Match Profiler
software engineering

CREDIT RISK ASSESSMENT PLATFORM

PARC
CREDIT RISK ASSESSMENT PLATFORM

Within the Match Profiler Innovation Department, PARC is a platform for companies in the financial sector, which may be connected to the granting of credit, aiming to detect default according to the characteristics of the loans and those requesting them.

PLATFORM DEVELOPMENT

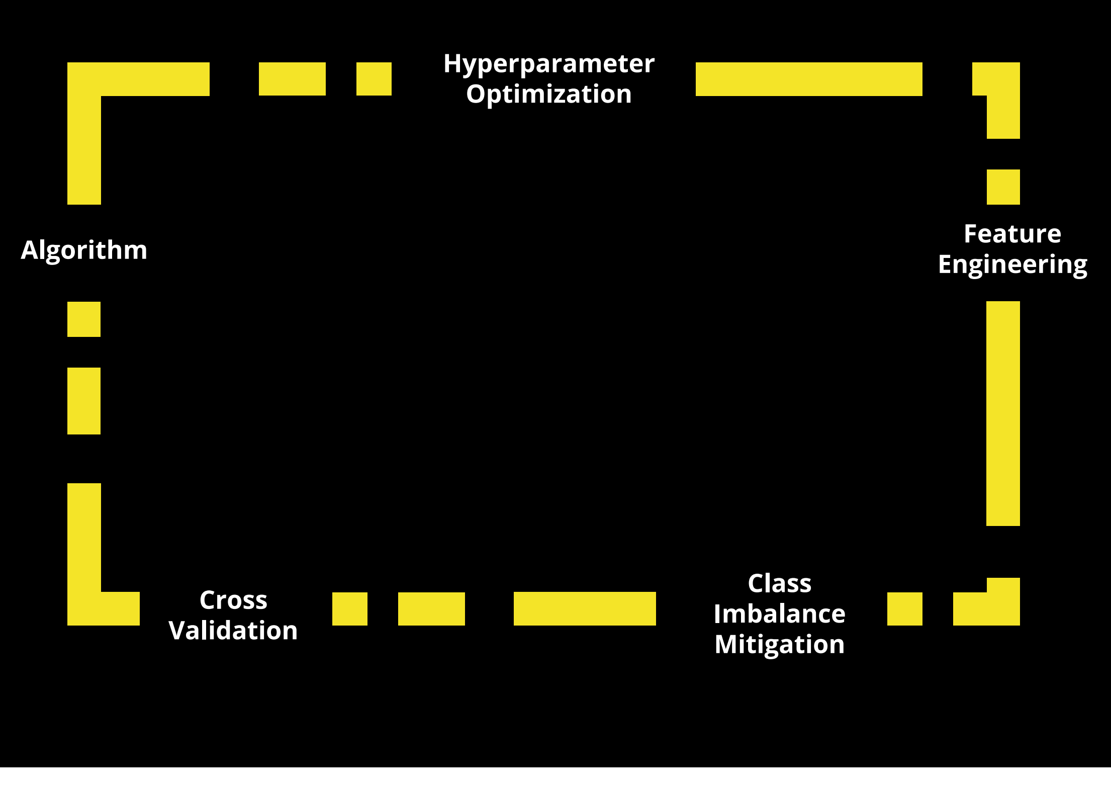
Selection of Datasets
Selection of datasets representing the problem under study. Anonymous, previously classified data.

Exploratory Data Analysis (EDA)
Data treatment, selection and transformation of variables: analysis of correlations, prevention of data leakage, treatment of outliers and missing values, treatment of categorical variables.

Benchmarking Classification Algorithms
In parallel with EDA, the choice of algorithm results from an iterative process at several levels in order to maximise model metrics. The characteristics of the datasets, with strong class imbalance, require special care in the choice and analysis of model performance metrics.

BENCHMARKING ALGORITHMS

- **Considered algorithms:**
  SVM, Random Forest, Gradient Boosting, LightGBM, CatBoost, XGBoost
- **Parameter optimisation:**
  Grid Search, Bayesian Optimization, Genetic Algorithms
- **Feature engineering:**
  Generation of new variables through transformations and aggregations
- **Class imbalance:**
  Mitigation of data imbalance (minority default) through SMOTE and oversampling
- **Cross validation:**
  Validation considering the ROC curve area and F1 Score as the most important metrics



WEB SERVICE

- The service core is a rating model optimised for maximum performance through C compilation.
- The rating model is available as a REST API in Web Services.
- The platform enables loan grading individually or in loan list batches.
- Considering the sensitivity analysis, there is an endpoint through which it is possible to obtain several score levels corresponding to variation ranges of amount and time.

CREDIT SCORE
WEB SERVICE – INPUT

The service is invoked by inputting a message composed of 21 numerical and qualitative variablest

- **Variables specific to the loan:**
  Amount, term, purpose, guarantor, co-petitioner, effort rate.
- **Applicant specific variables:**
  Credit history, employment status, qualifications, marital status, age, gender, number of dependents, number of loans taken out with the institution, other credits, foreign worker, residence, savings, existence of movable or immovable property, existence of service supply contracts (water, electricity, telecommunications, etc.).

CREDIT SCORE
WEB SERVICE – OUTPUT

The result of invoking the service includes a credit rating, combined with the metrics of the current version of the model that made the forecast.

- **Rating (default or otherwise)**
- **Model Metrics:**
  - (AUROC) – Area corresponding to the probability of the model classifying a true default case correctly. An area of 0.5 corresponds to a random model.
  - Precision $P = \frac{Tp}{Tp + Fp}$: within the universe of cases classified as positive (default) by the model, what is the proportion of correct cases. High P corresponds to low false positives.
  - Recall $R = \frac{Tp}{Tp + Fn}$: proportion of positive cases (default) correctly identified compared to the number of actual default cases. High R corresponds to low false negatives.
  - F1 Score $= 2\frac{P\,R}{P + R}$: harmonic mean of Precision and Recall.

ENDPOINTS

API has a set of specific scoring services, complemented with services related to the use of the platform itself, for audit trail, logging and invoicing. The scoring services consist of endpoints that allow interaction on a case-by-case basis, or alternatively, services capable of processing loan lists. The platform also offers the possibility of carrying out parametric studies, thus providing the study and evaluation of multiple risk scenarios.

Singular Endpoint:
Corresponds to a specific loan rating. It consists of the input of a message in JSON format.

Endpoint batch:
Allows the treatment of multiple loan cases. The input message consists of N cases. The answer contains the list of results composed of error or success codes and their classifications.

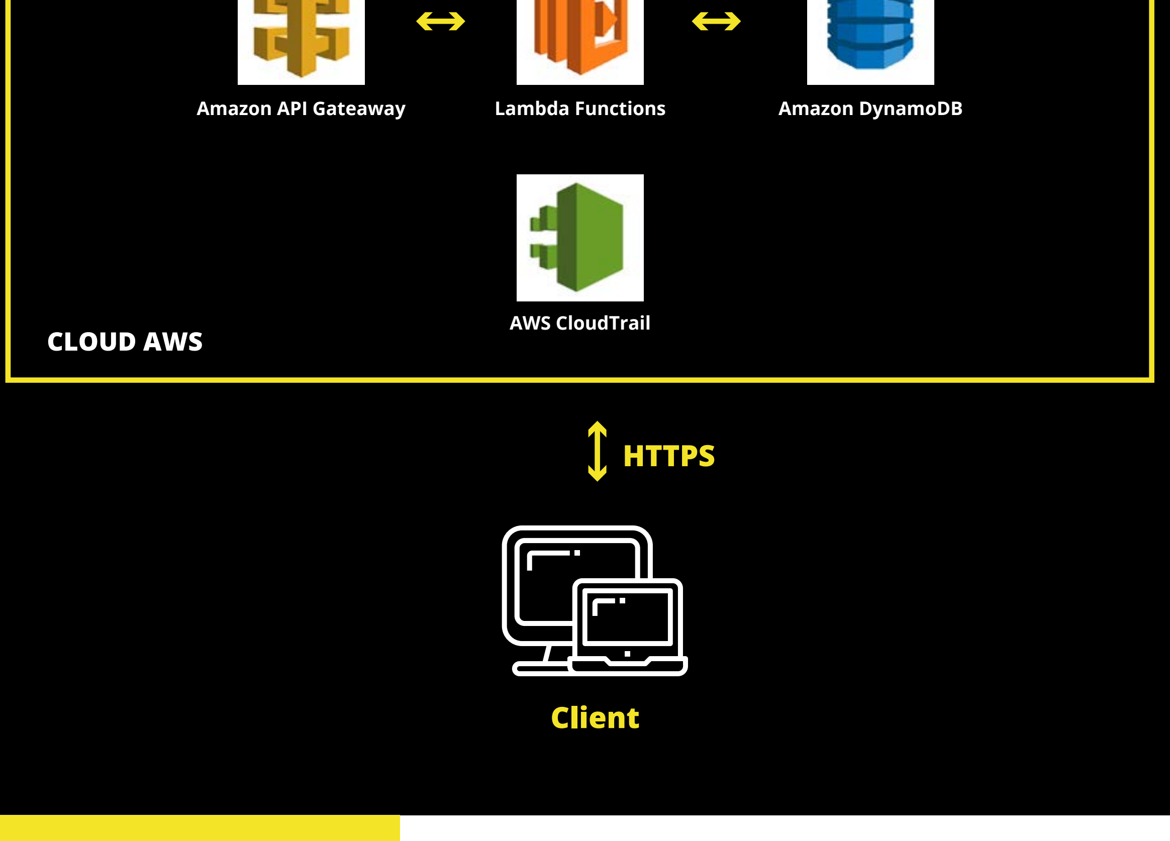
Endpoint simulation:
The input message includes 2 variation ranges for amount and deadline and the number of points to be considered in each range. The answer is a score map according to the amount and deadline.

TECHNICAL SOLUTION FEATURES

The API was developed in a serverless architecture following microservices standards.
Access is controlled by JWT or Amazon Cognito Authentication.
The communication is based on in-transit and persistence encryption mechanisms, i.e. on the database.
Invoking API is throttled in line with what is established in service contracts. Endpoints return appropriate code for back-off implementations by clients.

Serverless Solution:

- Scalable
- Resilient
- High availability



PARC
CREDIT RISK ASSESSMENT PLATFORM

TECHNICAL DATA

- XGBoost Algorithm
- Data processing with Oversampling and Deep Feature Synthesis
- Model compiled with Python Treelite
- AWS Serverless Architecture
- API REST

Cofinanciado por:
COMPETE 2020
PORTUGAL 2020
UNIÃO EUROPEIA
Fundo Europeu de Desenvolvimento Regional

Match Profiler
software engineering

CONTACT US