Итоговая работа

**Тема:** Использование машинного обучения для прогнозирования отправления железнодорожных вагонов в ремонт.

**Введение:**

Отправка вагона в плановый ремонт может происходить по разным причинам, как по регламенту(срок/пробег), так и из-за накопления мелких дефектов:

- было много текущих ремонтов;

- не было вариантов на погрузку и т.д.

Этих причин много, и все они влияют на возможность осуществления ремонта.

Вагон отправляют в ремонт после получения уведомления об их неисправности. К сожалению, текущий процесс не позволяет распределять нагрузку на ремонт депо, управлять последней заявкой на погрузку пред ремонтом и многое другое.

**Цель:** снизить нагрузку на железнодорожную систему путем своевременного обслуживания вагонов.

**Задача:** создать модель прогнозирования даты отправления вагона в плановый ремонт.

**Постановка задачи:**

1. найти закономерности и оценить значимые признаки;
2. спрогнозировать, что вагон отправится в ПР в течение месяца;
3. спрогнозировать, что вагон отправится в ПР в течение 10 дней.

**Обзор моделей:**

Наивная модель построена на правилах с использованием минимального набора данных, без применения машинного обучения.

Реальный процесс выглядит следующим образом - в начале месяца берется срез по парку по всем вагонам, за ремонт которых несёт ответственность ПГК. Для выбранных вагонов требуется установить, какие из них будут отремонтированы в текущем месяце. Данная информация помогает планировать нагрузку на вагоноремонтное предприятие (ВРП). Вторая модель определяет критичные вагоны, которые будут отправлены в ремонт в первую очередь (в ближайшие 10 дней). Это помогает фокусировать внимание диспетчеров.

Основными критериями, по которым вагон отправляется в плановый ремонт - является его остаточный пробег и срок до планового ремонта.

В регламентах РЖД используется следующее правило - если ресурс по пробегу не превышает 500 км, срок службы не превышает 500 дней, число текущих ремонтов больше 5 и/или плановый ремонт должен наступить через 15 дней (или меньше), то вагон может ехать только на ВРП.

Из этого регламента вытекают две особенности:

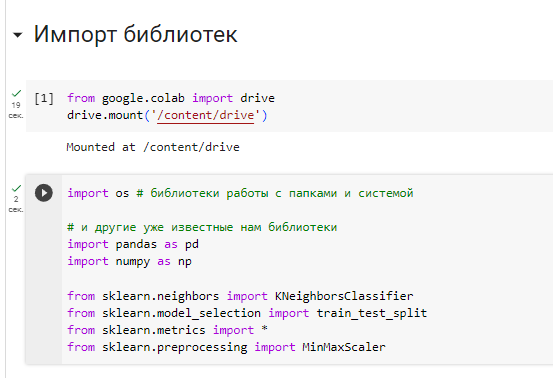
1. Диспетчер старается отправить вагон раньше положенных значений. Это позволяет выбрать предприятия, на которых ремонтироваться дешевле, а не ближайшее.

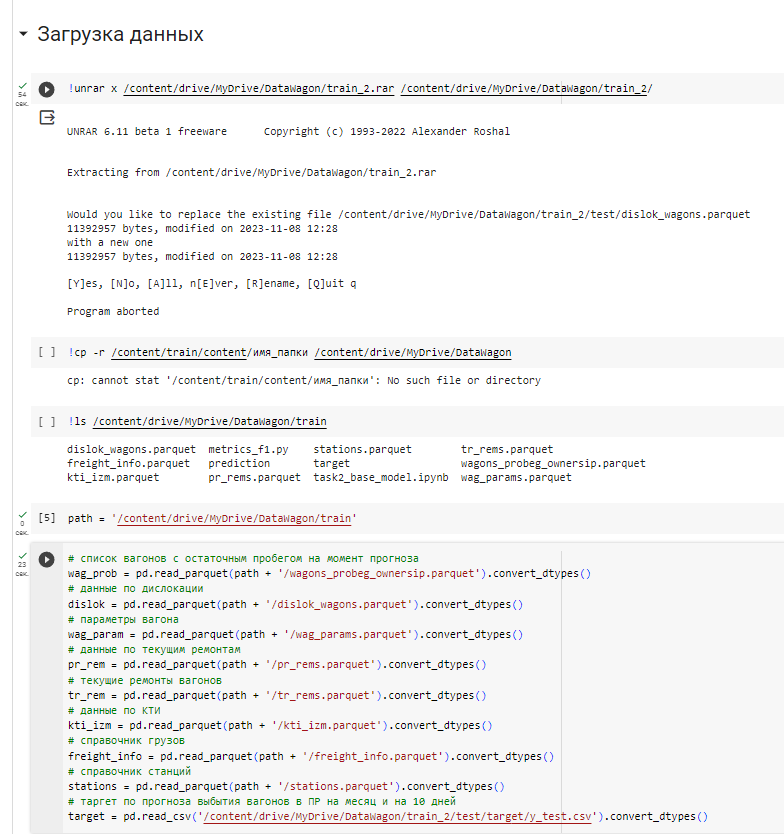
2. Компания-оператор может выбирать какому из нормативов нужно следовать - ремонтировать вагон по сроку, или по пробегу, или по обоим критериям сразу. Поэтому встречаются вагоны, у которых пробег может не отслеживаться.

Вагон может быть отправлен в плановый ремонт и раньше положенного. На это может влиять, например, история грузовых операций и количество текущих(мелких) ремонтов.

**Описание процесса решения (модель, тест, тренировка, результат):**

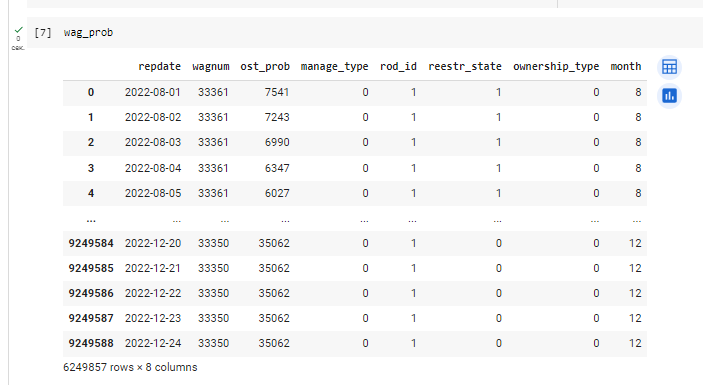
Изначально производится импорт необходимых библиотек (pandas, numpy, os) и загрузка данных.



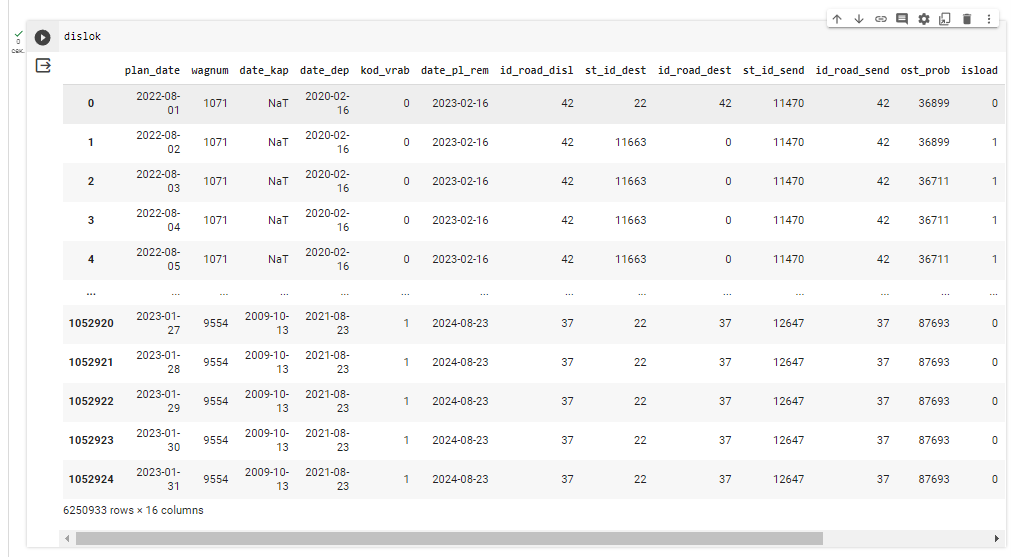


Исходные данные содержат 9 таблиц wag\_prob, dislok, wag\_param, pr\_rem, tr\_rem, kti\_izm, freight\_info, stations, target. Таблицы включают в себя следующее:

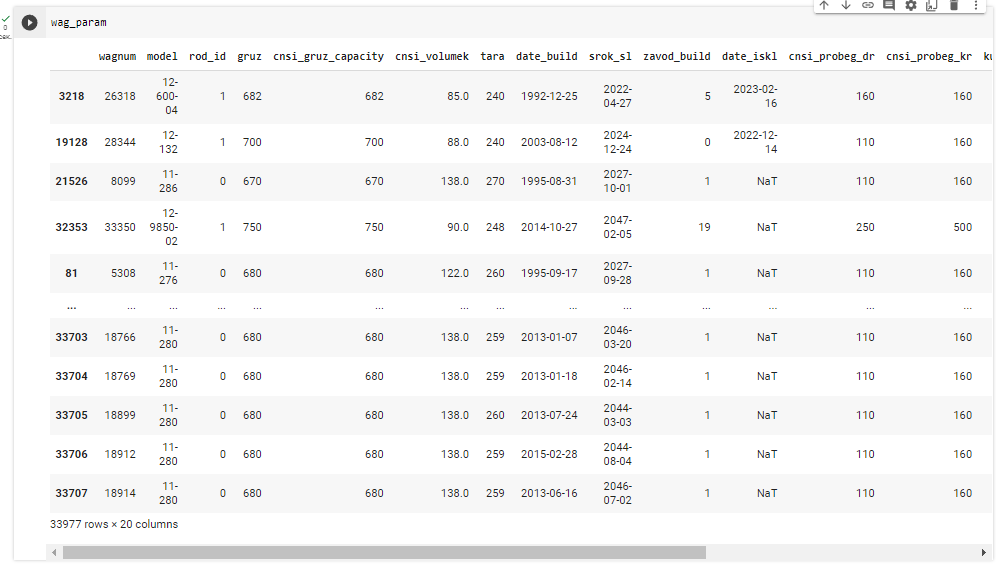
* Таблица wag\_prob. Список вагонов, по которым известен пробег и тип владения на дату среза: Дата среза, Номер вагона, Остаточный пробег, Вид управления по договору (1-собственноые, 2-лизинговые, 11-сервисные), Тип РПС, Состояние в реестре (1-слеженение, 3-вкдючение, 4-готовится к искл.), Тип собственности (1-собственные, 2-принятые в аренду, 4-привелеченный парк, 6-сервис), Месяц.



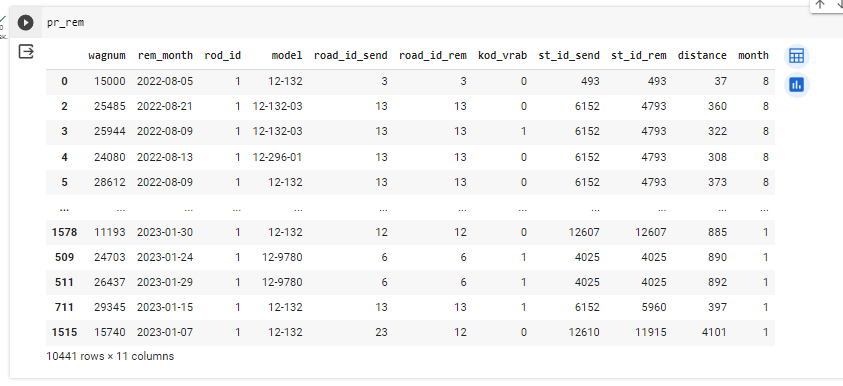
* Таблица dislok. Информация по дислокации: Дата отчета, Номер вагона, Дата последнего капитального ремонта, Дата последнего деповского ремонта, Код предстоящего планового ремонта, Дата предстоящего планового ремонта, Код дороги дислокации, Ид. станции назначения, Код дороги назначения, Индекс станции отправления, Код дороги отправления, Остаточный пробег, Груженый/порожний, Код груза, Ид. последнего груза, Расстояние между станцией отправления и прибытия.



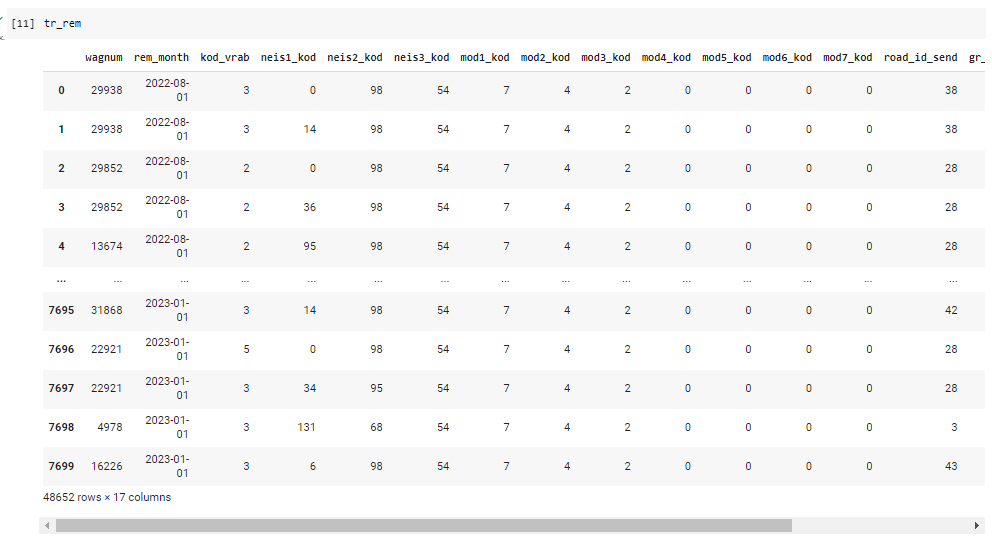
* Таблица wag\_param. Данные по характеристикам вагона: Номер вагона, Модель вагона, Тип РПС, Грузоподъемность (в центнерах), Предельная грузоподъёмность, Объем кузова, Масса тары в центнерах, Дата постройки, Дата окончания срока службы, Завод постройки, Норма пробега после ДР в тыс. км, Норма пробега после КР в тыс. км, Кузов, Код модели тележки, Тип тормозов, Тип воздухораспределителя, Межремонтный норматив пробега (<> 0, если вагон на пробеге), Признак передачи вагона в аренду (1-собст, 2-арендованный, 3-инвентарный).



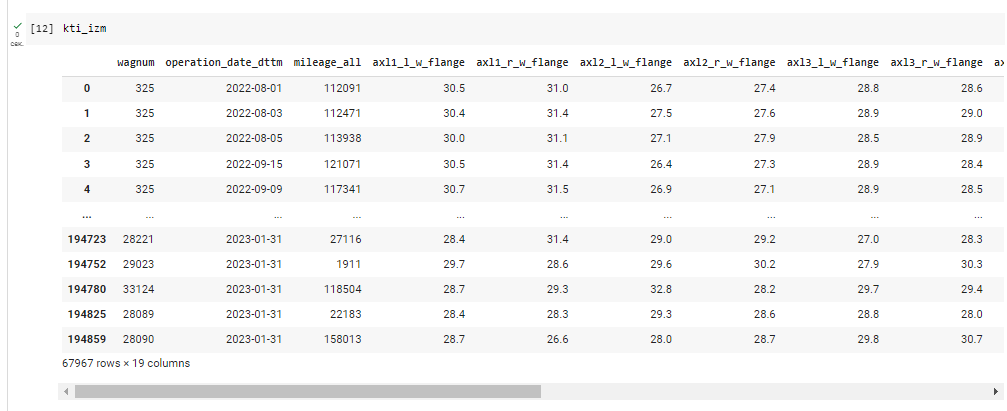
* Таблица pr\_rem. Данные по плановым ремонтам: Номер вагона, Месяц ремонта, Тип РПС, модель, Дорога выбытия в ПР, Дорога прибытия на ВРП, Код ремонта, Станция выгрузки, код станции ВРП, Расстояние со станции до ВРП, Месяц.



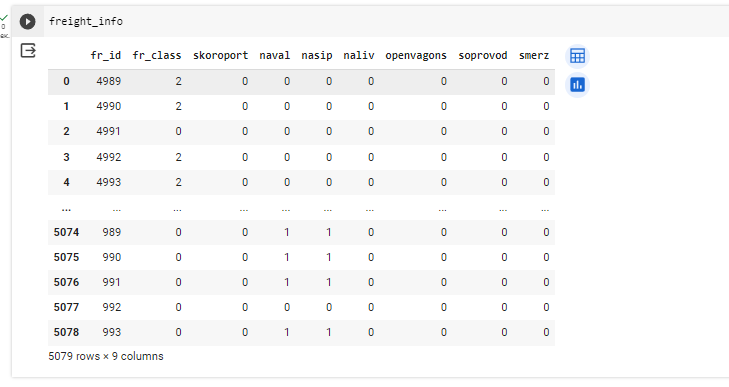
* Таблица tr\_rem. Данные по текущим ремонтам вагона: Номер вагона, Дата ремонта, Код работ, Код неисправности 1 (Заполняется по с.5353), Код неисправности 2 (Заполняется по с.5353), Код неисправности 3 (Заполняется по с.5353), Код модернизации 1 (Заполняется по с.5354), Код модернизации 2 (Заполняется по с.5354), Код модернизации 3 (Заполняется по с.5354), Код модернизации 4 (Заполняется по с.5354), Код модернизации 5 (Заполняется по с.5354), Код модернизации 6 (Заполняется по с.5354), Код модернизации 7 (Заполняется по с.5354), ID дороги перевода вагона в неисправные(nsi\_db.railway.RW\_ID), Пробег вагона в груженом состоянии, Код станции перевода вагона в неисправные (nsi\_db.station.ST\_CODE).



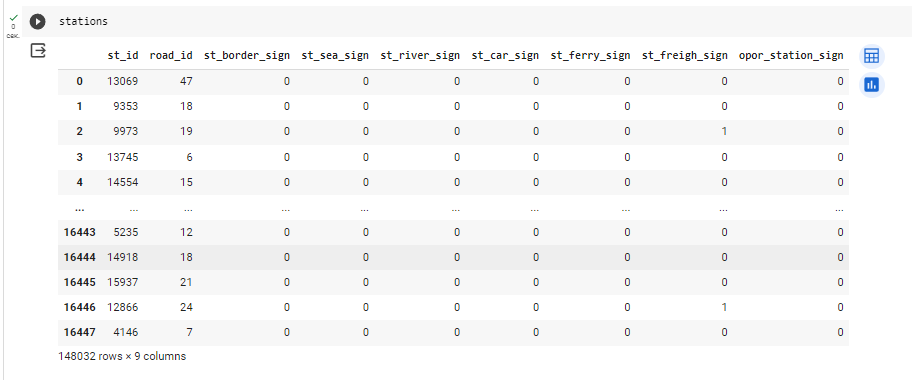
* Таблица kti\_izm. Данные по КТИ: Номер вагона, Дата и время измерения КТИ, Пробег общий, Толщина гребня 1 ось слева, Толщина гребня 1 ось справа, Толщина гребня 2 ось слева, Толщина гребня 2 ось справа, Толщина гребня 3 ось слева, Толщина гребня 3 ось справа, Толщина гребня 4 ось слева, Толщина гребня 4 ось справа, Толщина обода 1 ось слева, Толщина обода 1 ось справа, Толщина обода 2 ось слева, Толщина обода 2 ось справа, Толщина обода 3 ось слева, Толщина обода 3 ось справа, Толщина обода 4 ось слева, Толщина обода 4 ось справа.



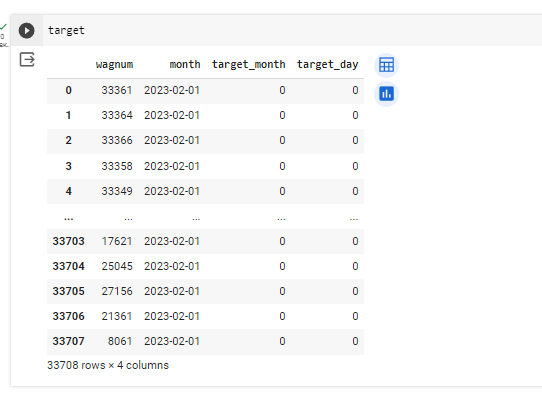
* Таблица freight\_info. Справочник грузов: Ид груза, Класс груза, Признак скоропортящегося груза, Признак погрузки навалом, Признак погрузки насыпи, Признак погрузки наливом, Перевозится в открытом подвижном составе, Груз требует сопровождения, Признак смерзающегося груза.



* Таблица stations. Справочник станции: Ид станции, Ид дороги, Признак граничной станции, Признак перевалки на море, Признак перевалки на реку, Признак перевалки на автотранспорт, Признак перевалки на паромную переправу, Признак станции, открытой для грузовой работы, Признак опорной станции.



* Таблица target: Номер вагона, Месяц ремонта, Ремонт в течение месяца (1-был ремонт, 0-не было ремонта), Ремонт в первые 10 дней (1-был ремонт, 0-не было ремонта). Данная таблица содержит две целевые переменные, так как стоит задача независимо друг от друга определить выбытие вагона в ремонт в первые 10 дней и в течении месяца.



**Подготовка данных:**

Перед обучением модели необходимо подготовить данные. Для этого:

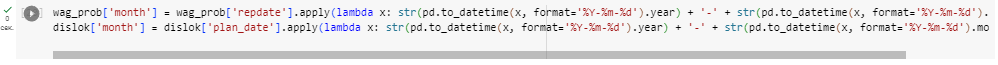
* Произвелась оценка среднесуточного пробега по пробегу вагона, на тот случай если данных по нормативу нет.



* Удалены дубликаты номеров вагонов из всех таблиц.



* Из таблиц по фильтру отобраны все значения на дату обучения.



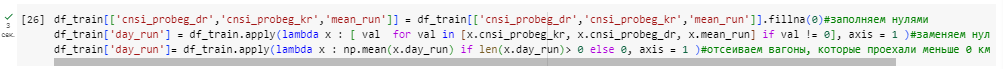
* Посчитано сколько текущих ремонтов было за прошедший период.



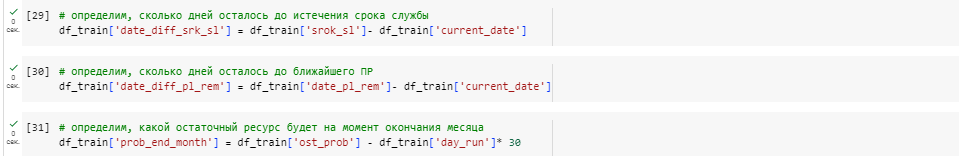
* Выбраны необходимые колонки из таблиц wag\_prod\_train, dislok\_train, wag\_param\_pred, pr\_rem\_train, kti\_izm\_train, tr\_term\_train и объединены в один общий датафрейм для обучения модели.



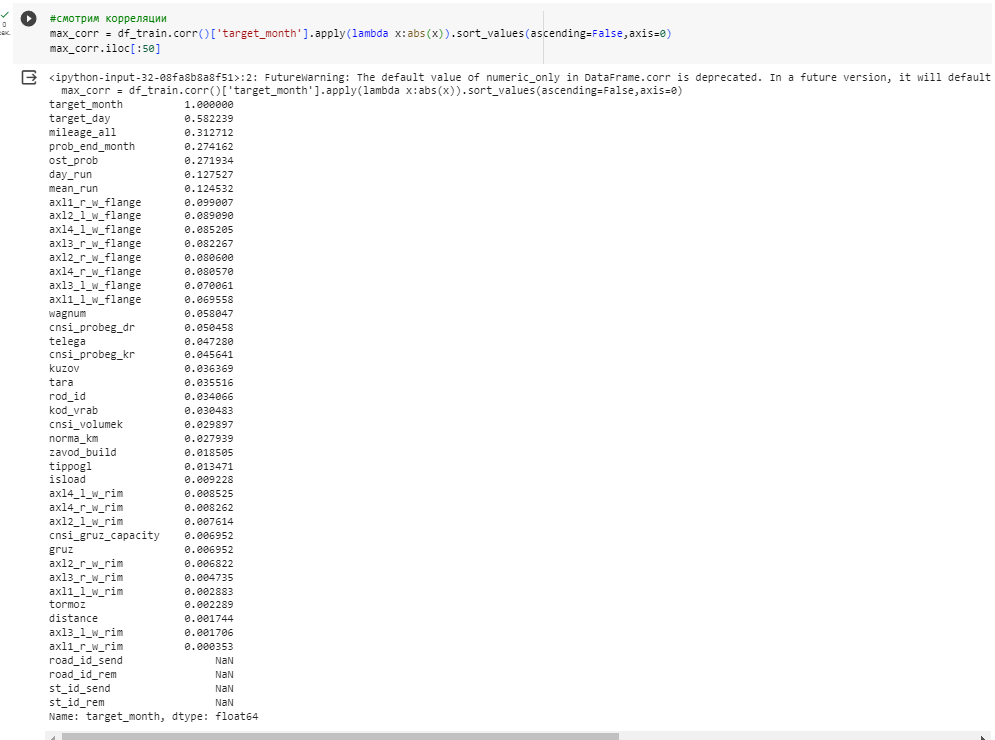
* Отсеяны вагоны, которые проехали 0 км в день, т.к. по ним невозможно обучить модель.



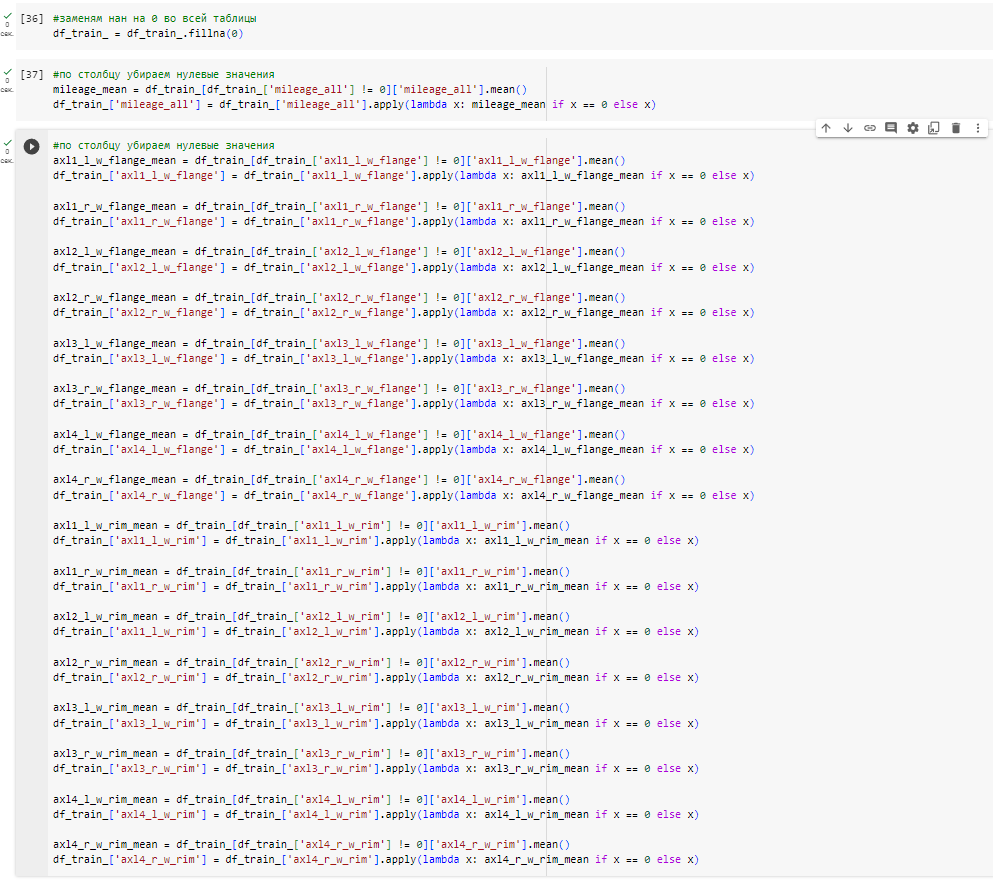
* Также рассмотрено количество дней, которое осталось до истечений срока службы вагона, дней до ближайшего ПР и какой остаточный ресурс будет на момент окончания месяца.



* Найдены параметры, которые больше всего коррелируют с target\_month (будет ли в этом месяце ремонт).



* Удалены нулевые значения по столбцам. Взяты 24 лучших параметров, которые коррелируют с target\_month.



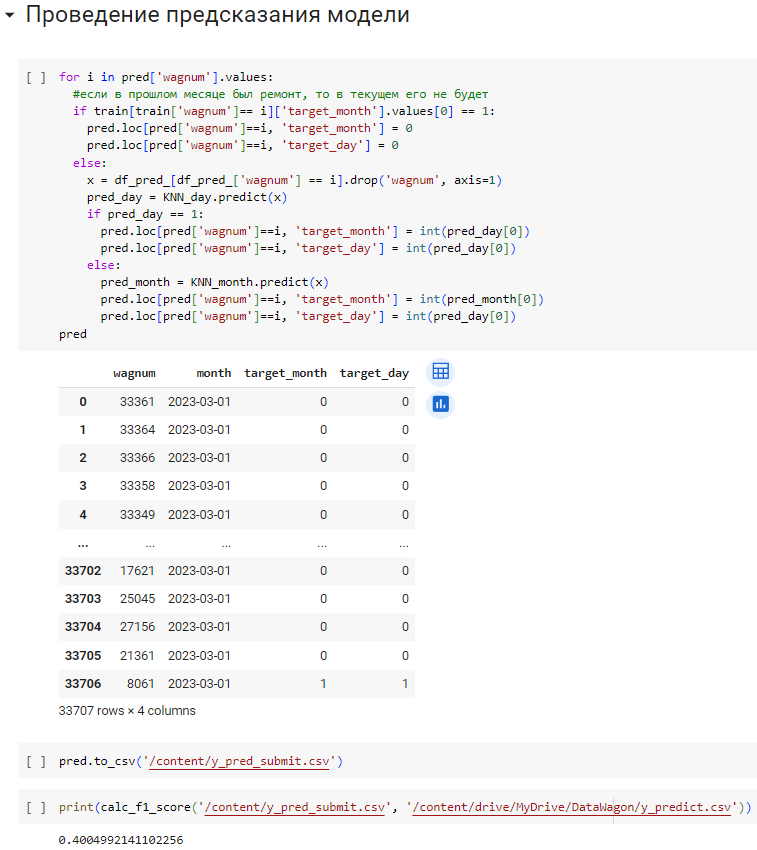


**Обучение модели:**

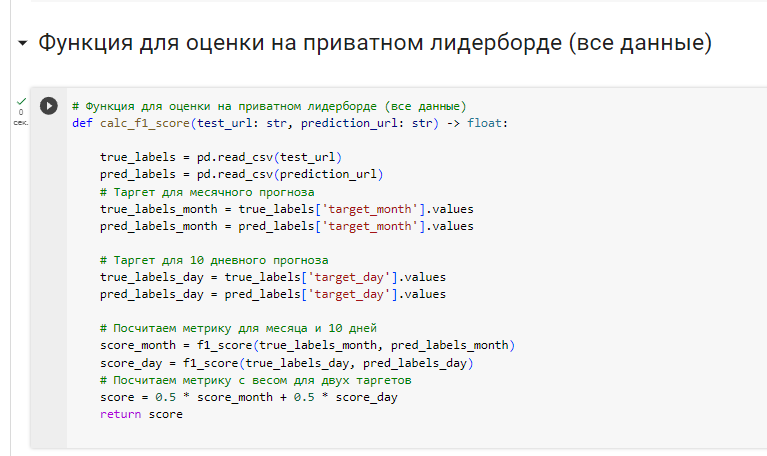
Для прогнозирования выбытия вагона в плановый ремонт применяется модель К ближайших соседей (KNN). Количество соседей выбралось 3 в связи с лучшей сходимостью.



Обработка данных для предсказания производилась так же, как и для обучения модели. Перед прогнозированием проверялось условие на наличие ремонта вагона в прошлом месяце. Если в прошлом месяце у вагона производился плановый ремонт, то в текущем месяце его не будет.



В результате предсказания общий счет модели равен 40%. Счет модели складывался из половины точности предсказания на месяц и половины точности предсказания на день.



**Заключение:**

Исходя из работы можно сделать вывод, что модель имеет удовлетворительную прогностическую способность и для дальнейшего ее использования в производстве необходима доработка, например, обучение на большем количестве данных, так как в представленном наборе в большом количестве строк отсутствовали значения.