

Avaliando previsões de *default* em empréstimos

João Pedro Abreu de Souza¹

¹Instituto de Computação – Universidade Federal Fluminense (UFF)

jp-abreu@id.uff.br

Abstract. *This paper model a decision's tree on the problem of determine default on loans, the fundamental problem on any credit institution.*

Resumo. *Este artigo modela uma árvore de decisão no problema de determinação de default em empréstimos, o problema fundamental em qualquer instituição fornecedora de crédito.*

1. Introdução

Dentro do mercado de crédito, protagonizado por Bancos, cooperativas de crédito e demais instituições financeiras, a determinação do risco de inadimplência (i.e. *default*) é crucial pois desse risco advém toda a decisão de conceder crédito, e caso conceda, a qual custo deve ser fornecido, de forma que as pessoas ou instituições adimplentes suportem a perda das pessoas ou instituições inadimplentes. O custo mínimo, dado um cohort específico, que deve ser acrescido ao custo dos adimplentes, é dado, segundo [1], $EAD \times PD \times LGD = \text{Expected Loss}$. Este artigo relata a utilização de um conjunto de 22999 registros de empréstimos disponível publicamente em [2] para treinar uma árvore de decisão usando [3], [4] e [5] para tratamento dos dados. Os dados para reprodução do artigo encontram-se em [7]

2. Limpeza dos dados

O dataset escolhido por esse artigo[6] necessitou de uma fase de limpeza bem curta, removendo do csv uma primeira linha de header que estava logicamente duplicada. Como é possível constatar [2], essa linha possuía a mesma informação do header seguinte, porém com colunas chamadas X1, X2, etc. De posse do dicionário de dados que a página fornece, é possível interpreta-lo, mas foi mais simples apenas não usa-lo em prol de legibilidade. Originalmente foram considerados para esse artigo outros 4 datasets, porém descartados na fase de limpeza, pois possuíam características discriminatórias, como local de moradia, ou possuíam múltiplas colunas nulas, a fim de contemplar os múltiplos pagamentos ou a falta deles. Além disso os datasets descartados, que permanecem nos documentos fornecidos junto ao presente artigo, possuíam uma cardinalidade muito menor que o escolhido. Como o modelo deve ser capaz de ser explicado para quem tiver seu crédito negado, além de atender a restrições legais e éticas, o dataset atual foi finalmente escolhido. A divisão entre as classes é de 17826 casos de default e 5173 casos de adimplência.

3. Divisão treinamento

A divisão dos dados originais em treinamento e teste ficou dividido em 30% para testes e 70% para treino. A divisão entre teste e treino obedece a divisão padrão do scikit, que é por padrão de 25%, mas considerando um conjunto maior em comparação aos datasets descartados, foi fornecido um pouco mais aos testes, para aproveitar o tamanho.

4. Avaliação de performance

| Classe | Precisão | Revocação | F1-Score | Suporte |
|-----------------|----------|-----------|----------|---------|
| 0 | 0.83 | 0.80 | 0.81 | 5375 |
| 1 | 0.37 | 0.41 | 0.39 | 1525 |
| Precisão Média | | 0.60 | | |
| Revocação Média | | 0.60 | | |
| F1-Score Médio | | 0.60 | | |
| Acurácia | | 0.71 | | |
| Suporte Total | | 6900 | | |

Tabela 1. Relatório de classificação para o arquivo Loan-data-sem-primeira-linha-change-dependent-variable.csv

5. Conclusão

Arvore de decisão é uma escolha apropriada para estimar problemas que precisem de explicação concreta e não abstrata, porém mesmo com um dataset com pouco desbalanço comparado a um dataset maior, ainda oferece uma performance relativamente ruim com os parâmetros utilizados. A avaliação foi extremamente rápida, então o trade-off entre performance e baixo custo de operação, seja energético ou ambiental, deve ser levado em conta.

Referências

- [1] Investopedia. *What Is Exposure at Default (EAD)? Meaning and How To Calculate*. Acesso em: 20 de junho de 2024. Disponível em: https://www.investopedia.com/terms/e/exposure_at_default.asp
- [2] kaggle. *Loan Data*. Acesso em: 20 de junho de 2024. Disponível em: <https://www.kaggle.com/datasets/jakeshbohaju/loan-data>
- [3] scikit-learn. *scikit-learn*. Acesso em: 20 de junho de 2024. Disponível em: <https://scikit-learn.org/stable/>
- [4] pandas. *pandas*. Acesso em: 20 de junho de 2024. Disponível em: <https://pandas.pydata.org/>
- [5] numpy. *numpy*. Acesso em: 20 de junho de 2024. Disponível em: <https://numpy.org/>
- [6] dataset. *dataset*. Acesso em: 20 de junho de 2024. Disponível em: <https://www.kaggle.com/datasets/jakeshbohaju/loan-data>
- [7] repositório. *repositorio*. Acesso em: 26 de junho de 2024. Disponível em: <https://github.com/petrolifero/trabalhoML>