Resumo

Nos dias de hoje, nota-se um crescimento exponencial da concessão de crédito a nível mundial devido às mudanças de comportamento dos clientes, dos avanços tecnológicos e de políticas financeiras. A elevada procura pela concessão de crédito pode contribuir para a rentabilidade das instituições financeiras, contudo, esse crescimento também pode trazer consequências, como o aumento do risco de inadimplência (ou de incumprimento). Deste modo, surge a necessidade de utilização de ferramentas para pontuar o risco de crédito, *credit score*, sendo que neste estudo, o score representa a probabilidade do indivíduo não cumprir as suas obrigações conforme o seu comportamento passado.

Nesta dissertação serão utilizados dados reais disponibilizados na plataforma Kaggle pela American Express (Amex) na competição "American Express - Default Prediction", cujo objetivo da competição é determinar o risco de o cliente entrar ou não em incumprimento.

Assim, serão explorados os dados disponibilizados, através de uma descrição detalhada, e identificados os processos necessários no seu pré-processamento para, posteriormente, poder aplicar-se modelos de Machine Learning (ML) como a rede neuronal Feedforward, Long Short-Trem Memory (LSTM), Gated Recurrent Unit (GRU), Transformer e Light Gradient Boosting (LGBM).

De acordo com os dados disponibilizados pela Amex, constatou-se que estes encontram-se desequilibrados, ou seja, a proporção de clientes cumpridores é muito superior à proporção de clientes incumpridores. Consequentemente, para enfrentar este desafio foi explorada a possibilidade da aplicação da métrica Amex como função custo. No entanto, esta métrica não poderá ser implementada como função custo pois não é diferencial, logo, será somente aplicada na avaliação dos modelos.

Relativamente, aos modelos explorados verificou-se que a GRU e LGBM foram os modelos que obtiveram um melhor desempenho na generalização para novos dados, com uma métrica Amex para dados de teste de 80,02% e 79,94%, respetivamente. Porém, a combinação de modelos, modelos ensemble, foram os que alcançaram os melhores resultados, destacando-se o modelo que combina a GRU e LGBM com uma métrica Amex de 80,36%.

Palavras-chave: Credit Score, Risdo de Inadimplência, Machine Learning, Métrica Amex

Abstract

Nowadays, there has been an exponential growth in the granting of credit worldwide due to changes in customer behavior, technological advances and financial policies. The high demand for lending can contribute to the profitability of financial institutions, but this growth can also have consequences, such as increasing the risk of default. Thus, the need arises to use tools to score credit risk, credit score, and in this study, the score represents the probability of the individual not fulfilling their obligations according to their past behavior.

This dissertation will use real data made available on the Kaggle platform by Amex in the "American Express - Default Prediction" competition, the aim of which is to determine the risk of a customer defaulting or not.

Thus, the data provided will be explored, through a detailed description, and the necessary pre-processing processes will be identified so that ML models such as the Feedforward neural network, LSTM, GRU, Transformer and LGBM can subsequently be applied.

According to the data provided by Amex, it was found to be unbalanced, i.e. the proportion of compliant customers is much higher than the proportion of non-compliant customers. Consequently, to address this challenge, the possibility of applying the Amex metric as a cost function was explored. However, this metric cannot be implemented as a cost function because it is not differential, so it will only be applied when evaluating the models.

Regarding the models explored, it was found that GRU and LGBM were the models that performed best when generalizing to new data, with a Amex metric for test data of 80.02% and 79.94%, respectively. However, the combination of models, ensemble models,

achieved the best results, with the model combining GRU and LGBM standing out with a Amex metric of 80.36%.

Keywords: Credit Score, Default Risk, Machine Learning, Amex metric