03-13	1	2024-03-13	38	10016	38	10389	38	17839	1	7
917139	2022-12-17	29930	NaT	2021-03-13	1	2024-03-13	38	22	38	10016	38	17440	0	7
917140	2022-12-18	29930	NaT	2021-03-13	1	2024-03-13	38	22	38	10016	38	17440	0	7

- **Главными помощниками оказались**:
- 1) **Дата планового капитального ремонта**
- 2) **Оставшийся пробег**
- 3) **Последняя остановка в месяце**
- 4) **Статистики** (суммарный путь, путь под нагрузкой, средний пробег в день, дисперсия) за предыдущий месяц.
- **Плановый капитальный ремонт и пробег – часть нормативов РЖД**. Не удивительно что они оказались сильными фичами нашего решения. Так же их советовал бейзлайн.

Номер последней посещённой станции в месяце





Картинка считалась по табличке прошедших ремонтов

Является географическим признаком. Мы заметили, что 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.

Долгие ремонты
Длинные забеги

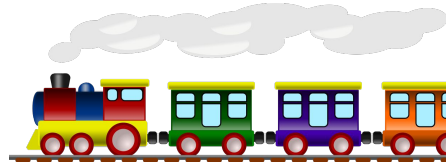
Вагон в предыдущие месяце мог испытывать экстраординарную нагрузку. Это может сыграть свою роль в принятие решения отправить вагон на ТО заранее. **Статистики** помогают детектировать такие кейсы. Также они помогают детектировать случаи, когда скрутка пробега и постановка метки происходят в разное время. Кроме того с такими случаями нам помогает бороться **время до предыдущего ремонта**.

2022-12-01	44394	12	91
2022-12-02	44394	12	91
2022-12-03	44394	12	91
2022-12-04	44394	12	91
2022-12-05	44394	12	91
2022-12-06	44394	12	91
2022-12-07	44390	12	91
2022-12-08	44122	12	91
2022-12-09	44122	12	91
2022-12-10	44122	12	91
2022-12-11	44122	12	91
2022-12-12	159999	12	91
2022-12-13	159860	12	91
2022-12-14	159731	12	91
2022-12-15	159722	12	91
2022-12-16	159722	12	91

```
target[target.wagnum == 91]
```

[24] ✓ 0.0s

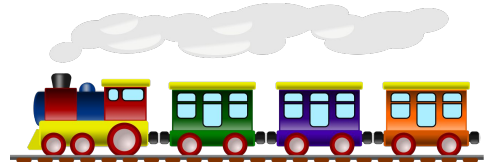
	wagnum	month	target_month	target_day
26366	91	2023-01-01	0	0
60342	91	2022-08-01	0	0
94317	91	2022-09-01	0	0
128293	91	2022-10-01	1	1
162269	91	2022-11-01	0	0
196245	91	2022-12-01	0	0



Параметры вагона

- Окончание срока службы. Возраст вагона был самым важным параметром влияющим на ответ

- Остальные параметры давали небольшой прирост порядка двух процентов



Кирпичики нашего результата

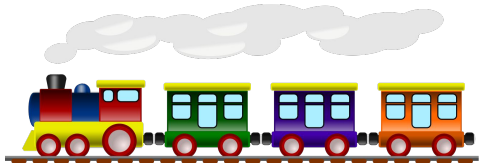
Самыми значимыми фичами оказались:

- накопленные статистики вагона за предыдущий месяц;
- станции отправления;
- время с последнего ремонта;
- нормативы РЖД из Baseline.

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

Самым простым улучшением, которое одновременно сильно повысило качество нашей модели, оказалось улучшение таргета на 10 дней.

Взяв бустинг на нормативах РЖД и добавив все наши фичи, мы получаем улучшение предсказания таргета на 10 дней с 0.2 до 0.67 и на месяц - с 0.55 до 0.69



Интересные особенности. Аккуратнее со счётчиками.

Дата среза	Оставшийся пробег
2023-01-12	121470
2023-01-13	121437
2023-01-14	160000
2023-01-15	160000
2023-01-16	160000
2023-01-17	160000
2023-01-18	160000
2023-01-19	160000
2023-01-20	160000
2023-01-21	160000
2023-01-22	160000
2023-01-23	160000
2023-01-24	160000
2023-01-25	160000
2023-01-26	160000
2023-01-27	160000
2023-01-28	119385
2023-01-29	119385
2023-01-30	119385
2023-01-31	119385

Разные типы ремонта нужно не путать.

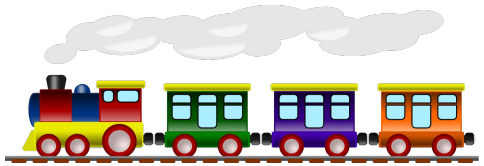
Норма пробега после ремонта для вагона №33350

	wagnum	norma_km
0	26318	110000
1	28344	0
2	8099	160000
3	33350	250000
4	5308	160000
5	16521	110000

Шум в счётчике ресурса хода.

	repdate	wagnum	ost_prob
0	2022-12-01	32353	596
1	2022-12-02	32353	596
2	2022-12-03	32353	159999
3	2022-12-04	32353	159999
4	2022-12-05	32353	159999
5	2022-12-06	32353	159876
6	2022-12-07	32353	159619

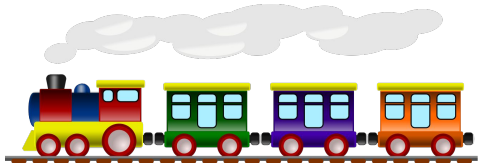
Фактическая норма пробега вагона №33350 после ремонта



22	2614926
7	2505
12	965
23	276
13	270
18	245
15	45
5	6
6	3
21	2
14	2

road_id	st_border_sign	st_sea_sign	st_river_sign	st_car_sign	st_ferry_sign	st_freigh_sign	opor_station_sign
19	0	0	0	0	0	1	0
5	0	0	0	0	0	1	0
24	0	0	0	0	0	1	0
3	0	0	0	0	0	1	0
19	0	0	0	0	0	1	0
24	0	0	0	0	0	1	0
22	0	0	1	0	0	1	0
5	0	0	0	0	0	1	0
7	0	0	0	0	0	1	0
17	0	0	0	0	0	1	0

Особенности станций

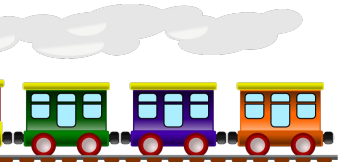


У большинства станций нет никаких особенностей, только у 23 из 16 тысяч. Среди них частотная особенность только возможность грузовых работ.

Другие результаты

- Данные о tr_rem не отображаются в таргете.
- Только первых три типа не константы
- Генерация фичей из этой таблицы не принесла прироста.

- Данные в таблице kti есть только за последний месяц, и нам показалось такое количество малым для использования в ML
- Генерация фичей из грузов и ownership не принесли результатов



Бизнес применимость

- Модель простая
- Генерация данных сделана аккуратно без подглядываний в будущее
- Значения совпадали на валидации по февралю, валидации по 5 фолдам и лидерборде
- Фичи не смещаются по времени, а значит модель можно использовать продолжительное время без необходимости вмешательства датасаентиста, например, для переобучения
- Высокое качество предсказания и, как следствие, экономия денег бизнеса за счёт более аккуратного распределения вагонов по ВПР с использованием нашей модели