

PROPOSTA DE *FRAMEWORK* PARA A MODELAGEM DE REGRESSÃO LOGÍSTICA NA CLASSIFICAÇÃO DO RISCO DE CRÉDITO

André Aoun Montevechi

andremontevechi@unifei.edu.br

Rafael de Carvalho Miranda

rafael.miranda@unifei.edu.br

André Luiz Medeiros

andremedeiros@unifei.edu.br

Jonathan Serafim Lúcio

serafim.epr@gmail.com

Universidade Federal de Itajubá
Avenida BPS, 1303, Itajubá – MG

RESUMO

A concessão de crédito é uma atividade vital da indústria financeira. É indispensável portanto, não só para o seu funcionamento e sucesso, mas para a manutenção do equilíbrio do sistema creditício, que a modelagem de risco de crédito seja precisa e capaz de avaliar bem a probabilidade de inadimplência de potenciais devedores com base em dados históricos. Conhecido como *credit scoring*, trata-se de um problema comumente abordado por meio de algoritmos de classificação, sendo a regressão logística a referência na indústria. A modelagem do problema, contudo, estende-se muito além da escolha do classificador, e existem diversas etapas e métodos capazes de aperfeiçoar a performance de modelos de *credit scoring*. Desta forma, neste artigo são analisados os efeitos de diferentes etapas de modelagem sob a regressão logística para a construção de modelos de *credit scoring*, buscando entender como criar modelos eficientes partindo desde algoritmo de classificação, e propor diretrizes para tal.

PALAVRAS-CHAVE. *Credit Scoring, Machine Learning, Regressão logística.*

Tópicos. SIM - Simulação

ABSTRACT

Credit granting is a vital activity in the financial industry. Models that are capable of accurately assessing the likelihood of default by potential borrowers based on historical data are of utmost importance for their success, as for the balance of the credit system as a whole. The credit scoring problem is commonly addressed through classification algorithms, with logistic regression being the industry standard. Modelling this problem, however, involves more than choosing a classifier, and there are several different steps and methods capable of improving the performance of credit scoring models. Consequently, in this article, we analyze the effects of different modeling steps on logistic regression for building credit scoring models, aiming to understand how to create efficient models starting from this popular classification algorithm and proposing guidelines for doing so.

KEYWORDS. First keyword. Second keyword. Last keyword.

Topics. SIM - Simulation

1. Introdução

A modelagem do risco de crédito tem papel importante na gestão de riscos financeiros. Através de um processo conhecido como *credit scoring*, a indústria financeira, por meio de diferentes métodos, avalia o risco de inadimplência de seus devedores, conhecido como risco de crédito, empregando diferentes modelos matemáticos que, com base em dados históricos, determinam se um empréstimo será pago ou não. Esta tarefa, inicialmente, é um tradicional problema de classificação, considerando que devedores devem, a partir de suas características individuais, ser divididos em classes (mais particularmente, entre bons e maus pagadores).

Modelos clássicos de *credit scoring* abordam este problema usando abordagens estatísticas [Marqués et al. 2013], com destaque para regressão logística ou *logistic regression* (LR), que é inclusive o padrão na indústria [Dumitrescu 2021]. Em razão da menor precisão de técnicas estatísticas, muitos pesquisadores têm também proposto novos modelos partindo dos campos de *data mining* e *Machine Learning* (ML) [Siami et al. 2014]. Podemos dizer, em suma, sejam estatísticos ou de ML que modelos de classificação do risco de crédito estão inseridos no campo do aprendizado supervisionado.

O aprendizado supervisionado, em essência, significa que um subconjunto de treinamento contendo dados acompanhados de suas respectivas classes é utilizado para construir regras que permitirão a identificação das classes a que pertencem novos dados [Antonakis e Sfakianakis 2009]. Desta forma, durante a fase de aprendizado, um algoritmo de classificação ou classificador, neste caso, é treinado com dados históricos (em geral, dados demográficos e financeiros dos devedores, que chamamos de atributos ou *features*), para então prever o comportamento de novas amostras (classificando novo devedor como bom ou mau pagador).

Entretanto, uma modelagem precisa de um problema de classificação supervisionada vai além do treinamento e avaliação do classificador. Para que sejam obtidos bons resultados, especialmente no contexto de *credit scoring*, é necessário que sejam implementadas outras etapas (também características dos campos de *data mining* e ML), que auxiliam no preparo, treinamento e otimização do modelo. O emprego apropriado de etapas como o pré-processamento e balanceamento de dados, *feature selection* e otimização de hiper parâmetros (OHP), que serão brevemente discutidas na Seção 2, são inclusive mais importantes para o aperfeiçoamento dos resultados da classificação do que a escolha do classificador. Isto significa que a modelagem diligente e adequada de um classificador de maior simplicidade costuma trazer resultados melhores do que apenas um algoritmo de classificação excessivamente complexo, mas impróprio para o problema em mãos.

Consequentemente, este artigo tem como objetivo, partindo de bases históricas de dados relativos à empréstimos bancários, avaliar o impacto de diferentes etapas de modelagem sobre uma LR, o *benchmark* para *credit scoring* na indústria financeira, para então propor diretrizes para a sua modelagem eficiente neste contexto. Para isso, este trabalho está dividido em 5 seções. A Seção 2 apresenta o referencial teórico. A Seção 3 descreve brevemente os passos para a condução da presente pesquisa. A Seção 4 apresenta a aplicação do método e a discussão dos resultados. A Seção 5 traz as conclusões, e por fim, a Seção 6 traz os agradecimentos e as referências utilizadas.

2. Fundamentação

2.1 Regressão logística

A LR é um processo de modelagem estatística da probabilidade de uma variável discretizada saída (variável dependente ou classe) dada uma ou mais variáveis explanatórias de entrada (variáveis independentes ou atributos). Com aplicações em uma grande variedade de áreas [Honorio e Ortiz 2015], trata-se de uma ferramenta útil para problemas de classificação, onde se busca determinar em que classe se encontra uma nova amostra [Edgar e Manz 2017].

Como na regressão linear, a LR produz um modelo da relação entre múltiplas variáveis [Seufert 2014] em um conjunto de dados. Essencialmente, usa-se uma função logística para a modelagem de uma variável binária de saída [Tolles e Meurer 2016], para que então seja prevista a sua probabilidade de assumir um valor no intervalo de 0 a 1.

Assim, a probabilidade da variável dependente Y ocupar determinada classe, pode ser estimada diretamente a partir de um conjunto de n variáveis independentes $X_1, X_2, \dots, X_n$ e a LR pode ser usada tanto para a classificação quanto para a estimativa da probabilidade de classes. Por esta razão seria mais apropriado que se chamasse classificação logística [Seufert 2014], uma vez que, apesar de seu nome, trata-se de um modelo de classificação e não regressão propriamente dita [Subasi 2020]. O modelo é descrito pela Eq. 2.1:

$$P(Y = 1) = \frac{1}{1 + e^{-g(x)}} \quad (2.1)$$

Em que:

$$g(x) = B_0 + B_1 * X_1 + \dots + B_n * X_n \quad (2.2)$$

Sendo os coeficientes $B_0, B_1, \dots, B_n$ da Eq. 2.2 estimados a partir do conjunto de dados na busca de uma combinação que maximize a probabilidade de a amostra ter sido observada.

Analisando as Eqs. 2.2 e 2.3, é possível notar que quando $g(x)$ tende a $-\infty$, $P(Y=1)$ tende a 0 e quando $g(x)$ tende a $+\infty$, $P(Y=1)$ tende a 1 [Hosmer e Lemeshow 1989]. Desta forma, a curvade probabilidade apresenta um formato sigmoide (forma de S) (Figura 2.1).

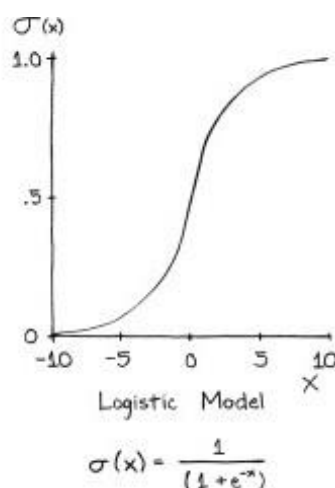


Figura 2.1 – Curva sigmoide.
 Fonte: Adaptado de [Seufert 2014]

Trata-se de um dos modelos estatísticos mais utilizados para algoritmos de classificação [Abdou et al. 2016]. Em ML, conforme [Gudivada et al. 2016], vetores ponderados dos atributos são mapeados na função sigmoide e interpretados como a probabilidade de uma amostra pertencer a determinada classe. O algoritmo, por sua vez, deve ajustar os coeficientes – o que pode ser feito de diferentes maneiras – para então classificar amostras de treinamento e possibilitar que o modelo seja aplicado em novas amostras.

Conforme [Seufert, 2014], é um método extremamente robusto e flexível para classificação preditiva dicotômica; ou seja, para a previsão de resultados ou estados binários. Em razão de sua relativa simplicidade em comparação a outros algoritmos, não fazendo suposições rigorosas [Takawira e Mwamba 2015], costuma ser a primeira opção para problemas de classificação [Gudivada et al. 2016]. Além de sua maior praticidade, a LR também é altamente explicativa [Bequé et al. 2017], tendo significativas vantagens em termos de interpretabilidade [Shen 2020], o que é requerido pelas agências financeiras regulamentadoras (tornando-a o *benchmark* na indústria de crédito [Dumitrescu 2021]).

2.2 Modelagem para aprendizado supervisionado

A modelagem para a classificação por meio de algoritmos de aprendizado supervisionado, por sua vez, vai muito além do classificador. Para resultados precisos e eficazes, é

necessário que o modelo seja preparado e otimizado em etapas bem definidas. Dentro do escopo deste estudo, serão exploradas as etapas de pré-processamento e balanceamento de dados, *feature selection* e OHP.

O pré-processamento de dados é uma etapa crucial na modelagem de algoritmos de classificação. Durante esta etapa é feita a preparação dos dados brutos, onde são aplicadas várias técnicas de limpeza e transformação de dados como, por exemplo, a remoção de valores ausentes e *outliers*, codificação de variáveis categóricas e normalização. O objetivo aqui é garantir que os dados estejam adequados e facilitem o aprendizado do modelo.

Outro desafio para o aprendizado do modelo pode surgir também com o desbalanceamento de classes. Especialmente em bases de dados de crédito, o número de amostras de uma classe usualmente supera o da outra (as bases costumam conter mais exemplos de bons pagadores do que maus pagadores). Conjuntos desbalanceados de dados, por sua vez, costumam causar *overfitting* nos modelos, cuja previsão se inclinará para a classe majoritária [Onel 2018]. Tal desequilíbrio entre classes costuma interferir negativamente nos resultados da classificação [Chen2021]. A forma mais comum de realizar o balanceamento das classes é feita, assim como o pré-processamento, no nível dos dados, por meio de novas amostragens que reduzem instâncias da classe majoritária (*undersampling*), ou aumentam as da classe minoritária (*oversampling*), etapa que é conhecida como *resampling*.

Outro desafio inerente aos dados de crédito é o excesso de atributos sobre devedores. Na era do *Big Data*, instituições de crédito detêm um grande volume de dados sobre os seus clientes, o que pode intensificar severamente a intensidade computacional do treinamento dos modelos [Zhou et al. 2021; Hu et al. 2018]. Em resposta, existem diversas metodologias para redução de dimensionalidade, que envolvem a filtragem ou remoção de variáveis de menor potencial preditivo [Ala'raj e Abbod 2016]. No campo de ML, a priorização de variáveis de maior impacto é conhecida como seleção de variáveis ou *feature selection*, processo importante para a modelagem do risco de crédito [Bou-Hamad 2017].

Metodologias de *feature selection* podem ser essencialmente divididas entre dois maiores grupos: métodos *filter* e *wrapper*. Essencialmente, *filters* usam testes estatísticos ou outras medidas de informação para avaliar a relevância individual de cada atributo (o que também é feito no nível dos dados) [Liu et al. 2018], sendo desconsiderada a interação entre o conjunto de variáveis escolhidas e o algoritmo de classificação [Zhang et al. 2013]. *Wrappers*, por outro lado, funcionam de maneira semelhante, com a exceção de usarem um algoritmo de classificação predefinido, avaliando conjuntos de variáveis de acordo com o classificador e frequentemente resultando em melhor performance [Wah et al. 2018].

Finalmente, de acordo com [Wu et al. 2019], algoritmos de aprendizado supervisionado comumente envolvem problemas de otimização. Estes algoritmos possuem hiper parâmetros, que controlam seu comportamento e desempenho. A etapa de OHP significa encontrar os melhores valores para esses parâmetros, a fim de maximizar o desempenho do modelo. Existem várias técnicas de otimização de hiper parâmetros, sendo que uma das mais simples e popular é a *grid search* [Dufour e Neves 2019]. Por meio deste algoritmo de OHP, o modelo é treinado exaustivamente com diversas combinações entre hiper parâmetros, até ser encontrada a melhor combinação, de acordo com uma métrica pré-definida [Wu et al. 2019].

3. Materiais e métodos

Para avaliar o efeito entre as etapas de modelagem discutidas sobre a LR, diversos modelos diferentes foram avaliados, contabilizando todas as etapas e possíveis interações entre elas. Combinando ou não LR, nosso classificador, com as etapas de pré-processamento de dados, *resampling*, *feature selection* e OHP, toda a modelagem foi feita em Python, por meio da biblioteca *scikit-learn*.

Cada um dos modelos foi avaliado em 3 diferentes *datasets* encontrados no repositório de ML da *University of California Irvine* (UCI). [Louzada et al. 2016] mencionam um estudo que indica que cerca de 45% dos artigos sobre análise do risco de crédito analisados levaram em conta os *datasets* *Australian*, *German* e *Japanese* da UCI. Estes *datasets* contém dados reais fornecidos

por instituições de crédito, listando a situação de contratos de empréstimos e informações de seus respectivos clientes. Desta forma, além de podermos avaliar resultados frente a incontáveis *benchmarks* na literatura, a avaliação com múltiplos *datasets* nos permite verificar melhor a robustez dos modelos. As principais características dos *datasets* utilizados, mais especificamente o número de amostras, atributos e percentual de inadimplentes, encontram-se abaixo, na Tabela 3.1:

	<u>Núm. Amostras</u>	<u>Núm. Atributos</u>	<u>% Inadimplentes</u>
<i>Australian</i>	690	14	44,49%
<i>German</i>	1000	64	30%
<i>Japanese</i>	670	46	55,22%

Tabela 3.1 – Descrição dos *datasets*.

A grande diferença entre o número de atributos se dá por causa da presença de variáveis categóricas nos *datasets German* e *Japanese*. Por isso, para prepará-las para o modelo, foi realizada a técnica de *one-hot encoding*, em que as variáveis categóricas são representadas de forma numérica e binária. Isto é alcançado criando colunas adicionais para cada categoria única da variável, recebendo o valor de 1 quando presente a categoria, e 0 caso contrário, o que aumenta significativamente o número de atributos nestes casos.

Além desta codificação de variáveis categóricas, a remoção de valores ausentes ou inadequados foi outra etapa de pré-processamento realizada preliminarmente, sabendo que, caso contrário, não é possível prosseguir para a classificação. Desta forma, modelos considerados como pré-processados contam, adicionalmente, com as etapas de remoção de *outliers* e de padronização (que converte as variáveis para uma escala com média zero e desvio padrão unitário). Nesta análise exploratória, é possível obter uma visão geral das bases, especialmente quanto ao balanceamento entre as classes. Pode-se notar o diferente grau de desbalanceamento entre os *datasets*, o que deve impactar a importância da etapa de *resampling* em cada um dos deles. Na etapa de *resampling*, foi utilizada uma combinação entre algoritmos de *oversampling* e *undersampling* (respectivamente, os algoritmos SMOTE e Tomek's links), igualando-se o número de amostras de cada classe.

Da mesma forma, na etapa de *feature selection*, os resultados podem variar significativamente dependendo do número de atributos escolhidos. Neste estudo, foi implementado tanto um método *filter* (*correlation based feature selection* - CFS) quanto um *wrapper* (*sequential backward selection* - SBS). Também se optou pela escolha dos 5 melhores atributos, o que não garante que qualquer outra combinação de escolha das variáveis tenha performance superior. Neste sentido, este estudo se beneficia por conduzir a avaliação dos modelos em múltiplos *datasets*. No entanto, o objetivo é ter uma visão geral do efeito que as etapas de modelagem têm sobre o desempenho da classificação por meio de LR.

Para garantir a reprodutibilidade dos experimentos e garantir a consistência de nossa análise comparativa, o parâmetro *random_state* da LR foi fixado igualmente para todos os modelos. Os demais parâmetros pertinentes, bem como a sua faixa de variação na etapa de OHP, em que foi implementada uma *grid search*, encontram-se na Tabela 3.2.

<u>Parâmetro</u>	<u>Amplitude</u>
<i>penalty</i>	['l1', 'l2', 'elasticnet']
<i>C</i>	[0.01, 100]
<i>class_weight</i>	[dict, 'balanced']
<i>max_iter</i>	[100, 1000]
<i>solver</i>	['lbfgs', 'newton-cg', 'newton-cholesky']

Tabela 3.2 – Parâmetros para LR.

4. Resultados e discussão

Na Tabela 4.1 encontram-se os resultados de cada um dos modelos para as métricas de acurácia, erros tipo I e II, e AUC. Cada um dos modelos foi treinado e avaliado em cada um dos 3 *datasets* utilizando a técnica de *10-fold cross-validation*, em que os dados são divididos em múltiplos subconjuntos, para que o modelo seja treinado e avaliado em cada um deles.

<i>Dataset</i>	Modelo	Acurácia	Erro Tipo I ↓	Erro Tipo II ↓	AUC
<i>Australian</i>	LR	0,804	0,124	0,289	0,793
	LR_WFS	0,855	0,203	0,070	0,864
	LR_OHP	0,874	0,149	0,094	0,878
	LR_WFS_OHP	0,855	0,203	0,070	0,864
	LR_PP	0,871	0,131	0,119	0,875
	LR_PP_WFS	0,849	0,203	0,084	0,857
	LR_PP_OHP	0,872	0,151	0,094	0,878
	LR_PP_WFS_OHP	0,857	0,203	0,066	0,865
	LR_FFS	0,858	0,160	0,117	0,861
	LR_FFS_OHP	0,865	0,168	0,090	0,871
	LR_RES	0,807	0,135	0,256	0,805
	LR_RES_WFS	0,860	0,198	0,080	0,861
	LR_RES_OHP	0,866	0,151	0,118	0,865
	LR_RES_WFS_OHP	0,860	0,198	0,080	0,861
	LR_PP_FFS	0,859	0,165	0,110	0,863
	LR_PP_FFS_OHP	0,862	0,167	0,100	0,866
	LR_PP_RES	0,889	0,136	0,090	0,887
	LR_PP_RES_WFS	0,889	0,147	0,079	0,887
	LR_PP_RES_OHP	0,893	0,115	0,102	0,891
	LR_PP_RES_WFS_OHP	0,889	0,147	0,079	0,887
	LR_RES_FFS	0,868	0,167	0,098	0,867
	LR_RES_FFS_OHP	0,871	0,167	0,093	0,870
	LR_PP_RES_FFS	0,885	0,140	0,092	0,884
	LR_PP_RES_FFS_OHP	0,885	0,140	0,092	0,884
<i>German</i>	LR	0,742	0,135	0,541	0,662
	LR_WFS	0,749	0,080	0,643	0,639
	LR_OHP	0,721	0,292	0,255	0,727
	LR_WFS_OHP	0,674	0,352	0,268	0,690
	LR_PP	0,748	0,139	0,510	0,675
	LR_PP_WFS	0,749	0,073	0,661	0,633
	LR_PP_OHP	0,720	0,288	0,269	0,721
	LR_PP_WFS_OHP	0,658	0,379	0,250	0,685
	LR_FFS	0,741	0,097	0,644	0,630
	LR_FFS_OHP	0,678	0,346	0,275	0,690
	LR_RES	0,771	0,255	0,201	0,772
	LR_RES_WFS	0,707	0,211	0,373	0,708
	LR_RES_OHP	0,773	0,267	0,187	0,773
	LR_RES_WFS_OHP	0,712	0,202	0,374	0,712
	LR_PP_FFS	0,743	0,094	0,644	0,631

Japanese	LR_PP_FFS_OHP	0,698	0,298	0,318	0,692
	LR_PP_RES	0,761	0,254	0,223	0,761
	LR_PP_RES_WFS	0,695	0,241	0,364	0,698
	LR_PP_RES_OHP	0,763	0,263	0,210	0,763
	LR_PP_RES_WFS_OHP	0,696	0,242	0,363	0,698
	LR_RES_FFS	0,722	0,300	0,253	0,724
	LR_RES_FFS_OHP	0,723	0,285	0,266	0,724
	LR_PP_RES_FFS	0,714	0,318	0,251	0,716
	LR_PP_RES_FFS_OHP	0,714	0,318	0,251	0,716
	LR	0,869	0,105	0,152	0,872
	LR_WFS	0,860	0,060	0,208	0,866
	LR_OHP	0,869	0,088	0,166	0,873
	LR_WFS_OHP	0,863	0,052	0,208	0,870
	LR_PP	0,873	0,108	0,142	0,875
	LR_PP_WFS	0,860	0,060	0,208	0,866
	LR_PP_OHP	0,869	0,078	0,175	0,873
	LR_PP_WFS_OHP	0,863	0,052	0,208	0,870
	LR_FFS	0,695	0,000	1,000	0,500
	LR_FFS_OHP	0,410	0,669	0,376	0,478
	LR_RES	0,868	0,093	0,170	0,869
	LR_RES_WFS	0,863	0,059	0,215	0,863
	LR_RES_OHP	0,878	0,081	0,167	0,876
	LR_RES_WFS_OHP	0,868	0,050	0,215	0,868
	LR_PP_FFS	0,860	0,060	0,208	0,866
	LR_PP_FFS_OHP	0,860	0,060	0,208	0,866
	LR_PP_RES	0,903	0,061	0,132	0,903
	LR_PP_RES_WFS	0,891	0,040	0,173	0,894
	LR_PP_RES_OHP	0,904	0,067	0,124	0,905
	LR_PP_RES_WFS_OHP	0,893	0,040	0,170	0,895
	LR_RES_FFS	0,860	0,061	0,218	0,860
	LR_RES_FFS_OHP	0,862	0,061	0,215	0,862
	LR_PP_RES_FFS	0,893	0,042	0,167	0,895
	LR_PP_RES_FFS_OHP	0,893	0,042	0,167	0,895

Tabela 4.1 – Resultados para os modelos em cada *dataset*.

*Abreviações significam: PP = Pré-processamento de dados; RES = *Resampling*; FFS = *Filter feature selection*; WFS = *Wrapper feature selection*; OHP = Otimização de hiper parâmetros.

** ↓ Significa que quanto menor, melhor.

É possível observar não só que alguns modelos de LR têm ótimo desempenho em alguns dos *datasets*, como o fato de que os resultados melhoram conforme o modelo é incrementado das etapas discutidas. A avaliação destes resultados, por sua vez, deve ser feita com cautela, sabendo que, por si só, a interpretação das métricas empregadas pode ser enganosa. Analisando, por exemplo, a performance do modelo LR_FFS no *dataset Japanese*, nenhum erro tipo I pode nos parecer vantajoso quando, na verdade, visto em conjunto com as demais métricas, isso significa que ele sofre de *overfitting*, tendo classificado todas as amostras como parte da classe majoritária (o que é confirmado pelo erro tipo II, que é igual a 1, o que indica que o modelo falhou na classificação de todas as amostras de inadimplentes). Isto pode ser enxergado em AUC, métrica que, para o valor de 0,5, indica a ausência de potencial de discriminação.

Desta forma, considerando um *ranking* médio das métricas de performance, os melhores modelos para os *datasets Australian*, *German* e *Japanese* foram, respectivamente: LR_PP_RES, LR_RES_OHP, LR_PP_RES_FFS_OHP. Dado o objetivo deste estudo, entretanto, analisar quais foram os modelos de melhor performance é ainda insuficiente. Desta forma, por

possível observar como cada uma das etapas, bem como as interações entre elas, afeta o *ranking* de cada modelo de LR entre os *datasets*.

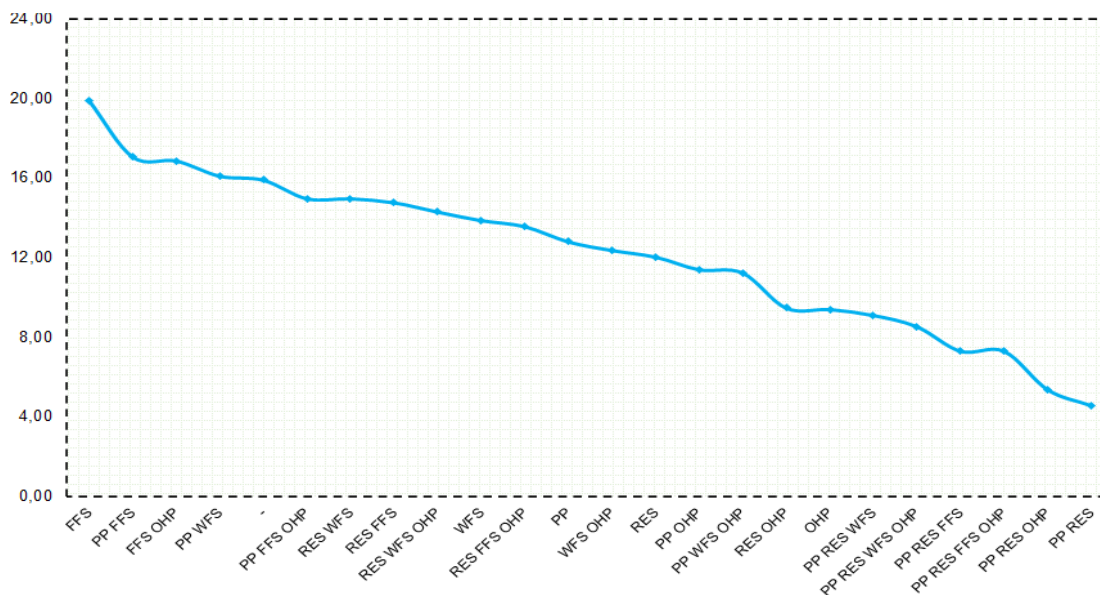


Figura 4.1 – *Ranking* médio dos modelos LR entre os *datasets*.

Analisando a Figura 4.1, nota-se que, na modelagem de LR, as etapas de pré-processamento de dados e resampling não só são as que mais causam variações nos modelos, como são os principais fatores para o aumento em sua performance, estando sempre presentes entre os melhores modelos, isto é, aqueles mais à direita, que detém um melhor *ranking* médio entre os *datasets* (quando pré-processado, há irrisório decréscimo em relação ao melhor modelo, no caso de LR_RES_OHP no *dataset German*). Embora inconsistentes, as etapas de FS também impactaram positivamente alguns dos modelos.

4.1 Desenvolvimento do *framework*

Para consubstanciar as observações possíveis acerca dos resultados dos modelos, bem como embasar o desenvolvimento de um *framework* para a modelagem de LR na classificação dorisco de crédito, foi, por fim, feita uma análise de importância das etapas. Neste caso, foi realizada uma regressão, por meio do algoritmo de ML *Random Forest* (RF) (uma agregação de 100 árvores de decisão), para prever a posição de um modelo no *ranking* com base nas etapas de modelagem que o compõe. Desta forma, RF é capaz de avaliar quais etapas de um modelo mais impactam sua posição em seu respectivo *ranking*. O resultado da análise de importância de etapas por meio de RF encontra-se abaixo, na Figura 4.2.

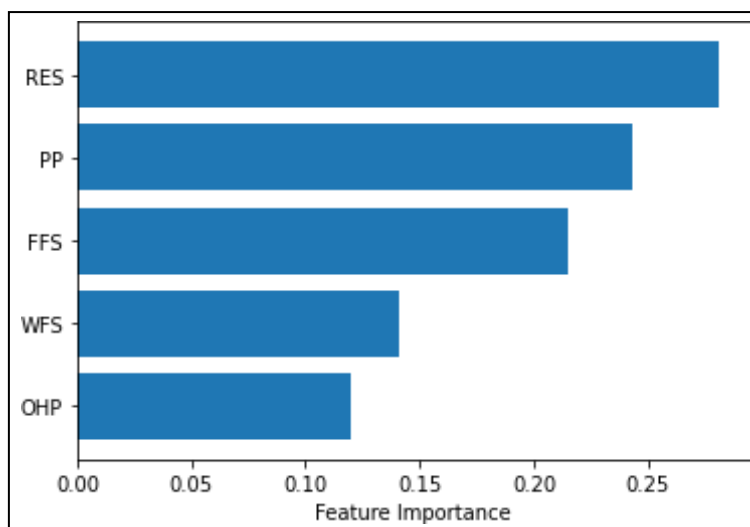


Figura 4.2 – Resultado da análise de importância das etapas de modelagem.

Como esperado, em conformidade com a Figura 4.1, a análise de importância das etapas também aponta para o pré-processamento e balanceamento dos dados como as etapas mais importantes de modelagem, considerando os 3 *datasets* utilizados. Além do efeito inconsistente das etapas de *feature selection*, a etapa de OHP, apesar de encontrar-se presente em alguns dos melhores resultados, não parece, por si só, colaborar significativamente para o aperfeiçoamento do resultado dos modelos.

Como discutido, poderíamos experimentar ainda diferentes tipos e tamanhos de subconjuntos na etapa de *feature selection*, bem como diferentes formas de conduzir as etapas de pré-processamento, *resampling* e OHP. Contudo, dado o nosso objetivo em propor diretrizes gerais para a modelagem eficiente de LR na classificação do risco de crédito, os resultados da análise dos experimentos já são o suficiente para fundamentar uma proposta de modelagem generalizada, simples e eficiente, ilustrada pelo *framework* da Figura 4.3.

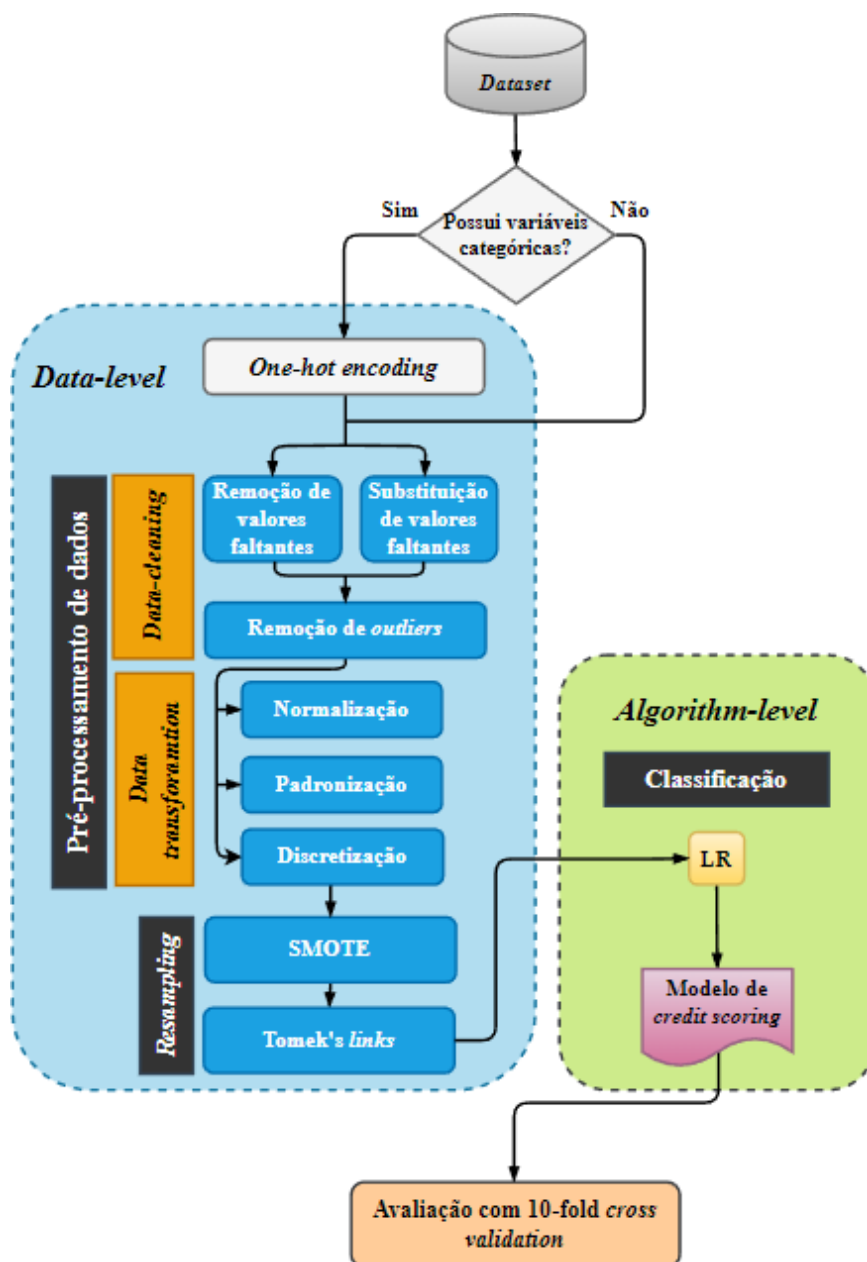


Figura 4.3 – *Framework* proposto para a modelagem do risco de crédito por meio de LR.

Em consonância com a análise dos resultados para os modelos de LR avaliados neste estudo (especialmente nas Figuras 4.1-4.2), o *framework* proposto incentiva a priorização das etapas de pré-processamento e balanceamento de dados, cujo impacto no aperfeiçoamento de performance foi mais pronunciado. Desta forma, apesar do modelador ter ainda a opção de incluir novas etapas, bem como optar por alternativas para o pré-processamento e *resampling*, acredita-se que o *framework* da Figura 4.3 possa servir como uma boa referência, trazendo bons resultados sem excessiva complexidade ou intensidade computacional, o que é especialmente importante no caso de grandes volumes de dados.

5. Conclusões

As diligências de instituições financeiras durante o processo de concessão de crédito são essenciais para o seu sucesso. A concessão de crédito é uma atividade fundamental para o desenvolvimento social e econômico do país, e é crucial que sejam tomados os esforços necessários para atender às expectativas legítimas da indústria e seus clientes, garantindo o efetivo adimplemento dos empréstimos tomados.

Modelos de aprendizado supervisionado oferecem possíveis soluções para o problema de *credit scoring*, abordando-o usualmente como um problema de classificação. Neste contexto, a LR é amplamente utilizada na modelagem de risco de crédito pela indústria financeira. Contudo, inovações na área de ML têm trazido novos algoritmos e técnicas mais avançadas para classificação, geralmente em troca de mais recursos e menor interpretabilidade dos modelos.

Mais importante que a escolha do classificador, contudo, são as demais etapas de modelagem que devem acompanhar modelos de *credit scoring*. Isto significa que, por si só, a escolha de algoritmos complexos de classificação não costuma resultar em melhorias tão significativas quanto a implementação apropriada de etapas como o pré-processamento de dados, *resampling*, *feature selection* e OHP. Assim, neste estudo, através da avaliação de múltiplos modelos de LR em 3 diferentes *datasets* de crédito populares na literatura, foram avaliados os impactos que cada uma destas etapas tem na performance de LR na classificação do risco de crédito para, por fim, desenvolver um *framework* contendo diretrizes para a modelagem do mais popular algoritmo de classificação neste contexto.

Analisando os resultados, é possível ver como LR, apesar de sua relativa simplicidade, é capaz de discriminar bem as classes em problemas de *credit scoring*, especialmente quando modelada apropriadamente, sendo acompanhada das devidas etapas. Não só pode ser observada a melhoria no desempenho dos modelos conforme incrementados destas etapas, como também o fato de que o pré-processamento e balanceamento de dados foram os maiores responsáveis pelos aumentos, resultando no *framework* proposto, que permanece simples, direto e não demanda grande intensidade computacional. Por fim, para futuros estudos, sugere-se a remodelagem de LR com novos parâmetros e atributos, diferentes metodologias em cada uma das etapas, e sua comparação com algoritmos avançados de ML.

6. Agradecimentos

Os autores agradecem a CNPq, CAPES e FAPEMIG pelo apoio à esta e outras pesquisas.

Referências

- Abdou, H. A. et al. (2016). Predicting creditworthiness in retail banking with limited scoring data. *Knowledge-Based Systems*, v. 103, p. 89-103.
- Ala'raj, M., Abbod, M. F. (2016). Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems*, v. 104, p. 89-105.
- Antonakis, A. C., e Sfakianakis, M. E. (2009). Assessing naive Bayes as a method for screening credit applicants. *Journal of applied Statistics*, v. 36, n. 5, p. 537-545.
- Bequé, A., e Lessmann, S. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems with Applications*, v. 86, p. 42-53.

- Bou-Hamad, I. (2017). Bayesian credit ratings: A random forest alternative approach. *Communications in Statistics-Theory and Methods*, v. 46, n. 15, p. 7289-7300.
- Chen, T. (2021). Credit Default Risk Prediction of Lenders with Resampling Methods. In: 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). IEEE. p. 123-127.
- Dufour, J., e Neves, J. (2019). Finite-sample inference and nonstandard asymptotics with Monte Carlo tests and R. In: *Handbook of statistics*. Elsevier. p. 3-31.
- Dumitrescu, E. et al. (2021). Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds.
- Edgar, T., Manz, D. (2017). *Research methods for cyber security*. Syngress.
- Gudivada, V. N. et al. (2016). Cognitive analytics: Going beyond big data analytics and machine learning. In: *Handbook of statistics*. Elsevier. p. 169-205.
- Honorio, J., e Ortiz, L. E. (2015). Learning the structure and parameters of large-population graphical games from behavioral data. *J. Mach. Learn. Res.*, v. 16, n. 1, p. 1157-1210.
- Hosmer, D. W., Jovanovic, B., e Lemeshow, S. (1989). Best subsets logistic regression. *Biometrics*, p. 1265-1270.
- Hu, L. et al. (2018). Feature selection considering two types of feature relevancy and feature interdependency. *Expert Systems with Applications*, v. 93, p. 423-434.
- Liu, X. et al. (2018). A hybrid genetic algorithm with wrapper-embedded approaches for feature selection. *IEEE Access*, v. 6, p. 22863-22874.
- Louzada, F., Ara, A., e Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, v. 21, n. 2, p. 117-134.
- Marqués, A. I., García, V., e Sánchez, J. S. (2013). On the suitability of resampling techniques for the class imbalance problem in credit scoring. *Journal of the Operational Research Society*, v. 64, n. 7, p. 1060-1070.
- Onel, M. et al. (2018). Big data approach to batch process monitoring: Simultaneous fault detection and diagnosis using nonlinear support vector machine-based feature selection. *Computers & chemical engineering*, v. 115, p. 46-63.
- Seufert, E. B. (2014). The freemium business model. *Freemium Economics*, p. 1-27.
- Shen, F., Wang, R., e Shen, Y. (2020). A cost-sensitive logistic regression credit scoring model based on multi-objective optimization approach. *Technological and Economic Development of Economy*, v. 26, n. 2, p. 405-429.
- Siami, M., Gholamian, M. R., e Basiri, J. (2014). An application of locally linear model tree algorithm with combination of feature selection in credit scoring. *International Journal of Systems Science*, v. 45, n. 10, p. 2213-2222.
- Subasi, A. (2020). *Practical Machine Learning for Data Analysis Using Python*. Academic Press.
- Takawira, O., e Mwamba, J. W. M. (2022). Sovereign Credit Ratings Analysis Using the Logistic Regression Model. *Risks*, v. 10, n. 4, p. 70.
- Tolles, J., e Meurer, W. J. (2016). Logistic regression: relating patient characteristics to outcomes. *Jama*, v. 316, n. 5, p. 533-534.
- Wah, Y. B. et al. (2018). Feature Selection Methods: Case of Filter and Wrapper Approaches for Maximizing Classification Accuracy. *Pertanika Journal of Science & Technology*, v. 26, n. 1.

Wu, J. et al. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, v. 17, n. 1, p. 26-40.

Zhang, D. et al. (2013). Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications*, v. 37, n. 12, p. 7838-7843.

Zhou, Y. et al. (2021). Feature selection in credit risk modeling: an international evidence. *Economic Research-Ekonomska Istraživanja*, v. 34, n. 1, p. 3064-3091.