Как выиграть BEST Hack'23?

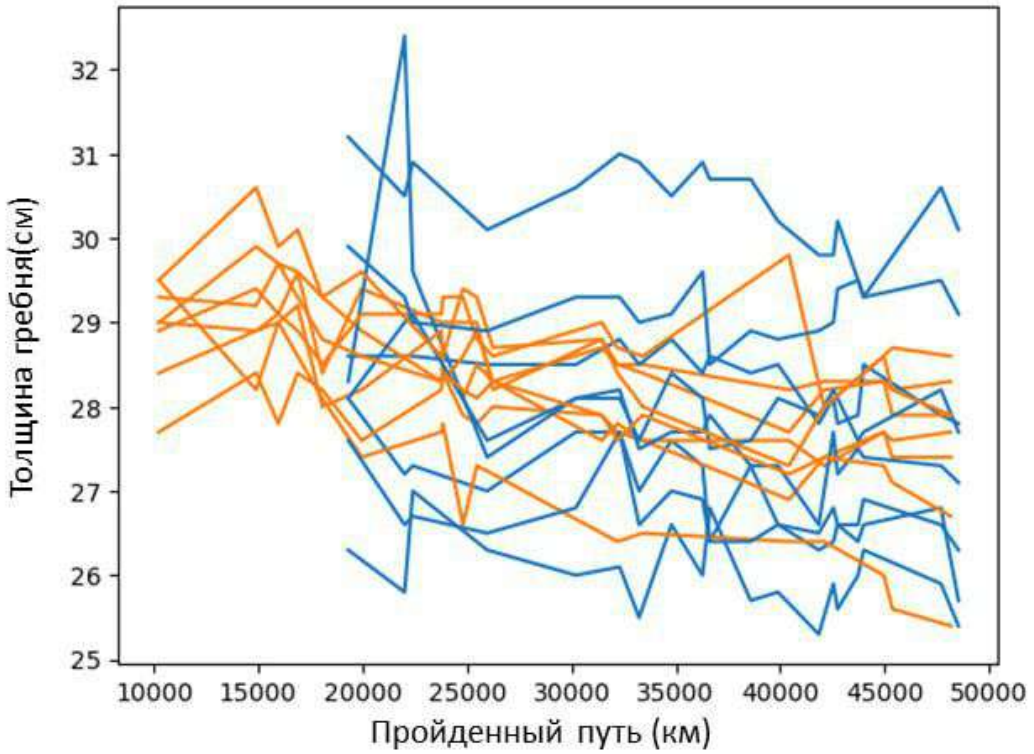
BEST Hack'23

В конце апреля проводился хакатон BEST Hack'23, на котором нам с командой удалось занять 1-ое место по направлению Data Science. Конкурс состоял из двух этапов: полуфинала и финала, продолжавшегося целые сутки без перерыва. Данные и задачи для этапов хакатона были предоставлены ЖД компанией ПГК.

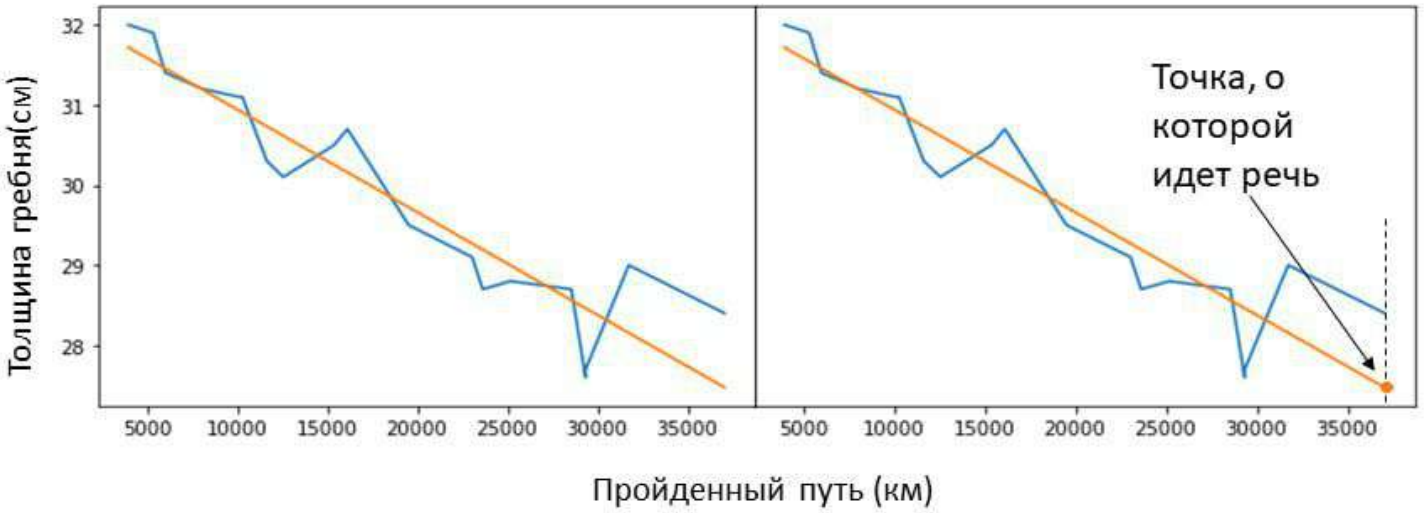
Полуфинал

Задачей в полуфинале было предсказание оставшегося ресурса хода (в км) грузового вагона по измеренной через некоторые промежутки времени толщине гребней на его колесах.

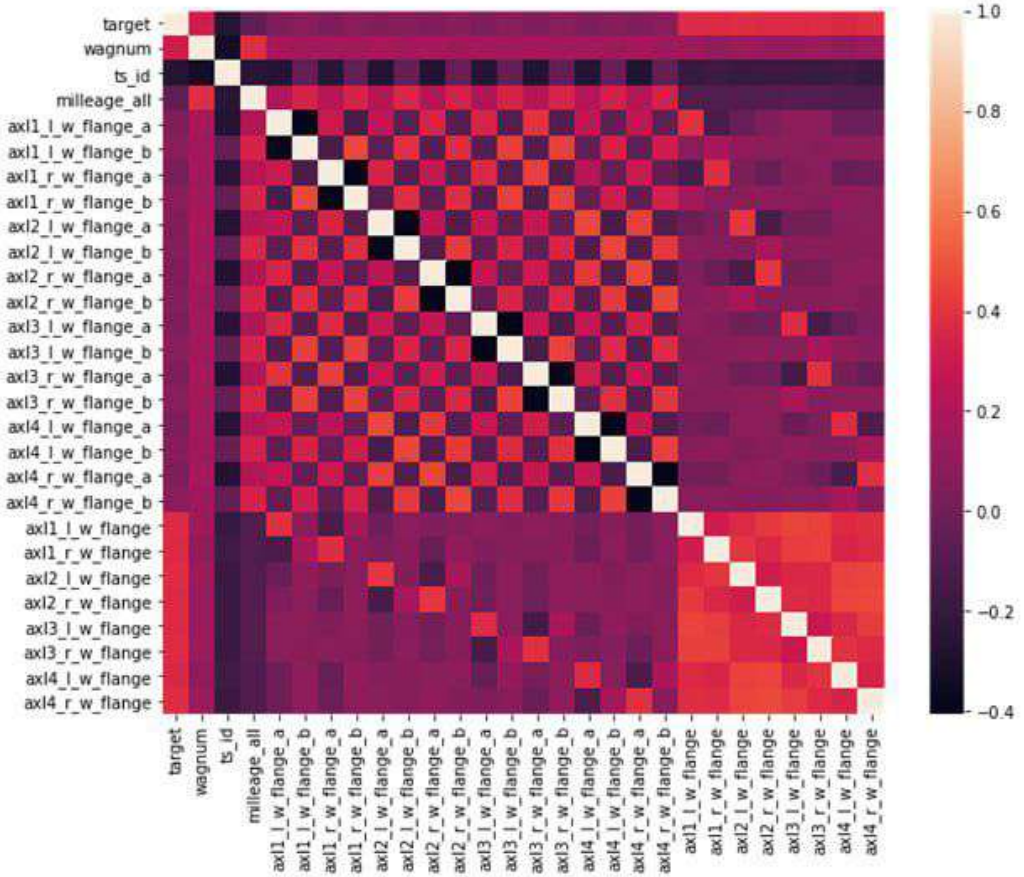
На графике видно, как меняется толщина гребней каждого из колес синего и оранжевого вагонов.



Так как измерения проводились быстро и недорогими датчиками, данные оказались засорены различными шумами и пропусками, видимыми на графиках (очевидно, что гребень стачивается с течением времени и не может «обрастать» металлом).
Мы решили заменить временные ряды из измерений наилучшей прямой, проходящий через точки данных, из которых он состоит. Это позволило использовать коэффициент наклона прямой и ее свободный член вместо 19-ти зашумленных измерений. Также мы вычислили крайнюю правую точку прямой, чтобы модели знали состояние гребня на момент последнего измерения.



Еще мы вычислили среднее значение ресурса хода для групп, с толщиной гребня 26-28 см, 28-30 см, 30-32 см, чтобы модель смогла понять зависимость между износом и оставшимся «временем жизни» вагона. Ниже представлена карта корреляций признаков и ресурса хода (target).



После предобработки признаков и пары минут обучения случайный лес (Random Forest Regressor) обеспечил нам проход в финал хакатона!

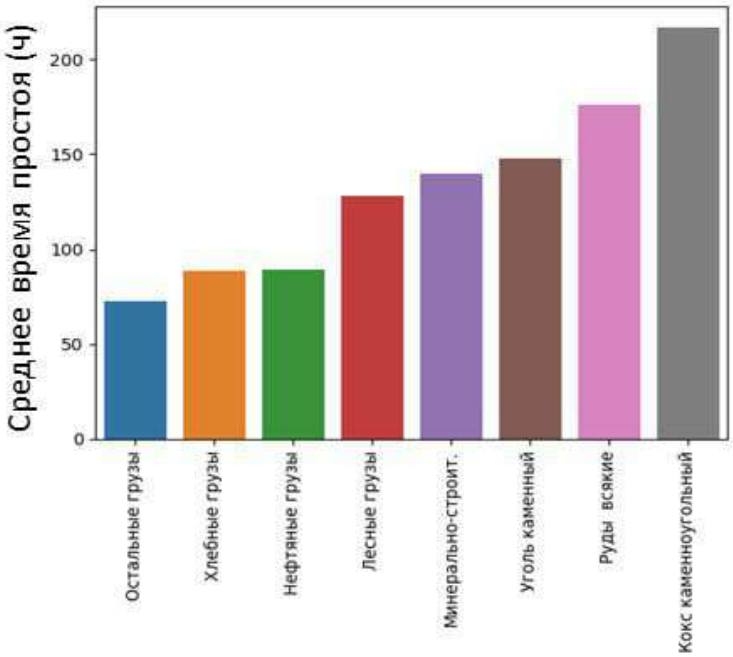
Финал

Финал, разумеется, тоже был посвящен ЖД тематике. Требовалось по данным накладных определить, сколько часов поезд пробудет на станции перед отправлением. Задача усложнялась многочисленными шумами в значениях признаков, отсутствием информации о состоянии поезда (он мог проехать станцию без остановок, а мог остаться на ремонте на несколько суток), однако нам были известны грузы, которыми наполнен поезд, компания, которой груз принадлежит, время отъезда поезда с прошлой станции и прибытия на текущую и еще много полезной информации.

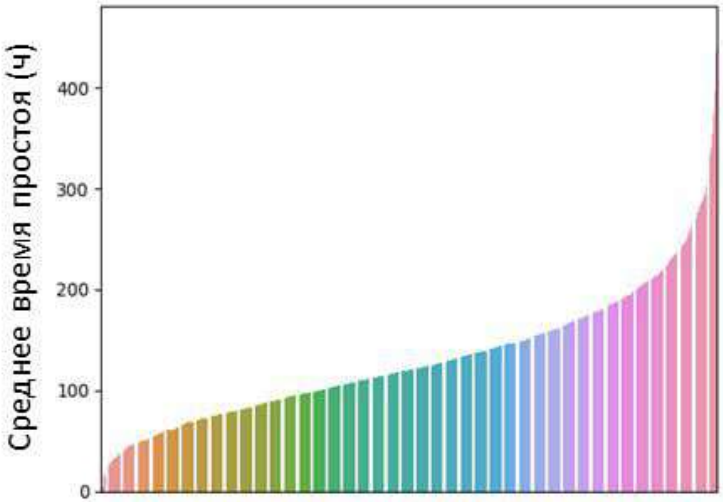
Сначала мы вычислили среднее значение времени стоянки поездов для наименования каждого груза, компании и других категориальных признаков (Target Encoding), надеясь, что подобные закономерности сохраняться на тестовой выборке. Данные средние значения неплохо разделяли выборку, показывая большую вариативность.

Например:

Для разных типов груза

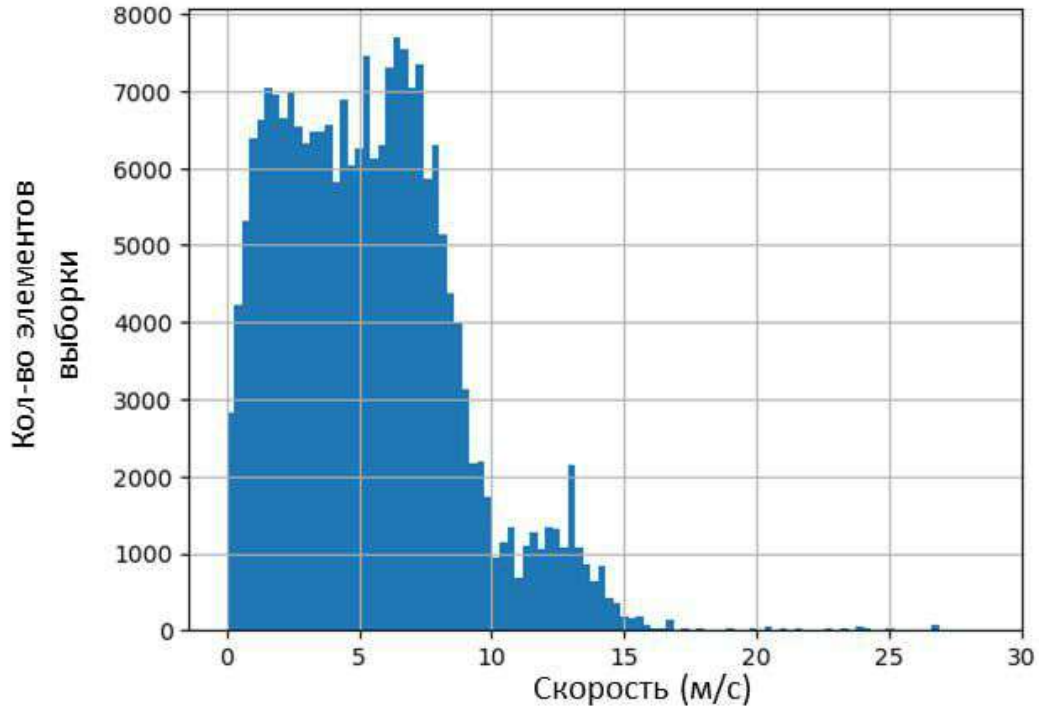


Для разных станций



Тут должны быть названия станций, но они не поместятся

После чего, воспользовавшись информацией о расстояниях между станциями и временем отправления, мы вычислили среднюю скорость поездов. Данный признак оказался крайне полезным, ведь, например, поезд, перевозящий цистерны с нефтью будет двигаться намного медленнее, чем пустой поезд. После очистки шумов данный признак показал нам трехгорбое распределение, говорящее о существовании некоторых 3-х групп рекомендуемых скоростей для поездов в зависимости от типов груза.



Также важно было не перемешивать данные перед обучением, так как использование модели на этапе тестирования предполагает предсказание будущего по данным из прошлого, а перемешивание во время разделения на тренировочную и проверочную выборку допускало бы случаи предсказания прошлого по настоящему, что гораздо проще.

Мы перепробовали множество моделей, в том числе и нейросетевых, но наилучшей из них оказался старый добрый случайный лес (Random Forest Regressor). Работающая параллельно с нами библиотека AutoML решила также, заодно подобрав удачные гиперпараметры для модели (глубину, количество деревьев в лесе и другие).

После обучения мы удалили из обучающей выборки объекты с самой большой ошибкой (чтобы модель не «распылялась», пытаясь установить шумовые, одиночные закономерности, а сосредоточилась на том, что она может объяснить и успешно рассчитать). Повторно обученная на обрезанной обучающей выборке модель показала почти вдвое лучший счет на этапе проверки!

Полученный таким образом лес выдал неплохой результат, который, в купе с дополнительными баллами за подбор признаков и оформление сайта для презентации работы алгоритма обеспечил нам победу.