Model Info Sheet

Section 1: Information about paper or report

1. Author(s): Names of the authors of the paper or report
2. Title of the paper or report which introduces the model
3. DOI or permanent link to the paper or report (for example, link to arxiv.org webpage)
4. License: Under which license(s) are the data and/or model shared?
5. Email address of the corresponding author

Section 2: Scientific claim(s) of interest

6. Does your paper make a generalizable claim based on the ML model? If yes, what is the scientific claim? For example, “Our ML model can be used to diagnose Covid-19 using chest radiographs of adult patients”.

If there are multiple claims, list each claim in a new line, along with a claim number.

We assume that the S-Bahn Stuttgart network does not change compared to the one related to the training dataset. Then, provided the knowledge of the delay at 2 stations before the station “event”, we can predict with >80% accuracy if the train will be on-time or not at the station “event”.

7. Is the scientific claim made about a distribution or population from which you can sample? If yes: (a) what is the population or distribution about which the scientific claim is being made? (b) What is the sample used for the study? For example, “(a) Population: adult patients with symptoms of Covid-19. (b) Sample: We use a random sample of adult patients who present at a U.S. based hospital between April 2020 and June 2020”.

If there are multiple scientific claims, list your answer for each claim in a new line, corresponding to their claim number in Q6.

Note: *A difference between the population and the set from which the sample is drawn could highlight potential generalizability failures, which are related to but distinct from leakage.*

a) The network of the S-Bahn Stuttgart if unchanged compared to the trained model.
b) Journeys of the S-Bahn Stuttgart between 01.09.2017 and 19.10.2017

8. Does the scientific claim only apply to certain subsets of the distribution mentioned in Q6? For example, “Our model works on chest radiographs of U.S.-based adult patients and might not generalize to radiographs taken in other places or using different machines.”

If there are multiple claims, list your answer for each claim in a new line, corresponding to their claim number in Q6.
No, the claim applies to all lines in the S-Bahn Stuttgart network.

Section 3: Train-test split is maintained across all steps in creating the model

9. Train-test split type: How was the dataset split into train and test sets? (For example, cross-validation; separate train and test sets).

If your model does not have a separate test set, it could suffer from leakage due to overfitting.

The dataset has been randomly split into 50% training and 50% test. For teaching purposes, the random seed has been fixed in order to always obtain the same results. Cross-validation has been used on the training dataset (90% actually training and 10% holdout).

10. Are there duplicates in the dataset? If yes, explain how duplicates are handled to ensure the train-test split.

If duplicates from the training set are included in the test set, your model could suffer from leakage. The higher the percentage of duplicates in the test set, the more severe the leakage.

Duplicated rows are removed from the dataset during the data preparation phase.

11. In case the dataset has dependencies (e.g., multiple rows of data from the same patient), describe how the dependencies were addressed (for example, using blockcross validation).

If dependencies across the train-test split are not addressed, your model could suffer from leakage. The higher the number of rows in the test set with dependencies, the more severe the leakage.

The events are dependent on each other (e.g. delayed arrival at station X of train Y impacts the arrival at station W of train Z), but still, the events are treated as being independent.

12. List all the pre-processing steps used in creating your model. For example, imputing missing data, normalizing feature values, selecting a subset of rows from the dataset for building the model.

All the mentioned actions are part of pre-processing. For more details, analyse NB1.

13. How was the train-test split observed during each pre-processing step? If applicable, use a separate line for each step mentioned in Q12.

If the train-test split is not maintained during all pre-processing steps, your model could suffer from leakage.

The train-test split is performed after the pre-processing steps in Q12. From that point on, the training and test datasets are handled separately.

14. List all the modeling steps used in creating your model. For example, feature selection, parameter tuning, model selection.

All the mentioned actions are part of modeling. For more details, analyse NB3-4 and "handlers".

15. How was the train-test split observed during each modeling step? If applicable, use a separate line for each step mentioned in Q14.

If the train-test split is not maintained during all modeling steps, your model could suffer from leakage.

The modeling steps in Q14 are applied to the training dataset only. The test dataset is not involved in the modeling, only in the evaluation.

16. List all the evaluation steps used in evaluating model performance. For example, cross-validation, out-of-sample testing.

Cross-validation is used during training. After training the model is evaluated on the separate test dataset.

17. How was the train-test split observed during each evaluation step? If applicable, use a separate line for each step mentioned in Q16.

If the train-test split is not maintained during all evaluation steps, your model could suffer from leakage.

See Q16: The training and test datasets are kept separate.

Section 4: Test set is drawn from the distribution of scientific interest.

18. Why is your test set representative of the population or distribution about which you are making your scientific claims?

If the test set distribution is different from the scientific claim of interest (listed in Q7), your model could suffer from leakage.

The test set is representative of the population.

19. Explain the process for selecting the test set and why this does not introduce selection bias in the learning process.

Selection bias (for example, only choosing data from a given geographic location but expecting your model's performance to generalize to all locations) can lead to leakage.

The test set is the result of a random split over the whole dataset.

20. In case your model is used to predict a future outcome of interest using past data, detail how data in the training set is always from a date earlier than the data in the test set.

In predictions about future outcomes of interest, using data from the future to predict in the training set the past in the test set is a form of leakage. Data in the training set should always have timestamps of an earlier time than those in the test set to avoid leakage.

The model does predict a future outcome (delay of upcoming trains). Yet, the training and test dataset are not split considering future/past events. The predicted quantity (delay) is assumed as “unknown” in the test dataset.

Section 5: Each feature used in the model is legitimate for the task

21. List the features used in the model, alongside an argument for their legitimacy. A legitimate feature is one that would be available when the model is used in the real world and is not a proxy of the outcome being predicted. You can also include this list in an appendix and reference the relevant section of your Appendix here.

Not relevant.

For example, “Patient age: We include this feature in our ML model for hypertension diagnosis since patient age is easily available in a clinical setting”.

An example of a feature that should not be included (for illustration only; you do not need to include these in your model info sheet): “Anti-hypertensive drugs: We do not include the use of anti-hypertensive drugs as a feature in our ML model for hypertension diagnosis since that information is only available after diagnosis and would not be available when a new patient presents with symptoms of hypertension.”

Note: *You do not need to list each feature used in your model here. However, you must provide an argument for the legitimacy of each feature included in your model to ensure that your model does not suffer from leakage due to illegitimate features. For example, “our model only uses data from the previous year as features. For instance, to predict civil war in 2017, we only use lagged features from the year 2016. Since these features are always available in advance of when we want to make predictions using our model, none of these features can lead to leakage.”*