# III. Model documentation and write-up

You can respond to these questions either in an e-mail or as an attached file (any common document format is acceptable such as plain text, PDF, DOCX, etc.) **Please number your responses.**

1. Who are you (mini-bio) and what do you do professionally?

Francisco Mendez

I have been working as a data scientist in Crescendo Technology for over a year. Most of my work is related to sport modelling, and customer behavior analysis.

Kunnal Kainth

I have been working as a data scientist in Crescendo Technology for more than 2 years. My work consist on building data science tools such as reports and packages in R that can help the company make better use of the data.

1. High level summary of your approach: what did you do and why?

The approach used consisted in summarizing the measurements of each process by stage, dividing each process recorded into different groups, according to the phases provided in the test set and training one random forest for each group.

1. Copy and paste the 3 most impactful parts of your code and explain what each does and how it helped your model.

The first code section take the average of the final rinse total turbidity liter by object id.

The second section is the function that summarizes all the impactful measures of a process by phase

The last code section discards the outliers in train (all process which the target variable is bigger than 7,500,000





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

We tried classifying processes that could be outliers and used in a separate step a model to predict the final turbidity.

One of the features of our final model was the object\_id average of the observed final rinse total turbidity liter, some objects have few observations, this might lead to an inaccurate estimation for future processes on this objects, so we try adjusting using a Bayesian approach to obtain a new version of the average but it didn’t work well on the model.

We tried creating new inputs in the models using the loadings of the principal component analysis

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

Yes, we played a lot with visualization charts, some made sense and some didn’t, we create time series for particular variables in a single process, created some biplots to understand the relation between variables.

1. How did you evaluate performance of the model other than the provided metric, if at all?

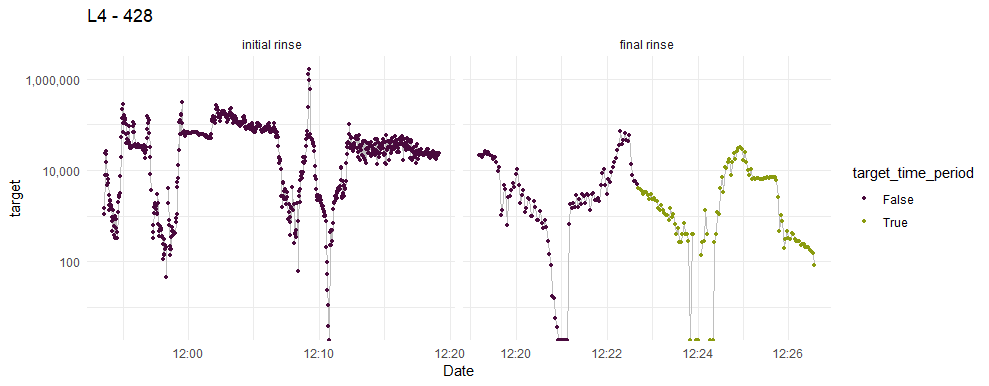
We used the MSE and MAE on addition to the metric provided, although all of the were highly correlated

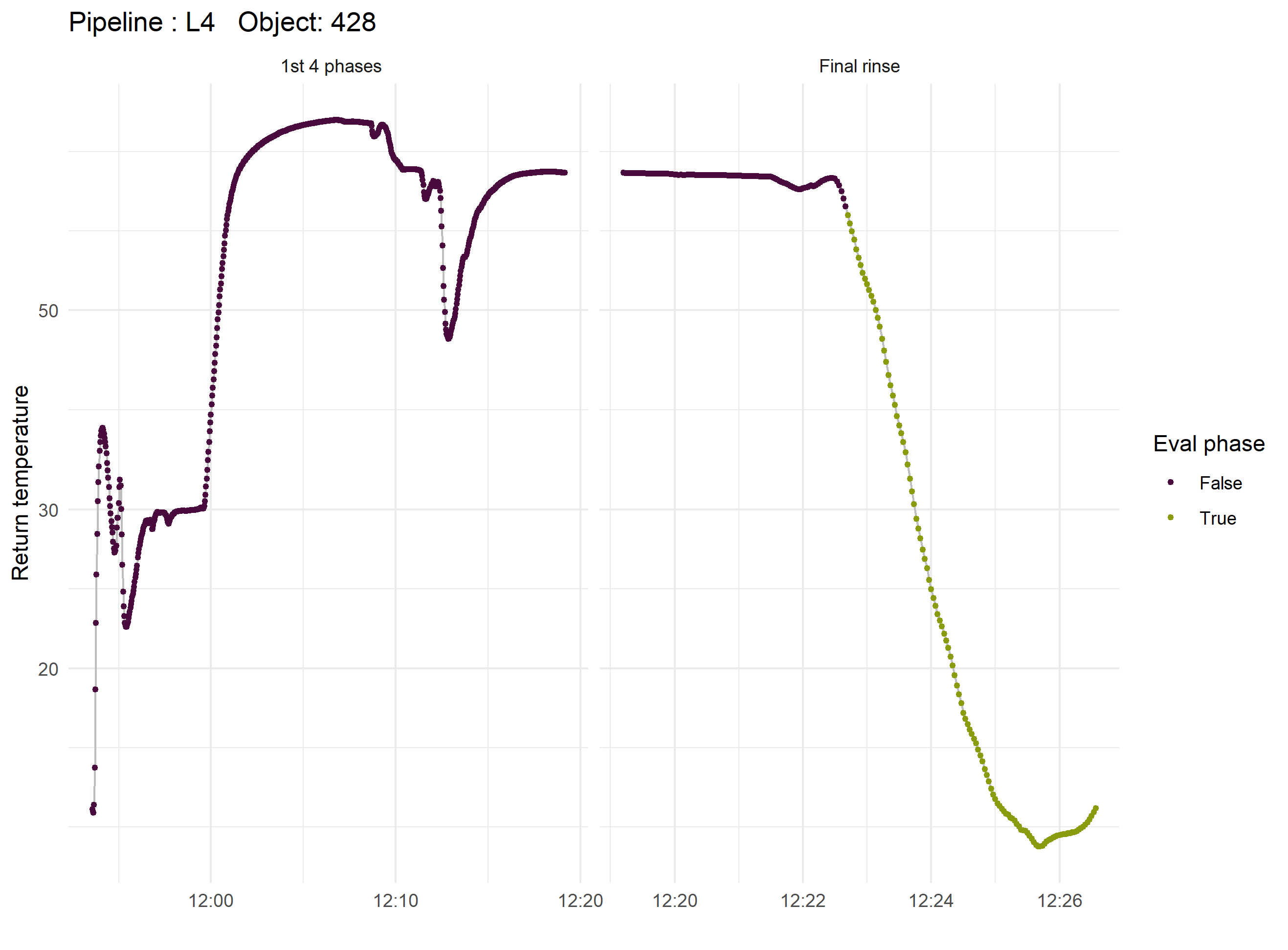
1. Anything we should watch out for or be aware of in using your model (e.g. code quirks, memory requirements, numerical stability issues, etc.)?

It is straightforward to use it and it should run without any problem considering that the libraries required are installed

1. Do you have any useful charts, graphs, or visualizations from the process?

Our most valuable chart were shared with you in the second stage report but we can share her an example of the visualization tool used to understand the target variable





1. If you were to continue working on this problem for the next year, what methods or techniques might you try in order to build on your work so far? Are there other fields or features you felt would have been very helpful to have?

We would work on understanding better the pipelines variable and perhaps creating separate models for processes in certain pipelines

Work on the outlier model until we could be able to incorporate to our current workflow