Model documentation and write-up

1. **Who are you (mini-bio) and what do you do professionally? If you are on a team, please complete this block for each member of the team.**

   Experienced Data Scientist specializing in time series and forecasting. Currently working in the IoT domain, focusing on elevating consumer experience and optimizing product reliability through data-driven insights and analytics. Previously worked in various tech companies in Indonesia.

2. **What motivated you to compete in this challenge?**

   I was motivated to compete in this challenge because of its unique setup. Unlike typical data science competitions, there's no predefined training dataset provided. This means participants must not only focus on modeling but also on finding the right data to be used. Additionally, the requirement to submit operational code adds another layer of complexity and practical application. I saw this as an exciting opportunity to test and expand my machine learning skills in a practical, real-world setting.

3. **If there are any particularly useful snippets of code when producing your communication outputs (calculations, visualizations, etc.) that would be useful to highlight, please copy-paste them here.**

   Get SHAP values directly from LightGBM model prediction
   explainability/shap.py line 67-70

```
67                    _pred_output = pd.DataFrame(
68                        mdl.predict(df_features[mdl.feature_name()], pred_contrib=True),
69                        columns=mdl.feature_name() + ["base_value"],
70                    )
```

Beautify table of SHAP summary using pandas styling
explainability/report.py line 845-896

```
845     _tbl = (
846         _res_final.style.applymap(style_negative, subset=_fc_cols, props="color:#c79854")
847         .format(to_positive, subset=_fc_cols, na_rep="")
848         .format(rounder, subset=_val_cols, na_rep="")
849         .bar(
850             subset=(_res_final[~_res_final[(" ", " ", "Feature")].isin(group_cols)].index, _fc_cols),
851             align=0,
852             height=50,
853             vmin=_res_final.select_dtypes(include="number").min().min(),
854             vmax=_res_final.select_dtypes(include="number").max().max(),
855             width=50,
856             props="width: 50px;",
857             color=["#DABB8E", "#1ca3ec"],
858         )
859         .set_table_styles(
860             [
861                 {"selector": "th.col_heading.level0", "props": [("text-align", "center")]},
862                 {"selector": "th.col_heading.level1", "props": [("text-align", "center")]},
863                 {"selector": "th.col_heading.level2", "props": [("text-align", "center")]},
864                 {"selector": "tr", "props": "line-height: 21px;"},
865                 {"selector": "td,th", "props": "line-height: inherit; padding-top: 0; padding-bottom: 0;"},
866             ]
867         )
```

Export styled pandas table using dataframe-image
explainability/report.py line 898-900

```
898     if save_tbl_img:
899         os.makedirs(f"figs/{site_id}", exist_ok=True)
900         dfi.export(_tbl, f"figs/{site_id}/tbl_{site_id}_{issue_date}.png", table_conversion="selenium", dpi=200)
```

4. **Please provide the machine specs and time you used to train your model and to produce the communication outputs.**

- CPU (model): Core i5-1135G7
- GPU (model or N/A): N/A
- Memory (GB): 8GB
- OS: Windows
- Train duration: ~3 hours for all 720 models (20-fold years x 9 model variants x 4 losses)
- Inference duration: less than 3 minutes for a single issue date and 26 sites (not including data download time)
- Software used to prepare the forecast summary: Python

5. **Anything we should watch out for or be aware of in using your model (e.g. code quirks, memory requirements, numerical stability issues, etc.)?**

In hindcast setting, we have the ideal condition where all the data is available, However, in operational setting, it will be more challenging since there's a possibility that the data will not be available on the forecast issue date because of the operational issue. Consequently, the list of features used for SHAP summary might vary from date to date depending on which models are used as the ensemble members.

6. **Did you use any tools for data preparation or exploratory data analysis that aren't listed in your code submission?**

   I use plotly and matplotlib to quickly explore the data and do post evaluation of the model

7. **What are some other things you tried that didn't necessarily make it into the final workflow (quick overview)?**

   Including more contextual data, using geo-visualization, and adding more details of how specific ensemble member model behaves. However, since the maximum forecast summary is only two pages, I only included the two most important summaries.

8. **If you were to continue working on your explainability/communication solution for the next year, what methods or techniques might you try in order to build on your work so far? Are there other metrics or visualizations you felt would have been very helpful to have?**

   Possibly, I will try to add a more interactive summary page and geo-visualization.