Julius Baer Challenge

TEAM TECH-NICK

Pipeline
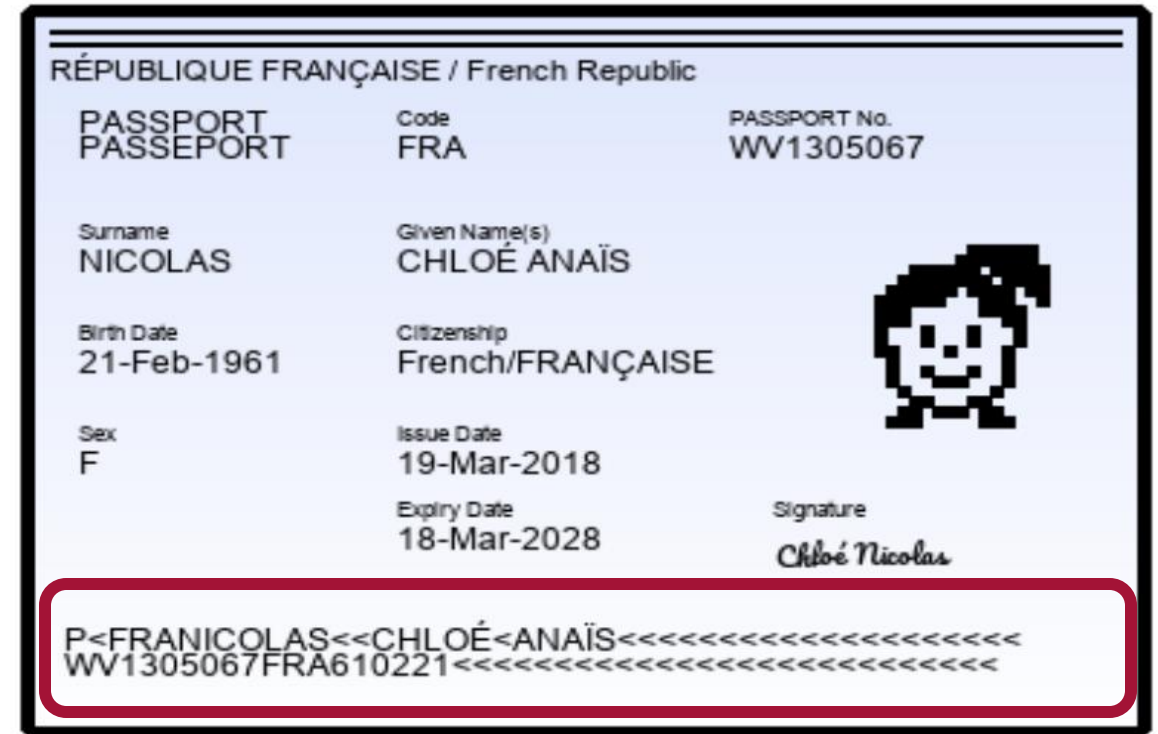
Jaro-Winkler Distance

- To compare different occurences for the **same information** type we use the **Jaro-Winkler** distance.

- It measures **edit distance** between strings.

- To compare the data in tabular form with the full texts we use **LLM** based **feature extraction**.

Feature Engineering

Additional hand-engineered features were added to account for

- Mismatches in passport MRZ
- Outdated passports
- Missing salary information
- Incorrect country codes

Model

- These features are then used with a **LightGBM** model
- Hyperparameters were tuned using cross-validation
- We achieved **0.95 accuracy** on a hold-out validation set, with most errors being False Negatives.