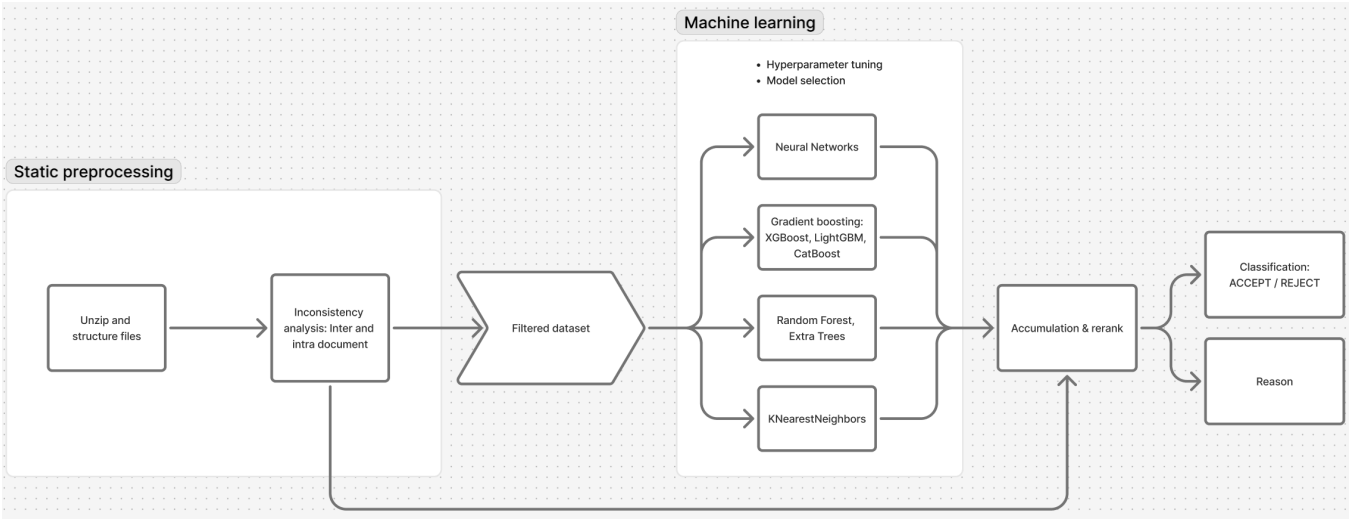
Solution report - creativity_is_all_you_need

Abstract

We present a composite solution pipeline for automated client evaluation in private banking. Our approach combines rule-based static preprocessing with machine learning-based decision-making to detect inconsistencies and assess client profiles. The preprocessing stage filters out faulty applications with high precision and explainability, while the learning stage leverages a diverse set of models to optimize classification performance. Despite challenges in feature relevance, our system achieves strong accuracy.

Solution pipeline

We constructed a composite pipeline that consist of a preprocessing stage, a machine learning stage and an accumulation of predictions for the final evaluation.



Static preprocessing

To check for faulty client documents, like forms with typos and deliberate false claims, we preprocess the client's JSON files to detect inconsistent data. This is a static analysis that not only filters out about 50% of rejected clients (and not a single accepted client), but also provides explainability at every step, improving customer experience. The preprocessing adds an entry to the client's data with whether they passed the checks and if not, all the flags raised by the analysis tool.
Intra document:

Document	Checks
Passport	Birth date is valid (valid format, in the past, at least 18 years old); Issue date is before expiry date and not in the future; Expiry date is in the future; Country, country code and nationality is consistent
Client profile	Address, email and phone number are correctly formatted; Education, higher education and employment history: Dates are consistent (chronological ordering, no overlap); If inheritance >0: inheritance details are not empty; If real estate >0: real estate details are not empty; Inheritance year in lifetime and in the past
Account form	First name + middle name + last name results in “name”; Address, email and phone number are correctly formatted; Country of domicile is consistent

Inter document:

Document set	Checks for consistency
Passport and client profile	Gender, nationality, passport issue and expiry date, birth date
Client profile and account form	Name, address, country of domicile, phone number, email address
Passport and account form	First name, middle name, last name
All three	Passport number

Machine learning methodology

To find the best machine learning method, we tested and tuned a large number of Machine Learning algorithms, including ones based on decision trees, neural networks and nearest neighbor search. The automated training of the methods, including hyperparameter tuning, is handled by `AutoGluon`. After the models are trained, a weighted composite of all model outputs is generated to enable cross-domain learning and best possible accuracy.

Results

- Average error on test split: 88.75%
- Average error on entire dataset: 93.25%
- False positives in static preprocessing: 162 (through noise)

Challenges and insights

We faced a variety of challenges while working on the problem. Our preprocessing detected 162 falsely accepted client whose passport was expired. Additionally, we had to do some data cleanup, for example due to non-existent currencies/not accepted currencies.

Approach for feature extraction

In an attempt to reduce dimensionality in our data and capture soft inconsistencies (like salaries that are way too high or low for certain positions and/or unexplained money sources), we crafted a feature engineering stage for our pipeline. The tool uses OpenAI's `gpt-4o-mini` in parallelized API calls and structured generation to extract information like (excerpt):

Client feature	Type	Examples	Default if not given
Degrees	one-hot encoded	Bachelor, other higher, Master, PhD, Postdoc	`[0,0,0,0,0]`
Maximum prestige of university degrees	range	1-5	0
Seniority level	range	Junior=1, Senior=2, Manager=3, Director=4, C-level=5, Chairman=6,	0
Consistency of professions with salaries	number	Difference to median salary for position divided by median	1

After the feauture extraction was complete, we realized that a lot of data points had very low correlation to the label. Indeed, while trying to fit our machine learning models to the extracted features, we couldn't achieve sufficient accuracy.