**Experimentation**

**Overview**

To test the data, a wide range of models are used. In each, their usage is compared to uses previously identified in the literature. The concept behind the model is explained and the method used is detailed. Results are then thoroughly discussed. Both classification – whether or not a train is delayed – and regression – by *how much* a train is delayed – are addressed. Finally, the influence of weather will be tested.

**Pre-processing**

Pre-processing is necessary before using the data output from the LOAD stage. This mostly takes the form of encoding categorical variables, dropping unnecessary columns, and setting datatypes to save space.

Special attention is paid to datetimes, as they are likely to have significant influence. Dates are split into their respective parts: year, month, day, day of week. Hour and minutes are encoded into a single minutes variable. Different variables are used for both origin and destination. These are encoded cyclically[[1]](#footnote-1) by converting a single column to two: one for and the other for . In this way a variable is encoded as a unit circle, so Monday and Sunday are adjacent, for instance, as are 11pm (23 hours) and 12am (0 hours).

Two target columns are then constructed using the difference between the scheduled arrival time (STA) and actual arrival time (ATA). The first, whether or not a train is delayed, is used for classification. The second, the amount by which a train is delayed, is used for regression. Erroneous values are also discarded. In about 1% of records, a error in the ETL pipeline resulted in ATAs a day *before* STAs. This occurred on a small number of trains that run over two days.

Some element of locality is also likely to improve results. Categorically encoding origin and destination TIPLOCs is impractical, as there are approximately 80,000. Nor is there any structure to their designations. Instead, their corresponding STANOX area is used. There are 89 STANOX areas[[2]](#footnote-2), proxies for geographical areas.

Dropped columns are mostly IDs, and those used to construct the columns discussed above. The actual time of departure (ATD) must also be dropped, as it could not be known in real-life data ahead of time. Also dropped are three metadata columns with too many NaN values: *sleepers*, *reservations*, and *branding*.

Finally, the following columns are one-hot encoded: *status, category, power\_type, timing\_load, ATOC\_code, seating, origin\_stanox\_area,* and *destination­\_stanox\_area*. *characteristic* and *catering* were encoded during schedule parsing in the TRANSFORM stage.

Numeric variables (*length, speed,* and all generated datetime columns) are then scaled to between 0 and 1. The datatype of each column is then minimised. For the majority, *np.uint8* is used. The resulting dataset comprises over 5 million rows and 284 columns.

**Metrics**

**Classification**

Accuracy is a misleading metric for imbalanced datasets. Approximately 7% of trains are delayed; models can simply always predict ‘not delayed’ to achieve roughly 93% accuracy, and duly learn to do so. This is known as the accuracy paradox. Instead, other metrics should be used. *Precision* is the ability of the classifier not to label as positive a sample that is negative. *Recall* is the ability of the classifier to find all positive samples. The fraction of samples from a class which are correctly predicted by the model. There is a natural trade-off between the two.

The -measure ( and measures) are a weighted harmonic mean of precision and recall. A measure reaches its best value at 1 and its worst score at 0. With , , and the recall and precision are equally important. True positives, false positives, true negatives, and false negatives may be succinctly represented in a confusion matrix and visualised using heatmaps.

The receiver operating characteristics (ROC) curve is a plot which shows the performance of a binary classifier as a function of its cut-off threshold (the true positive rate against the false positive rate for various thresholds).

**Regression**

Regression metrics include the mean absolute error (MAE), root mean squared error (RMSE), mean squared error (MSE), R-Squared (R2), mean squared percentage error (MSPE), mean absolute percentage error (MAPE), and root mean squared logarithmic error (RMSLE).

We will need the models for these metrics, alas. But we can pickle them once trained, because that’s the expensive part.

MAE is more robust to outliers.

RMSE is more sensitive to outliers.

**Classification**

Initial results (using only two days’ worth of data) were very promising: better than 90% accuracy using out-of-the-box sklearn classification models. This is misleading, however. Approximately 7% of trains are delayed. The models have simply learned to predict ‘not delayed’. Accuracy is a misleading metric for imbalanced datasets; instead, recall, precision and -measure should be used.

That said, decision trees often perform well on unbalanced datasets, and did so here.

For nominal: one-hot, hashing, leave-one-out, target. Avoid one-hot for high cardinality columns and decision-tree based algorithms.

For ordinal: binary, one-hot, leave-one-out, target

Nominal: no order.

Ordinal: order.

Use target / leave-one-out. Preferentially the later. Avoids contamination. Useful for nominal and ordinal data. Classification.

**Decision trees**

Decision trees reputedly suffer from one-hot encodings.

Characteristic and catering have to be one-hot encoded; it is their natural representation.

**Sources**

* <https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce>
* https://medium.com/@george.drakos62/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0

1. https://ianlondon.github.io/blog/encoding-cyclical-features-24hour-time/ [↑](#footnote-ref-1)
2. https://wiki.openraildata.com//index.php?title=STANOX\_Areas [↑](#footnote-ref-2)