Match Profiler
software engineering

PLATAFORMA DE AVALIAÇÃO DE RISCO DE CRÉDITO

PARC
CREDIT RISK ASSESSMENT PLATFORM

No âmbito do Departamento de Inovação da Match Profiler, o PARC é uma plataforma destinada às empresas do setor financeiro, que possam estar conectadas à concessão de crédito e que tem como objetivo detetar o incumprimento, tendo em conta as características dos empréstimos e de quem os solicita.

DESENVOLVIMENTO DA PLATAFORMA

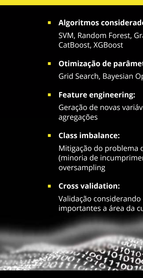
Seleção de Datasets
Seleção de datasets representativos do problema em estudo. Dados anónimos, previamente classificados.

Exploratory Data Analysis (EDA)
Tratamento de dados, seleção e transformação de variáveis: análise de correlações, prevenção de data leakage, tratamento de outliers e missing values, tratamento de variáveis categóricas.

Benchmarking de Algoritmos de Classificação
Em paralelo com EDA, a escolha do algoritmo resulta de um processo iterativo a diversos níveis, com vista a maximizar as métricas do modelo. As características dos datasets, com forte desequilíbrio (class imbalance) obrigam a um especial cuidado na escolha e análise das métricas da performance do modelo.

BENCHMARKING DE ALGORITMOS

- **Algoritmos considerados:**
  SVM, Random Forest, Gradient Boosting, LightGBM, CatBoost, XGBoost
- **Otimização de parâmetros:**
  Grid Search, Bayesian Optimization, Genetic Algorithms
- **Feature engineering:**
  Geração de novas variáveis através de transformações e agregações
- **Class imbalance:**
  Mitigação do problema de desequilíbrio dos dados (minoria de incumprimento) através de SMOTE e oversampling
- **Cross validation:**
  Validação considerando como métricas mais importantes a área da curva ROC e F1 Score



WEB SERVICE

- O core dos serviços é um modelo de classificação otimizado para máxima performance através de compilação em C.
- O modelo de classificação é disponibilizado na forma de API REST em WebServices.
- A plataforma permite a classificação de empréstimos de forma singular ou em batches de listas de empréstimos.
- Considerando a análise de sensibilidade, existe um endpoint através do qual é possível obter diversos níveis de score correspondentes a intervalos de variação de montante e prazo.

WEB SERVICE DE SCORE DE CRÉDITO – INPUT

A invocação do serviço faz-se através do input de uma mensagem composta por 21 variáveis numéricas e qualitativas.

- **Variáveis específicas do empréstimo:**
  Montante, prazo, propósito, fiador, co peticionário, taxa de esforço.
- **Variáveis específicas do solicitante:**
  Histórico de crédito, situação laboral, qualificações, estado civil, idade, género, número de dependentes, número de créditos contraídos na instituição, outros créditos, trabalhador estrangeiro, residência, poupanças, existência de bens móveis ou imóveis, existência de contratos de fornecimento de serviços (água, eletricidade, telecomunicações, etc.).

WEB SERVICE DE SCORE DE CRÉDITO – OUTPUT

O resultado da invocação do serviço é composto por uma classificação do crédito, juntamente com as métricas da versão corrente do modelo que realizou a previsão

- **Classificação (incumprimento ou não)**
- **Métricas do modelo:**
  - (AUROC) – Área correspondente à probabilidade de o modelo classificar corretamente um caso de verdadeiro incumprimento. Uma área de 0,5 corresponde a um modelo aleatório.
  - Precision $P = \frac{Tp}{Tp + Fp}$: dentro do universo dos casos classificados como positivos (incumprimento) pelo modelo, qual a proporção de casos corretos. P elevado corresponde a baixos falsos positivos.
  - Recall $R = \frac{Tp}{Tp + Fn}$: proporção de casos positivos (incumprimento) corretamente identificados face ao número casos reais de incumprimento. R elevado corresponde a baixos falsos negativos.
  - F1 Score $s = \frac{2PR}{P + R}$: média harmónica de Precision e Recall.

ENDPOINTS

A API tem um conjunto de serviços específicos de scoring, sendo complementada com serviços relativos à utilização da própria plataforma, destinados a audit trail, logging e faturação. Os serviços de scoring consistem em endpoints que permitem interação caso a caso, ou alternativamente, em serviços capazes de processar listas de empréstimos. A plataforma oferece também a possibilidade de se realizarem estudos paramétricos dando assim o estudo e avaliação de múltiplos cenários de risco.

Endpoint singular:
Corresponde à classificação de um empréstimo específico. Consiste no input de uma mensagem em formato JSON.

Endpoint batch:
Permite o tratamento de múltiplos casos de empréstimos. A mensagem de input consiste em N casos. A resposta contem a lista de resultados composta por códigos de erro ou sucesso e respetivas classificações.

Endpoint simulação:
A mensagem de input inclui 2 intervalos de variação para montante e prazo e o número de pontos a considerar em cada intervalo. A resposta é um mapa de score em função de montante e prazo.

BENCHMARKING DE ALGORITMOS

A API foi desenvolvida numa arquitetura serverless seguindo padrões de referência para microserviços.
O acesso é controlado por Autenticação JWT ou Amazon Cognito.
A comunicação assenta em mecanismos de encriptação em trânsito e em persistência, isto é, na base de dados.
As invocações à API são restringidas (throttled) em consonância com o estabelecido em contratos de serviço. Os endpoints retornam códigos apropriados para implementações de back-off por parte dos clientes.

Solução Serverless:

- Escalável
- Resiliente
- Alta disponibilidade

PARC
CREDIT RISK ASSESSMENT PLATFORM

FICHA TÉCNICA

- Algoritmo XGBoost
- Tratamento de dados com Oversampling e Deep Feature Synthesis
- Modelo compilado com Python Treelite
- Arquitetura Serverless AWS
- API REST

Colaboram por:
COMPETE 2020 | PORTUGAL 2020 | UNIÃO EUROPEIA