Temperature Calibration Challenge

Hong Tang

https://www.linkedin.com/in/hong-tang/

Summary and Recommendation

- **Why:** Primary objective is to output a value for air temperature that is more accurate and reliable than the raw value from our main air temperature sensors
- **Method: build regression model to calibrate Mark data from sensor to reference data (Gold standard)**
- **Findings and Challenges**
  - Missing value, outliers
  - Challenges on resampling and joint Mark data and reference data:
    - selection on join with tolerance of 10min
  - Strong collinearity among features Band downwelling watts (Potnetial overfitting)
  - Compare with lasso regression, Random Forest Model with pipeline achieves almost perfect score (R2, MSE)
  - Further work on cost function to investigate the residual (predict T vs reference T)
- **Lessons Learned and Best Practices**
  - Focus on feature engineering, and data QC
  - ML Pipeline avoid data leakage, makes modeling easy to read and maintain
  - Communication with stakeholders for business drivers
- **Plan Forward:**
- Improve feature engineering;
- Investigate unsupervised classification

ML Process

Cycle 1

Feature EDA

Existing numerical features

Model Selection

Simple Linear regression

Cycle 2

Feature EDA

Time features

Numerical features
Categorical features

More data cleaning

Joint resample Data decisions

Modeling

RandomForest Model

Pipeline

Validation

Cycle 3

Future feature engineering

Time Series Analysis

Different Cost function focus on residual trends

Model Deployment

Summary

recommendation

Summary

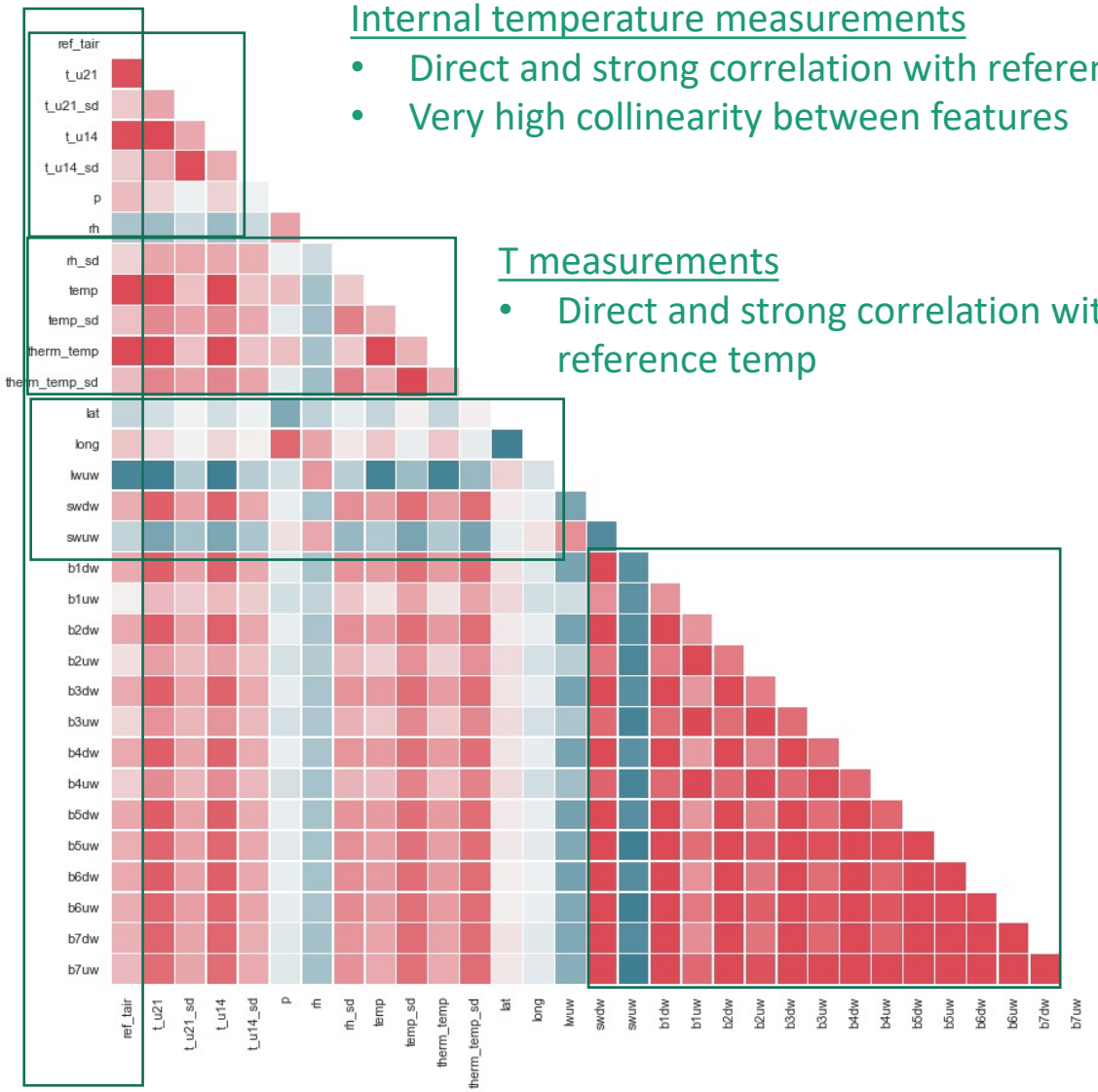
Internal T. and P, humidity

Temp Measure

Other Measure.

Band downwelling watts

Feature Correlation With Target



Internal temperature measurements

- Direct and strong correlation with reference temp
- Very high collinearity between features

T measurements

- Direct and strong correlation with reference temp

Other factors

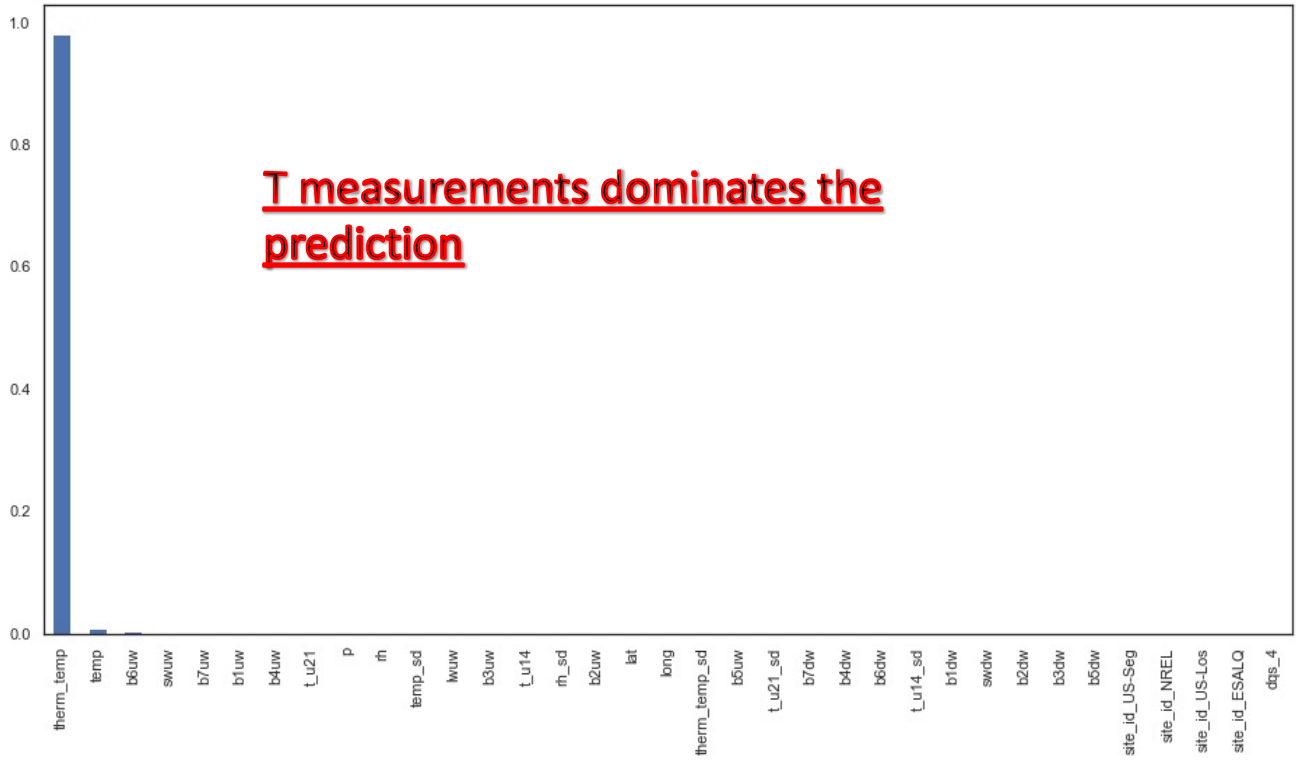
- Temp. measurements
- Decrease lwuw, swuw increases temp
- Latitude increases temp decrease(North cooler)

Increase bxdw increase T.

- Weak positive correlation with temp
- Strong colinearlity

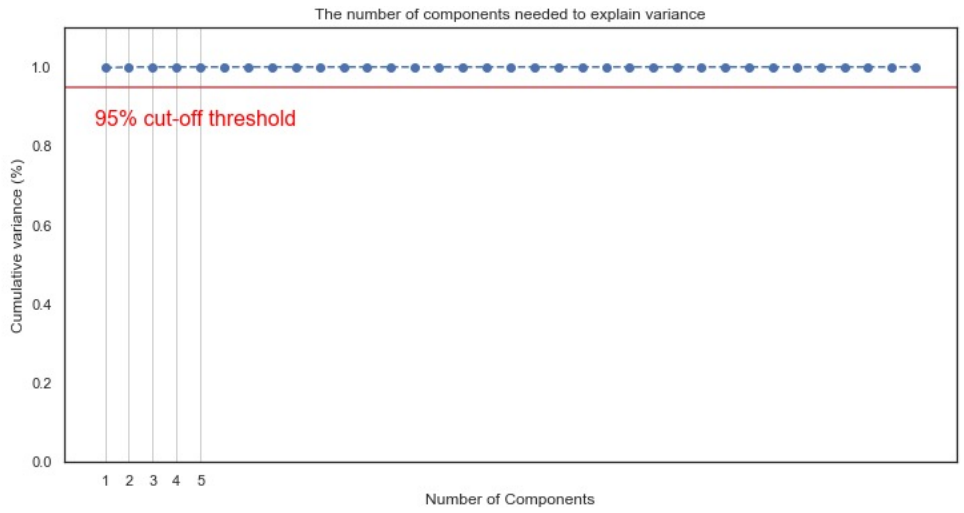
Importance of Influence and PCA component analysis

all indicate temp and therm_temp overshadow other features for reference temp prediction



T measurements dominates the prediction

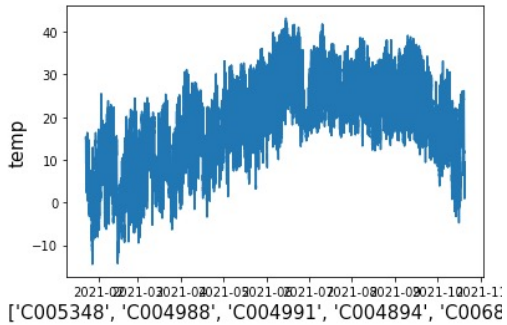
PCA no. comp. need to explain variance

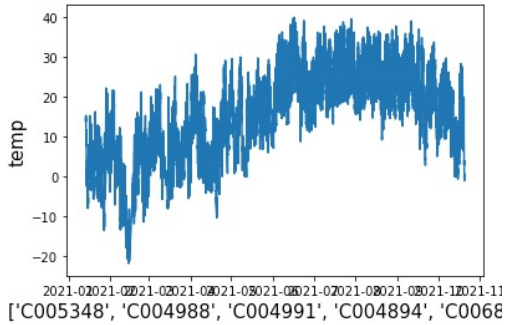


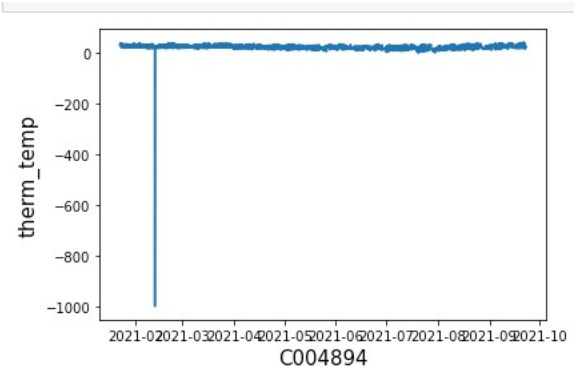
EDA and Feature Engineering

Batch Visualization

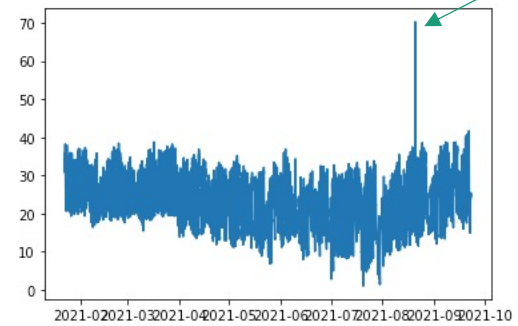
```
    plot all property:props plots by cats:catego
    '''
    for cat in cats:
        fig, ax = plt.subplots()
        plt.plot(df[df[cat_name]==cat][props])
        ax.set_xlabel(cats, fontsize=15)
        ax.set_ylabel(props, fontsize=15)
quick_look(df, 'device', cats, 'temp')
```





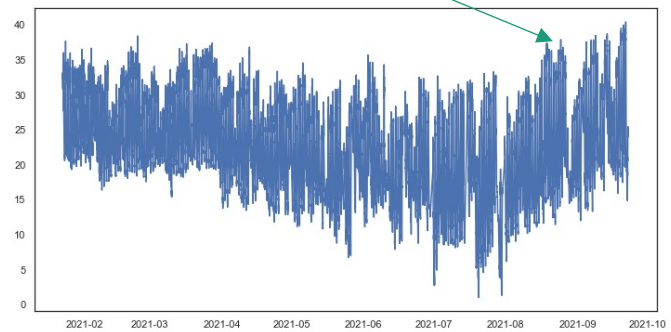



```
plot_temps(df,'C004894','temp' )
```



Remove outliers in batch

```
plot_temps(df,'C004894','temp' )
```

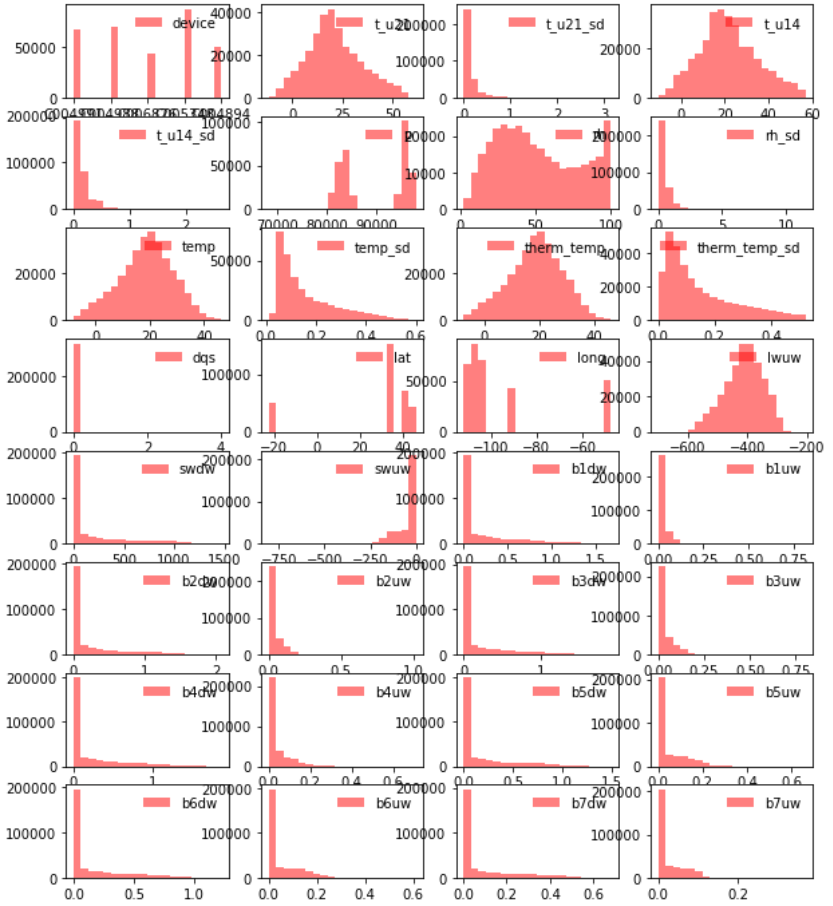


- Visual QC
- Drop lwdw due to more than 90% missing data
- Outliers removal
- rescaling

Feature EDA

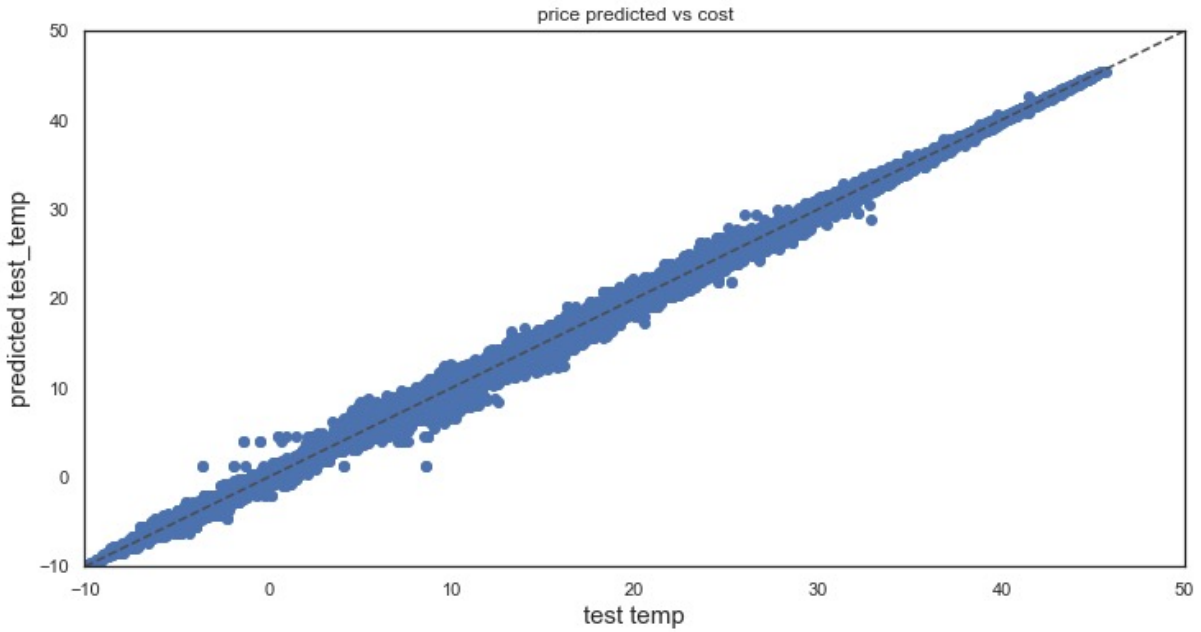
- IQR outlier treatment
- Internal tempratures are trianglar or “normal-ish” distribution
- Temp. standard deviation are skewed distribution

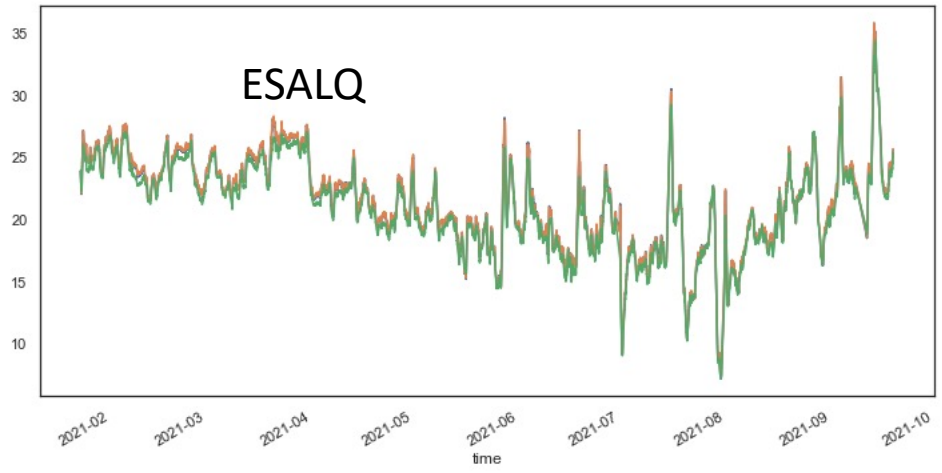
- Band downwelling watts are very simiar skewed distribution with long tails. Will be interesing to color code with temperature
- Wide range of values, rescaling is needed. Three scalers are tested decide to use standard scaler

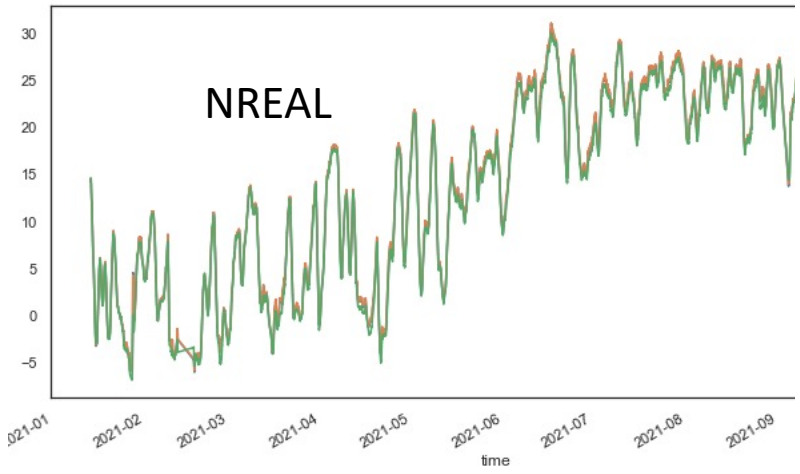


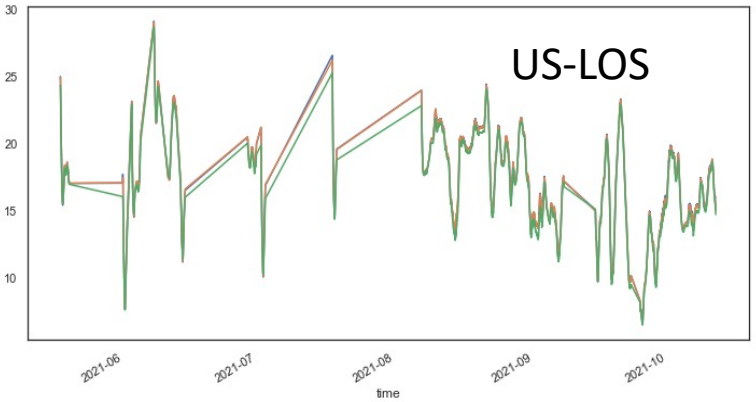
Model improvement from additional feature creation AUC improves from 90-94%

- Linear Regression does not have enough predictability (R2)
- Different Scaler and PCA tested
- RandomForest has great R^2 ,MSE for both train and test data









'funny looking data' need to check