Datathon Report

Team Tech-Nick

April 6, 2025

We have two types of data: tabular and free form text. In a first instance, we unified the data to a single format to allow for comparisons. We achieved this using LLM-based feature extraction with prompt engineering. The LLM parses the texts looking for pre-defined features to extract.

In a next step, we compare different entries containing the same information types using the Jaro-Winkler edit distance. This metric can find inconsistencies between data entries for string data. To avoid flagging mistakes which come down to additional whitespaces or leading/trailing whitespaces we manually filtered out such occurrences, since they generally did not lead to rejections in the training set. Names were concatenated to full names to account for incomplete orders (middle name before first name) in some data entries. In many cases, an edit distance different from 1 (one being a perfect match) would already result in a rejected client (i.e. a typo in an email or phone number). In other cases, it can be useful to compare results with the edit distance, for example when comparing nationality and country of passport (for example comparing French vs France).

To account for inconsistencies in the data which can not be captured by comparing entries in an edit distance, we hand-engineered features to catch mistakes in the data which we observed. The grounds for rejection which we identified and accounted for with such features are

- Incomplete salary information in employment history (Salary = 0)
- Expired passport (passport expiry date needs to be at least 6 months after 01.04.2025)
- Inconsistencies between MRZ and the data it contains
- Country codes inconsistent with nationality or international standard

We also compared the age listed in the client description (which we extracted with features) with the birth date listed in the tabular data but did not find an improvement in using this feature.

The resulting features, hand-engineered or edit distance based, were passed to a LightGBM model. For this model, we tuned the hyperparameters through cross- validation and tested on a holdout validation set.