**特征工程与模型进阶**

* **Polars与Pandas进阶**
* **时序特征提取方法**
* **树模型使用**

**新的特征工程思路**

[https://www.kaggle.com/competitions/predict-energy-behavior-of-prosumers/discussion/456369](https://www.kaggle.com/competitions/predict-energy-behavior-of-prosumers/discussion/456369" \t "_blank)

def new\_features(df):

for col in ['temperature', 'dewpoint', '10\_metre\_u\_wind\_component', '10\_metre\_v\_wind\_component', '10\_metre\_u\_wind\_component\_fl',

'10\_metre\_v\_wind\_component\_fl', '10\_metre\_u\_wind\_component\_fdw', '10\_metre\_v\_wind\_component\_fdw', '10\_metre\_u\_wind\_component\_flw',

'10\_metre\_v\_wind\_component\_flw']:

for window in [1]:

df[f"{col}\_diff\_{window}"] = df.groupby(["county", 'is\_consumption', 'product\_type', 'is\_business'])[col].diff(window)

df[f"{col}\_diff\_{window}"] = df.groupby(["county", 'is\_consumption'])[col].diff(window)

df[f"{col}\_diff\_{window}"] = df.groupby(["is\_business", 'is\_consumption'])[col].diff(window)

return df

数据划分的方法：

* 时间划分：验证集晚于训练集

dates = (df\_train['year'].astype(str) + df\_train['month'].apply(lambda x: str(x).zfill(2))).astype(int)

train\_index = dates < np.unique(dates)[-1]

val\_index = dates >= np.unique(dates)[-1]

* 按照prediction\_unit划分：验证集和训练集prediction\_unit 重合度80%

**如何进一步上分？**

1、理解现有的代码的逻辑，如何做特征？如何验证模型的？如何对训练集和测试集做特征？

2、改进方向：

* + 新的特征（收益最大）
    - 天气，从地理位置做新的特征？
    - 天然气或其他的自变量可以考虑做归一化、剔除离群点
    - ...
  + 模型调参
  + 换新的模型

**高分代码**

Pandas版本：[https://www.kaggle.com/code/finlay/enefit-prediction-from-0-to-1](https://www.kaggle.com/code/finlay/enefit-prediction-from-0-to-1" \t "_blank)

Polars版本：[https://www.kaggle.com/code/finlay/enefit-prediction-polars-101](https://www.kaggle.com/code/finlay/enefit-prediction-polars-101" \t "_blank)

* 特征工程思路、训练模型
* 内存上差异（Polars更加节约内存）

什么类型的比赛：时间序列 + 多表预测（不同的自变量分开存储）

时间序列是以设备为区分，需要同时预测多个设备在未来的电量消耗和电量产生。

现在高分代码有什么差异（72 -> 67/68）？

* 主要的差异：训练模型使用early stop、lightgbm的objective、target（生产和消耗）可以分开建模分开预测
* [https://www.kaggle.com/code/siddhvr/enefit-pebop-updated-feat-engg-and-lgbm-ensemble/](https://www.kaggle.com/code/siddhvr/enefit-pebop-updated-feat-engg-and-lgbm-ensemble/" \t "_blank)
* [https://www.kaggle.com/code/davero/enefit-lgb-with-regression-l1-objective](https://www.kaggle.com/code/davero/enefit-lgb-with-regression-l1-objective" \t "_blank)
* [https://www.kaggle.com/code/vitalykudelya/explain-dataset-and-baseline/](https://www.kaggle.com/code/vitalykudelya/explain-dataset-and-baseline/" \t "_blank)
* [https://www.kaggle.com/code/newtonbaba12345/lb-67-48](https://www.kaggle.com/code/newtonbaba12345/lb-67-48" \t "_blank)