Predicting a process

Dataset used:
BPI challenge 2012

2IOI0 - group 19

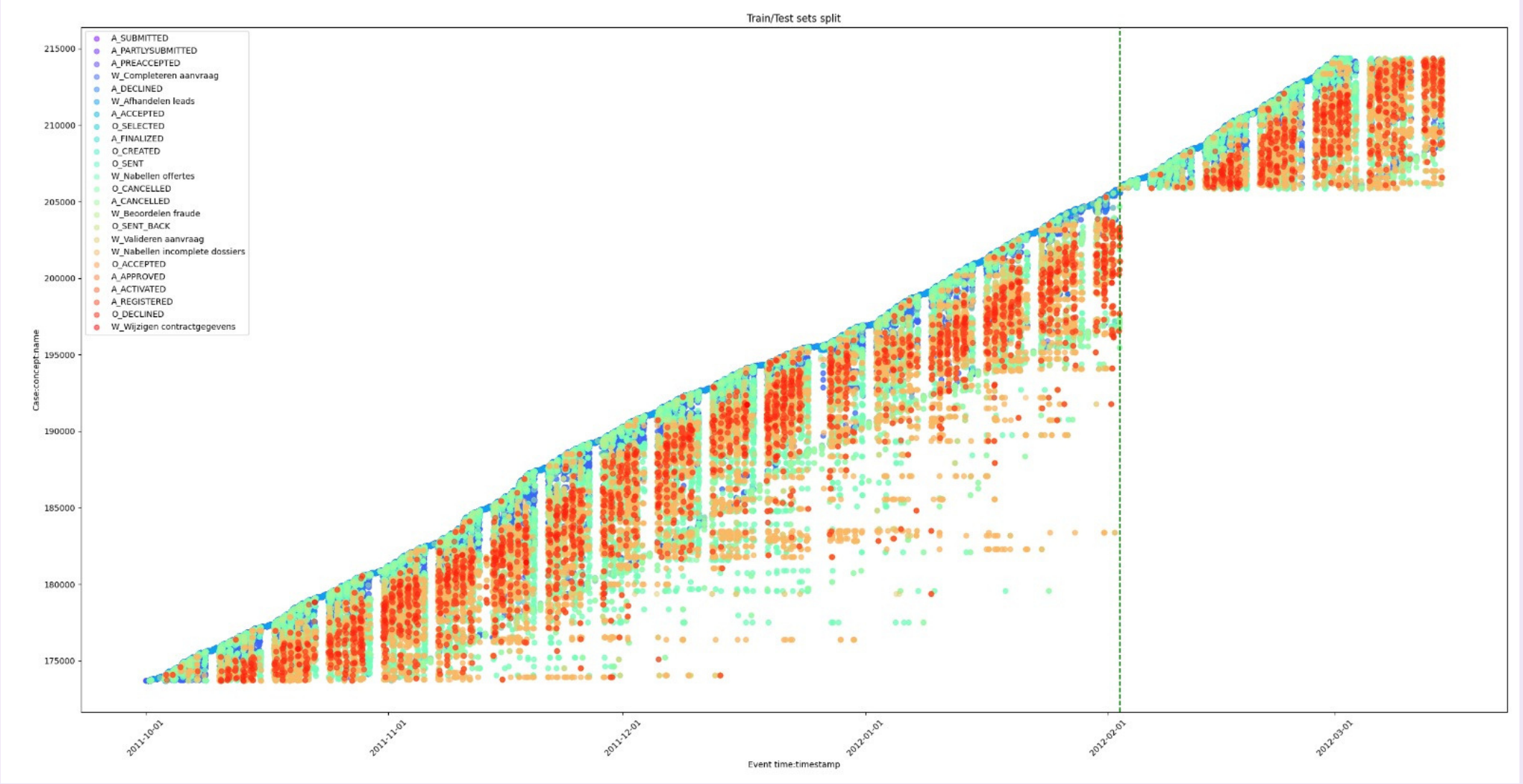
Maciej Bober	1809628	**Bartosz Wiewióra**	1852167
Jarmo Boer	1844482	**Efsane Yildiz**	1738777
Jennie Vermeijlen	1429337	**Misra Yilmaz**	1801511

Data exploration

- data comes from a loan application process at a bank
- most activity takes place on weekdays, but some on weekends
- a process prediction could improve the workflow of the bank
- some actions depend on work delivered by the bank, others depend on customers undertaking some actions, this could have consequences for the process
- when an (automated) action takes place outside working hours, it is repeated in the morning

Train/test split

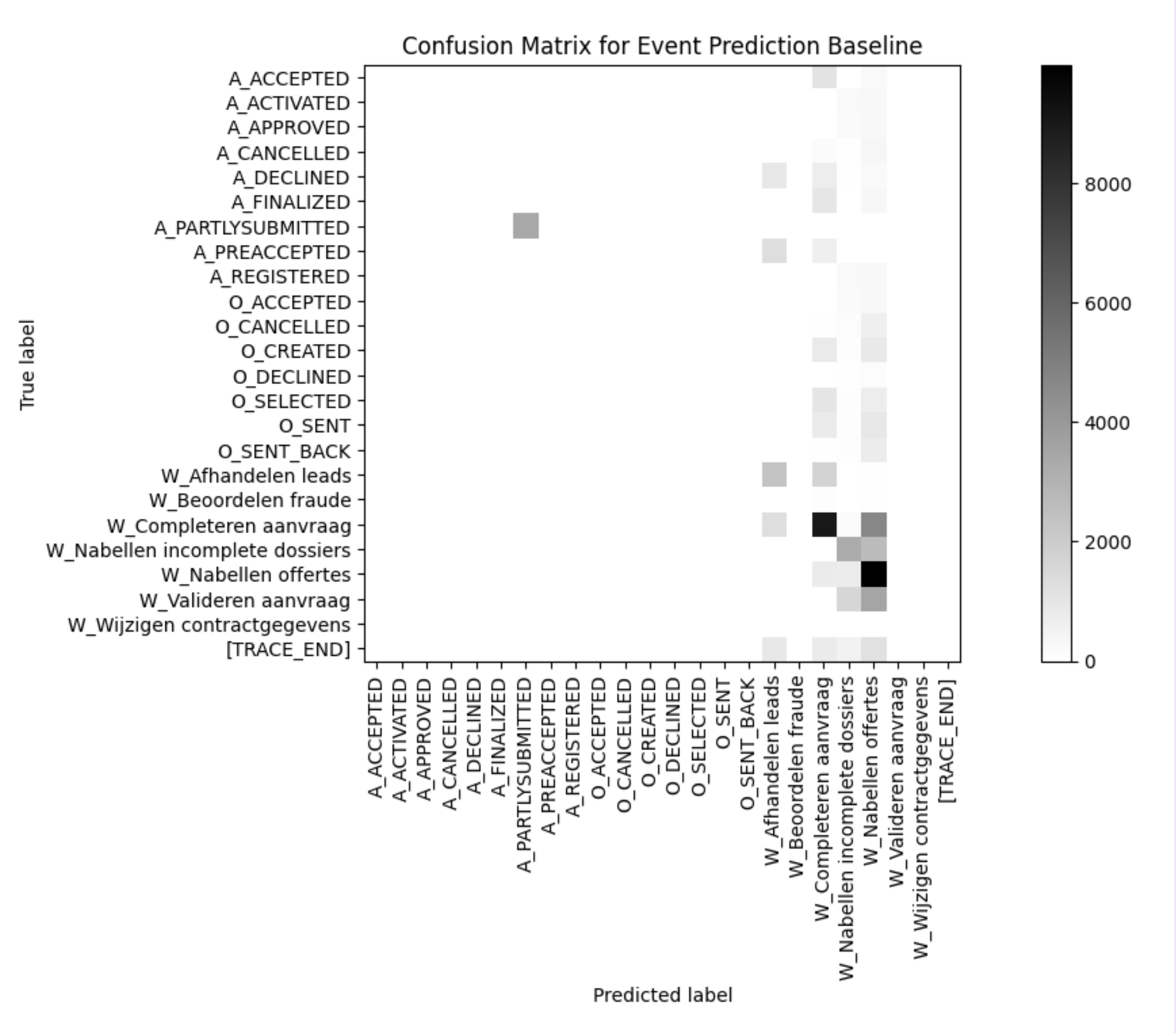
- made at 75%/25% order by time
- deleted traces that exist in both train and test set



Baseline

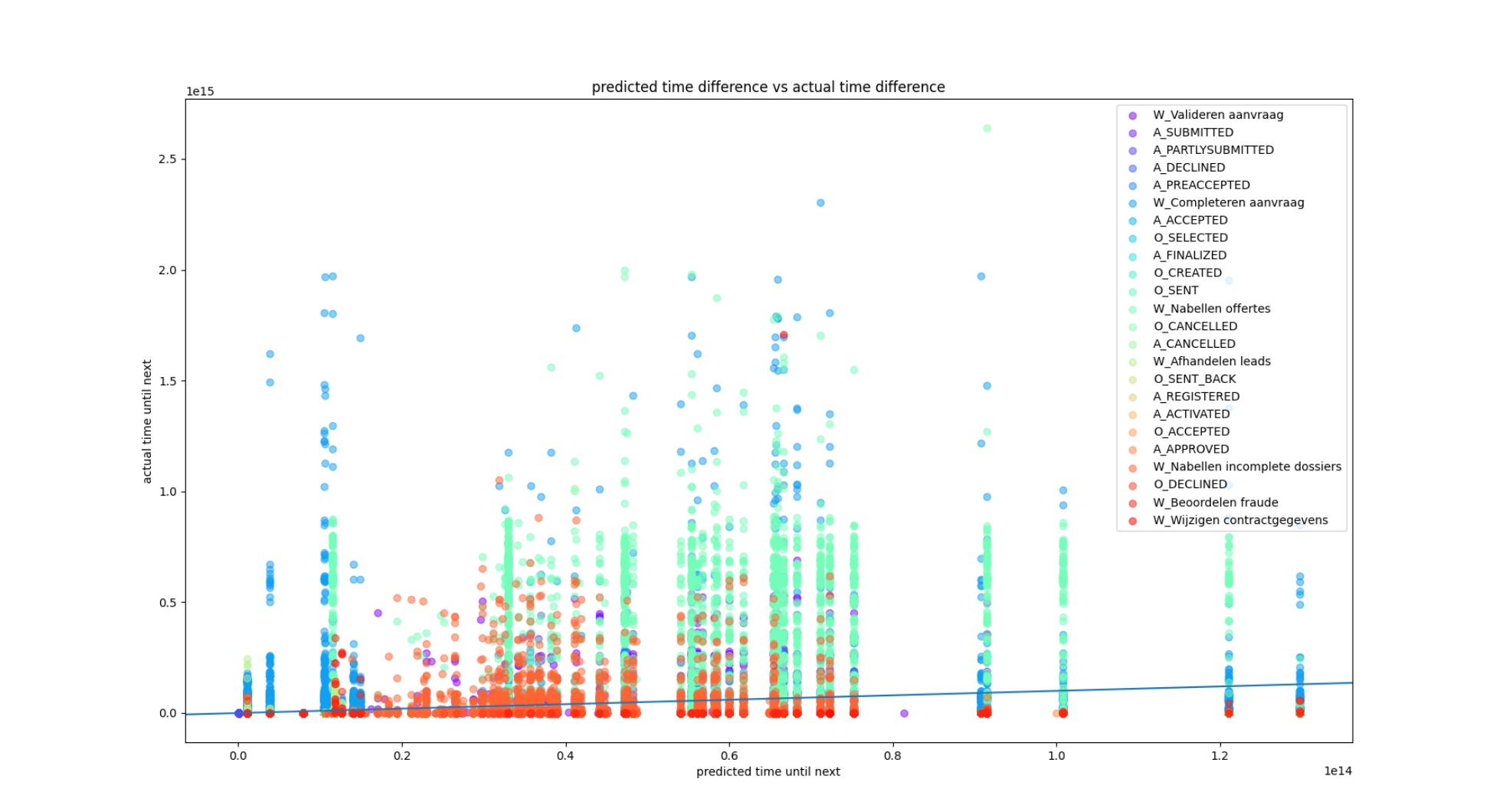
Baseline for activity

The baseline was made using the most common event per relative index. This resulted in a not very accurate prediction except for W_completeren aanvraag, which is one of the most occuring events in the dataset.



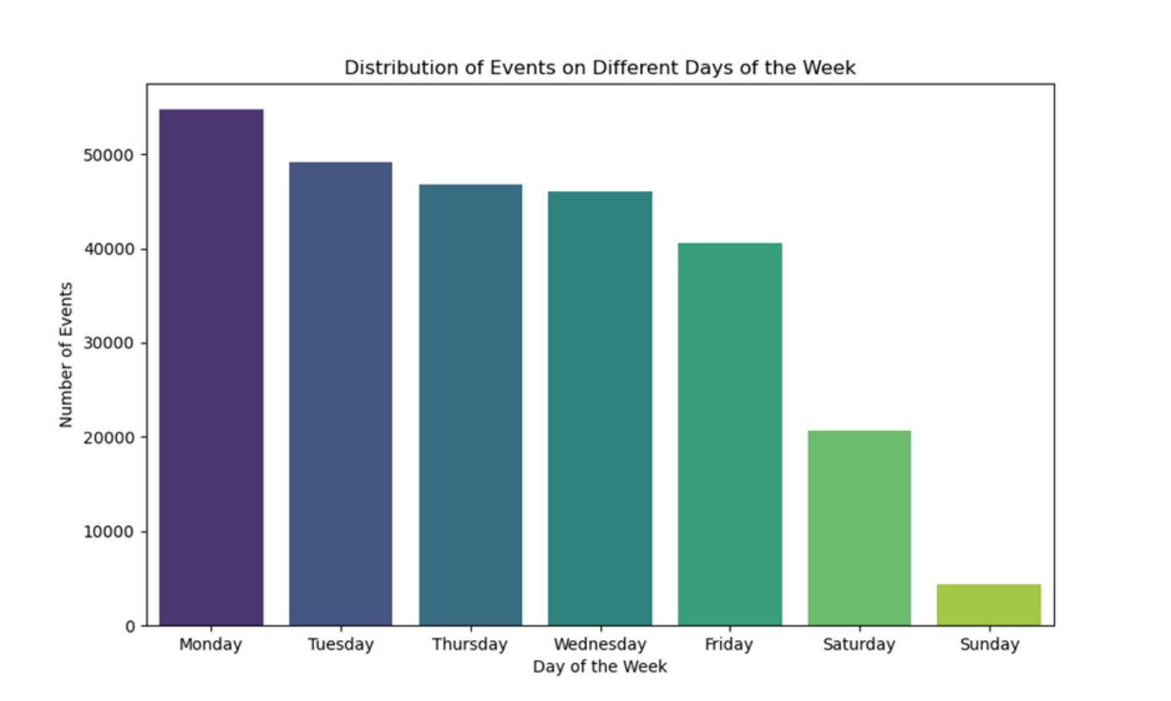
Baseline for time

- using average time per relative index
- not accurate
- line shows the perfect prediction



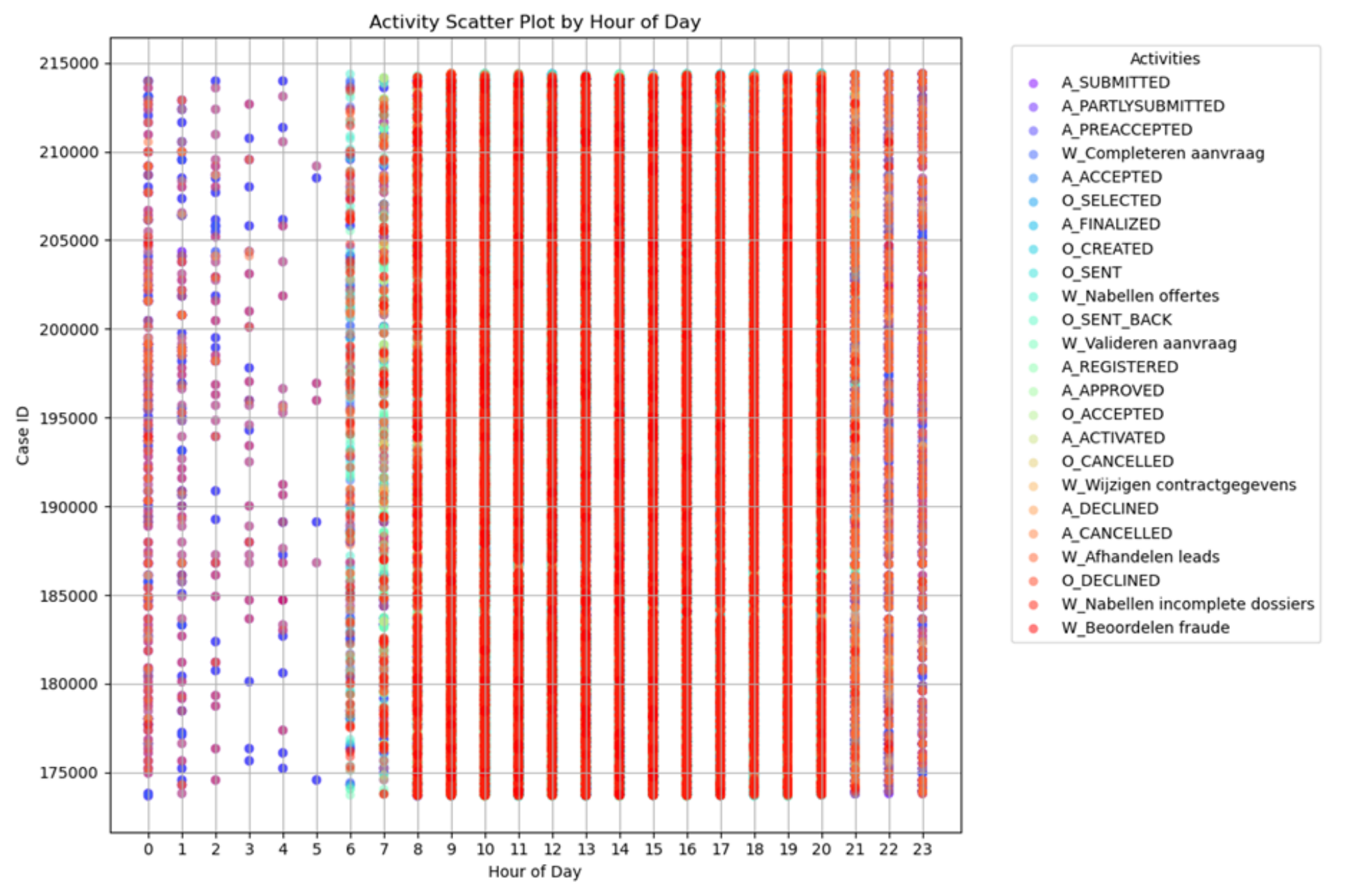
Feature engineering

The day of the week is an important feature, in both time and event prediction, since less events happen in the weekends and other events can happen. This is a logical feature, the data comes from a bank and those work less during the weekends than during the week.



The feature of the previous event is very valuable to event prediction. For the event prediction this gives a greater insight into what was happening before the prediction took place. This gives the tool a larger reach in overseeing what happened and increased the tools accuracy.

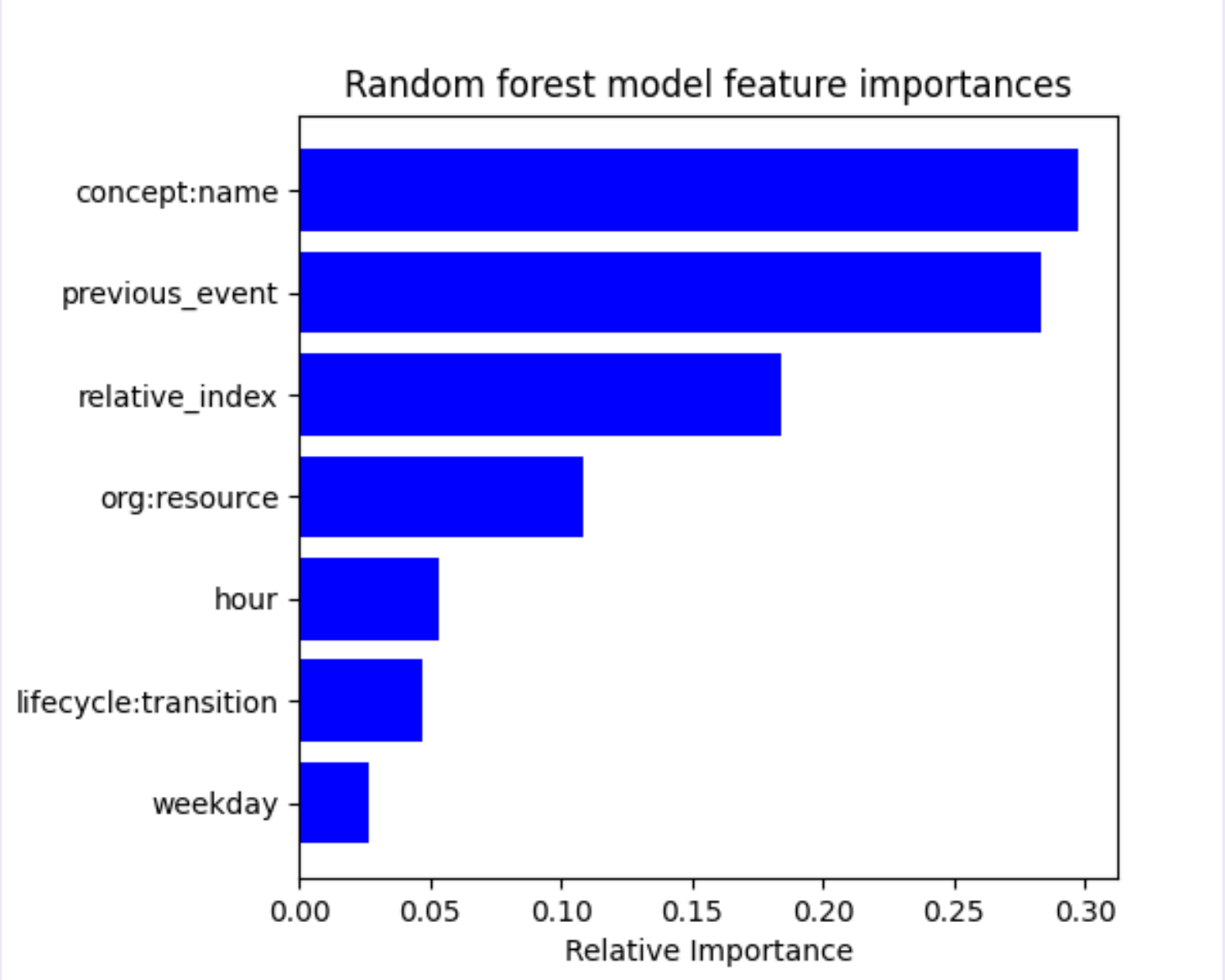
For time prediction we make use of the time of day. This feature is of importance because there are less actions that take place during the night than during working hours. This is also explainable, since most of the work is done by people, either the bank or the customers, and these generally sleep at night, less actions take place at this time.



Overall the plot does not clearly show how much difference there is during working hours and without. However, there is still a clear difference that can be seen.

Feature importance

- previous event is very important for the event prediction, as mentioned above
- the current event is also very important, this is because the follow up event needs to be predicted based on the current event. In this way the tool can "see" what is happening now.
- For event prediction the time of the action is of less importance but still adds value to the prediction, since for instance it is less likely that a customer responds when it is late in the day.



The Tool

Prediction for time

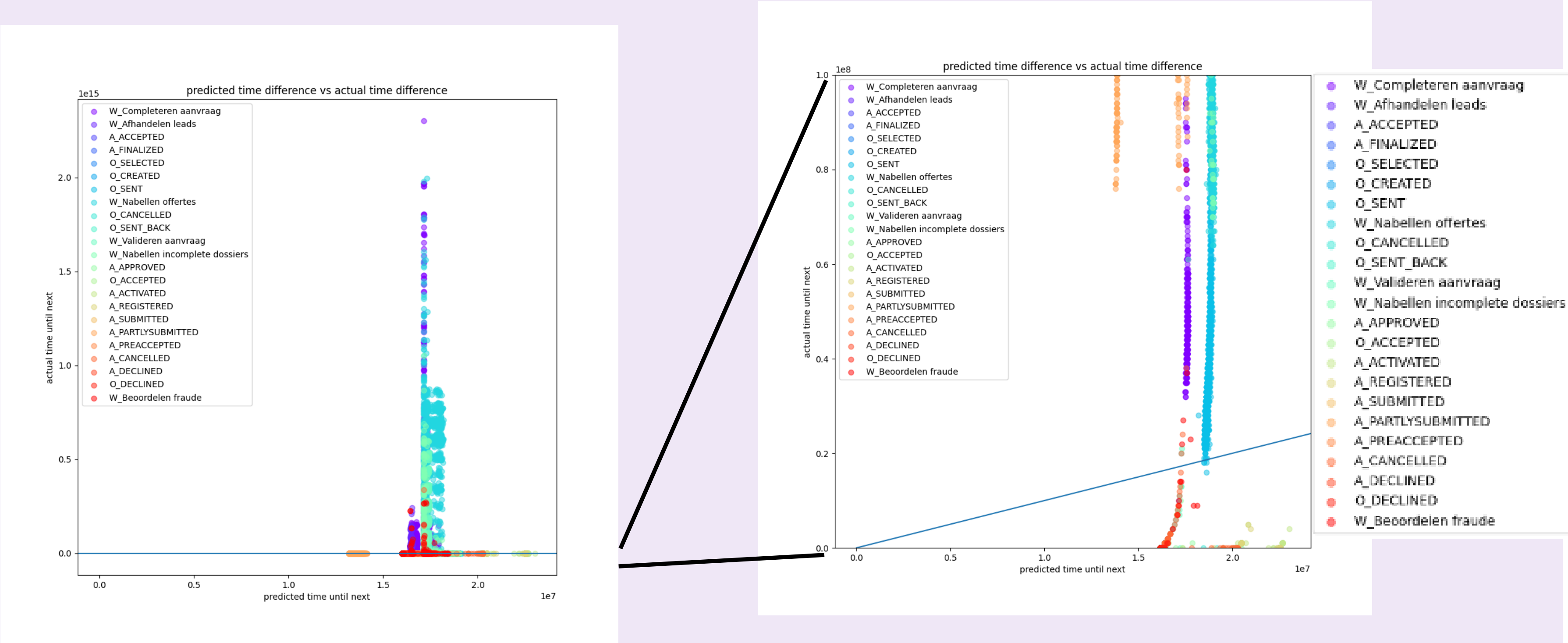
- Time range is very large (ms to days)
- It tends to skew towards bigger numbers
- No linear relation
- LSTM tends to keep on predicting the same value
- Featues make prediction worse

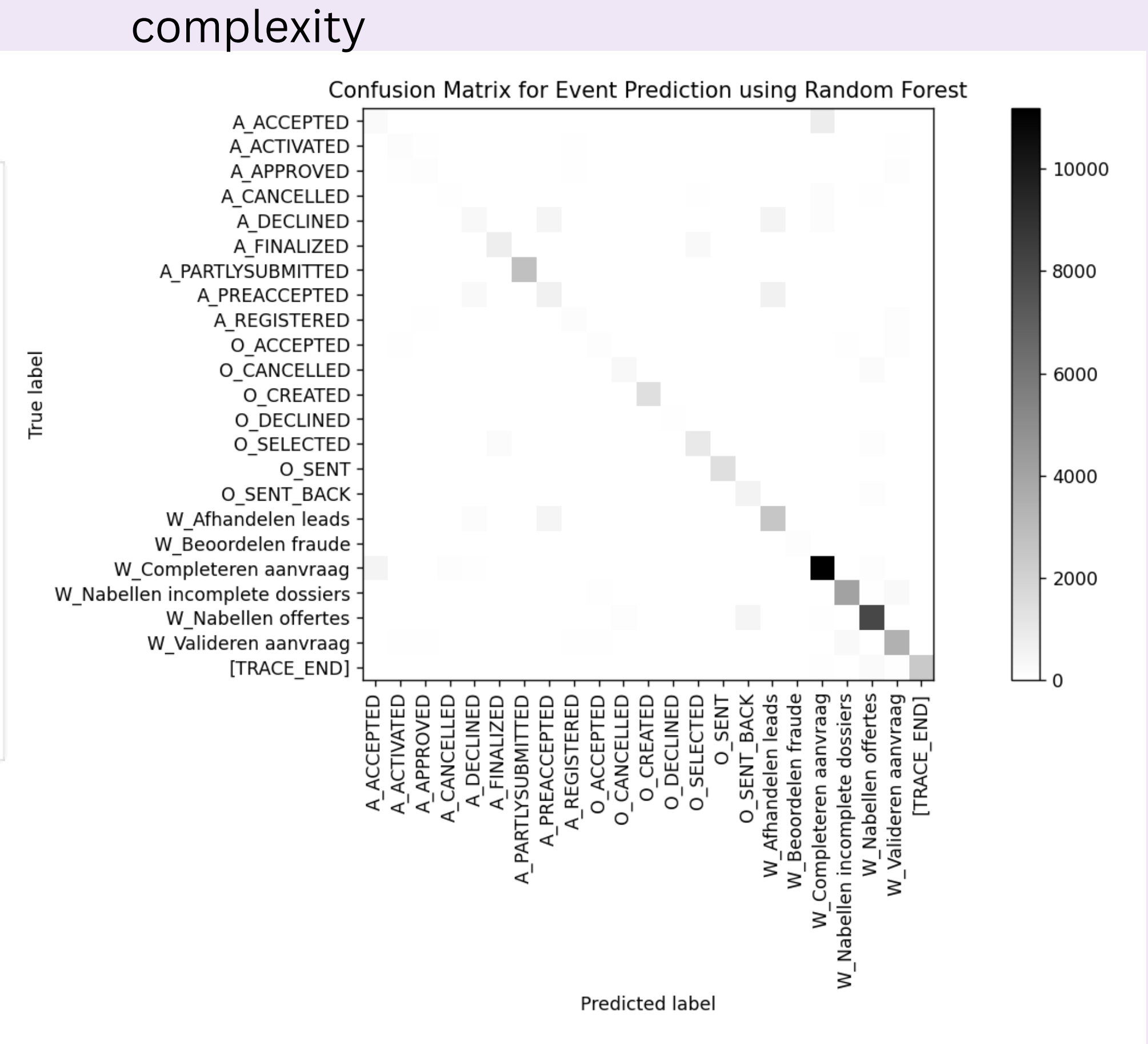
Prediction for activity

- random forest regression
  - aggrevated bootstrapping method
  - runs well on large datasets
- many trees correct for overfitting of one tree
- sklearn random forest with 100 trees
- good for predicting activities
- makes use of features given to the trees

Prediction for whole suffix

- LSTM model
- Suitable because of its characteristics like long memory storage or handling of sequential data
- it remembers inputs for a long time, so if something early in the process is of importance to later events it remembers
- High training time due to the model's complexity





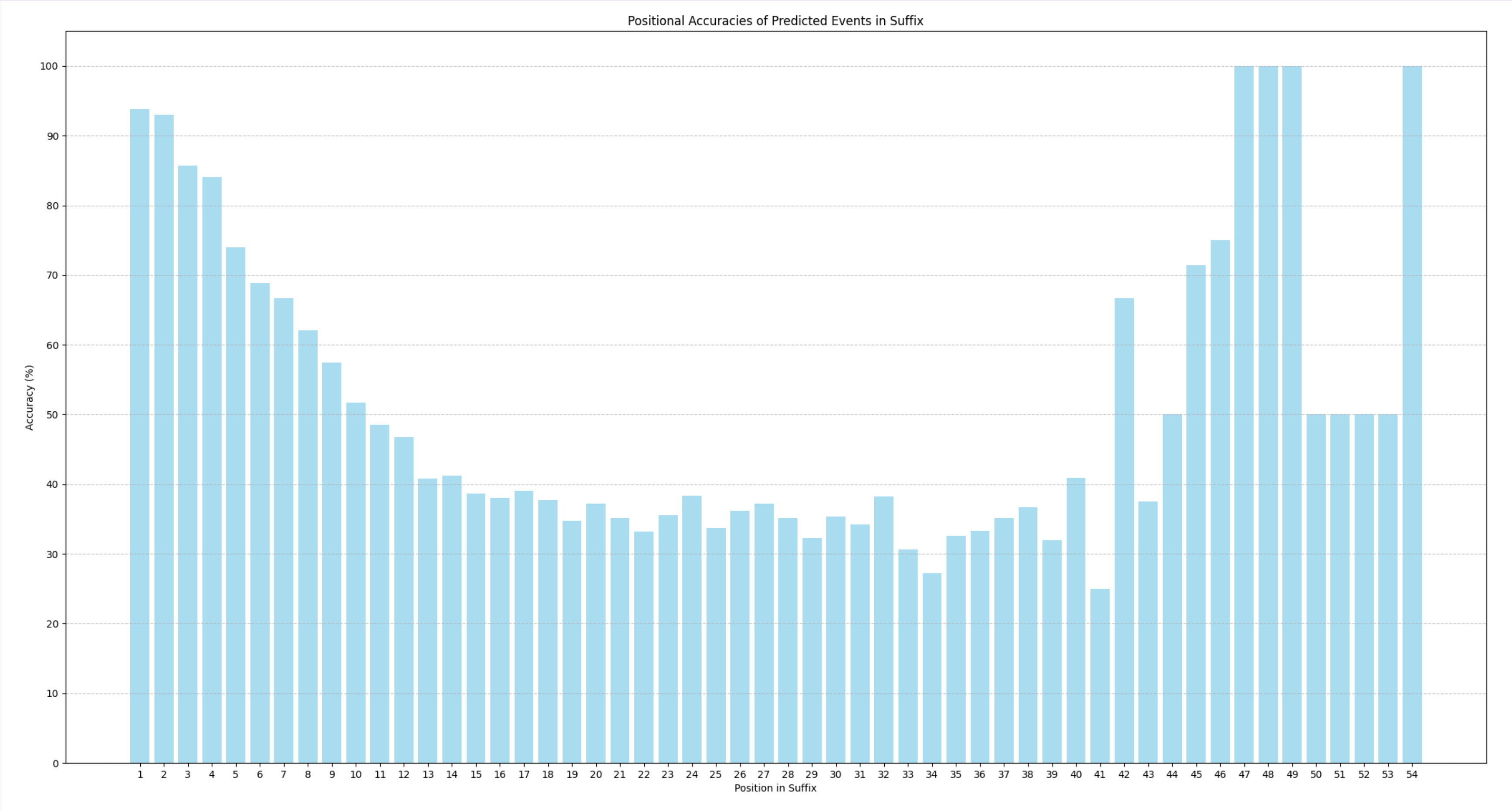
Evaluation - qualitative

When comparing the confusion matrix of the tool to the one of the baseline tool, an observation can be made that there are more activities being correctly predicted. The random forest tool greatly improved the ability in predicting the difference between W_nabellen offertes and w_complementeren aanvraag over the baseline tool. The tool however sometimes predict actions which in reality were A_DECLINED, this means that for importing our model into the buisness world this can be a hindrance

Evaluation - quantative

- Baseline: 43% accuracy
- Time prediction 18:17:43 MAE
- Next event prediction: 81.7%
- Suffix prediction: see graph

The suffix prediction model's accuracy is quite high, especially for the 1st, 2nd and even 3rd event, which all have higher accuracy than the random forest model. The accuracy decreases, as expected, with each next event, but remains better than the baseline up to the 12th position. As can be seen in the graph below, the accuracy starts to jump around a lot around 30+ this is because not many predictions were made, n is very small, and thus with one good prediction the accuracy is very high.



Reflection and limitations

Overall, we think that the chosen approach was accurate for the requirements of this project. The models performed reasonably well and their complexity was managable. However, the chosen models had their own downsides too. First of all, the suffix prediction could be performed with a stronger, more complex model. Also, we could investigate if the model does not overfit, as accuracy of around 92% seems rather high, even for the first event in the suffix.