# Технологический стек

Веб-приложение для запуска кода и визуализации данных Jupyter Notebook;

Язык программирования Python;

HTML + Bootstrap 5 + JavaScript + CSS.

# Работа и взаимодействие с моделью

Основной файл формата \*.ipynb рекомендуется запускать в Jupyter Notebook или других IDE, поддерживающих работу с ячейками.

Структура блокнота состоит из нескольких разделов. Прежде всего общий датасет разбивается на выборки по номерам эксгаустеров. Затем проводится анализ зависимостей имеющихся признаков.

### Анализ данных

С целью уменьшения вычислительной нагрузки была создана функция compute\_corr\_matrx(X\_train\_n, threshold), поэтапно вычисляющая значения корреляций между столбцами признаков X\_train\_n, и отбирающая столбцы со значением корреляции, превышающей пороговое threshold.

Так, например, для эксгаустера №4 имели значение корреляции более 0.8 следующие столбцы:

['ЭКСГАУСТЕР 4. ТОК РОТОРА2',

'ЭКСГАУСТЕР 4. ТЕМПЕРАТУРА ПОДШИПНИКА НА ОПОРЕ 2',

'ЭКСГАУСТЕР 4. ТЕМПЕРАТУРА ПОДШИПНИКА НА ОПОРЕ 3',

'ЭКСГАУСТЕР 4. ТЕМПЕРАТУРА ПОДШИПНИКА НА ОПОРЕ 4',

'ЭКСГАУСТЕР 4. ТЕМПЕРАТУРА МАСЛА В МАСЛОБЛОКЕ',

'ЭКСГАУСТЕР 4. ТЕМПЕРАТУРА ПОДШИПНИКА НА ОПОРЕ 4',

'ЭКСГАУСТЕР 4. ТЕМПЕРАТУРА МАСЛА В МАСЛОБЛОКЕ',

'ЭКСГАУСТЕР 4. ТЕМПЕРАТУРА МАСЛА В МАСЛОБЛОКЕ',

'ЭКСГАУСТЕР 4. ВИБРАЦИЯ НА ОПОРЕ 3. ПРОДОЛЬНАЯ.',

'ЭКСГАУСТЕР 4. ВИБРАЦИЯ НА ОПОРЕ 4. ПРОДОЛЬНАЯ.',

'ЭКСГАУСТЕР 4. ВИБРАЦИЯ НА ОПОРЕ 4. ПРОДОЛЬНАЯ.']

При этом повторение столбцов свидетельствует о том, что они коррелируют не с одним другим столбцом, а с несколькими.

Далее для каждого из признаков таргета выбирались самые коррелирующие столбцы с помощью функции compute\_corr\_target(X\_train\_n, y\_train\_n, threshold).

Таким образом, было отмечено, что на большинство параметров сильно влияют значения токов ротора.

Работа с пропусками производилась с помощью ресемплинга и последующей интерполяции

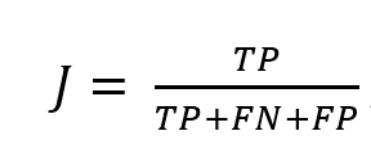
В следующем блоке выводятся графики рассеяния, гисттограммы и анализ сезонностей с помощью функции plot\_seasonality(df, start, end, res). Было установлено, что режим работы одного из самых важных признаков для анализа – тока ротора, имеет характерную зависимость. В целом наблюдаются минимальные колебания около определенного значения, но в течение трех лет само это значение несколько раз менялось. Особенно заметен скачок в пределах апреля 2020 года. При этом построение гистограмм показывает, что для значительной части признаков действительно наблюдаются отличия в режимах работы до и после этого периода.

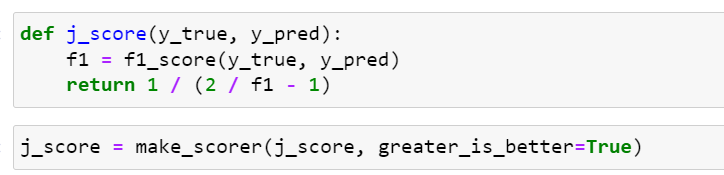
Анализ сезонности показал, что для эксгаустера №4 при ресемплинге в 1 час можно наблюдать ежедневную сезонность, в 1 день - еженедельную, а в 15 дней – ежегодную. Так, резкий пик работы приходится на середину июля, а спад на конец ноября. При этом для некоторых эксгаустеров ежегодная сезонность прослеживается не так сильно, но ежедневная и еженедельная четко наблюдаются.

### Обучение модели для предсказания поломок М3

Для оптимизации все этапы заворачивались в функции.

Прежде всего была создана пользовательская метрика на основе f1\_score





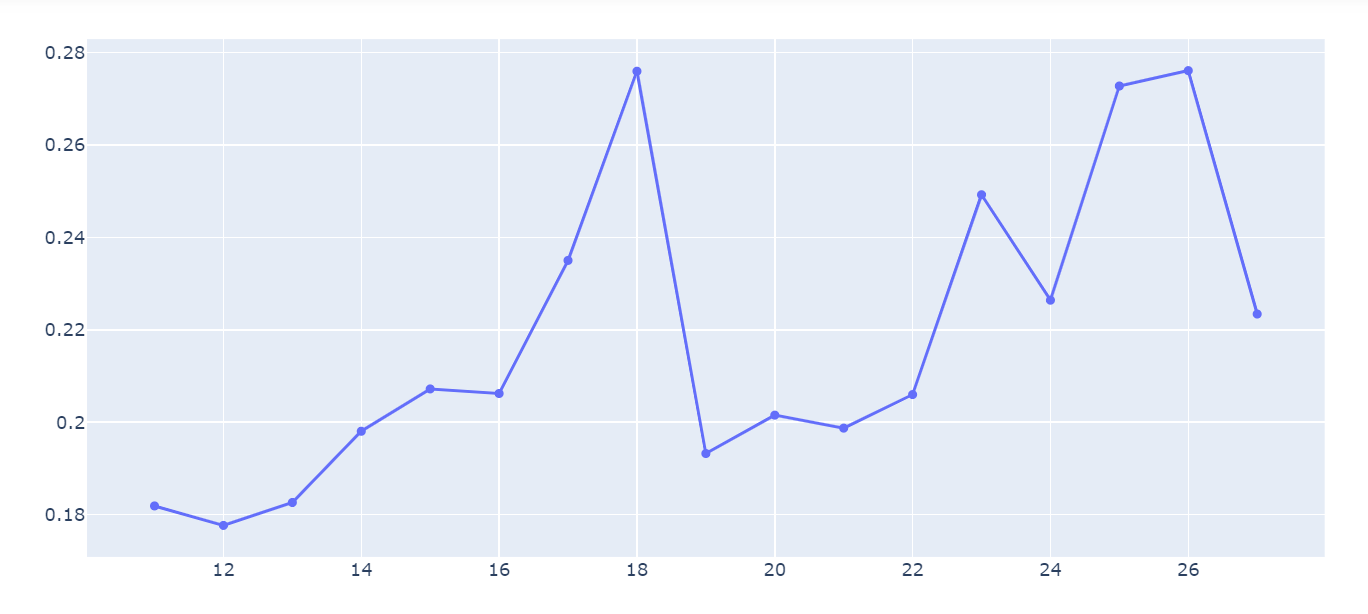
Затем для упрощения применения GridSearch для подбора лучших параметров модели была создана функция gsearch.

Преобразование данных начинается с ресемплинга resampling\_func для периодов в 1 день, 1 час и 5 минут. В дальнейшем для каждого из таргетов производился соответствующий ресамплинг с лучшими показателями метрики. Затем в общем случае добавляются различные окна для соответствующих признаков посредствами функции window\_add. После этого используются различные сдвиги функцией lag\_add.

Также для первого признака были исследованы дополнительные способы преобразования данных:

* Среднеквадратичное отклонение от окна window\_add\_std;
* Разница между максимумом и минимум window\_add\_std;
* Отклонение от среднего значения window\_mean\_diff;
* Отношение текущих значений к среднему window\_mean\_ratio.

При этом использовались дополнительные функции win\_tendency и lag\_tendency, выводящие графики метрик для диапазона значений окна и сдвига.



Пример графика зависимости окна (по оси х) от метрики (по оси у)

Из соображений ускорения процесса дополнительные методы в предсказании признаков не использовались.

# Потенциал решения в будущем

* Применение всех описанных способов преобразования и добавления столбцов в таблицу таргетов для остальных эксгаустеров;
* Надлежащая работа с пропусками разными способами и выбор наилучшего;
* Задача мультиклассификации поломок;
* Интегрирование модели в продакшн;
* Работа с блоком аномалий;
* Предсказание поломок типа М1: соответствующая обработка данных и применение различных моделей.