Relatório de Aprendizado de Máquina - Aula 3

Lucas Ribeiro da Silva - 2022055564
Universidade Federal de Minas Gerais
Belo Horizonte - Minas Gerais - Brasil

lucasrsilvak@ufmg.br

1 Introdução

Neste relatório, utilizamos a técnica de Regressão Logística para modelar a relação entre duas ou mais variáveis. Essa abordagem nos permite avaliar quando a Regressão Logística pode ser útil, sua precisão e até mesmo prever valores desconhecidos, ou classificar dados a partir das variáveis conhecidas.

2 Dados

2.1 Disposição

Pequena amostragem dos dados após tratamento.

Index	Inadimplente	Student	Balance	Income
0	0	0	729.53	44361.63
1	0	1	817.18	12106.13
2	0	0	1073.55	31767.14
3	0	0	529.25	35704.49
4	0	0	785.66	38463.50

Tabela 1: Amostra de Dados com Informações de Inadimplência, Estudante, Saldo e Renda

2.2 Dados Gerais

Métrica	Valor
Número de inadimplentes	333
Número de estudantes	2944
Saldo médio	835.37
Renda média	33516.98

Tabela 2: Estatísticas do Conjunto de Dados

3 Análise dos Coeficientes e Interpretação

3.1 Balance x Inadimplente

- **Acurácia:** 0.9733

		Predito	
		Adimplente	**Inadimplente**
Real	**Adimplente**	2896	10
	Inadimplente	70	24

Tabela 3: Matriz de Confusão

Classe	**Precisão**	**Revocação**	**F1-Score**
Adimplente	0.98	1.00	0.99
Inadimplente	0.71	0.26	0.38
Acurácia		0.97	
Macro avg	0.84	0.63	0.68
Weighted avg	0.97	0.97	0.97

Tabela 4: Relatório de Classificação

Coeficientes do Modelo

- **Coeficientes (β_1):** 2.66775044
- **Intercepto:** (β_0) -6.0528

Análise

A chance de inadimplência aumenta linearmente com coeficiente β_1 com o balanço, entretanto, a análise feita sem maiores alterações na estrutura dos dados originais, não nos permite inferir conclusões significativas sobre a classe de interesse (os inadimplentes), pois a chance de um inadimplente ser corretamente classificado como inadimplente é baixa (apenas 26%). Nesse caso, pode se concluir que o algoritmo está *overfitting* a classe dos adimplentes e *underfitting* a classe dos inadimplentes.

3.2 Student x Inadimplente

- **Acurácia:** 0.9687

		Predito	
		Adimplente	**Inadimplente**
Real	**Adimplente**	2906	0
	Inadimplente	94	0

Tabela 5: Matriz de Confusão

Classe	**Precisão**	**Revocação**	**F1-Score**
Adimplente	0.97	1.00	0.98
Inadimplente	0.00	0.00	0.00
Acurácia		0.97	
Macro avg	0.48	0.50	0.49
Weighted avg	0.94	0.97	0.95

Tabela 6: Relatório de Classificação

Coeficientes do Modelo

- **Coeficientes (β_1):** 0.53001335
- **Intercepto (β_0):** -3.5289

Análise

A chance de inadimplência aumenta linearmente com coeficiente β_1 com a possibilidade student, entretanto, assim como no caso anterior, a análise feita sem maiores

alterações na estrutura dos dados originais, não nos permite inferir conclusões significativas sobre a classe de interesse (os inadimplentes), pois a chance de um inadimplente ser corretamente classificado como inadimplente é zero (0%). Nesse caso, pode se concluir que o algoritmo está *overfitting* a classe dos adimplentes e *underfitting* a classe dos inadimplentes. Como o $\beta_1 x + \beta_0 < 0$, a chance de um estudante ser inadimplente ainda deve ser considerada baixa.

3.3 Income x Inadimplente

- **Acurácia:** 0.9687

		Predito	
		Adimplente	**Inadimplente**
Real	**Adimplente**	2906	0
	Inadimplente	94	0

Tabela 7: Matriz de Confusão

Classe	**Precisão**	**Revocação**	**F1-Score**
Adimplente	0.97	1.00	0.98
Inadimplente	0.00	0.00	0.00
Acurácia		0.97	
Macro avg	0.48	0.50	0.49
Weighted avg	0.94	0.97	0.95

Tabela 8: Relatório de Classificação

Coeficientes do Modelo

- **Coeficientes (β_1):** -0.15751613
- **Intercepto (β_0):** -3.3531

Análise

A chance de inadimplência diminui linearmente com coeficiente β_1 com o income, entretanto, assim como nos casos anteriores, a análise feita sem maiores alterações na estrutura dos dados originais, não nos permite inferir conclusões significativas

sobre a classe de interesse (os inadimplentes), pois a chance de um inadimplente ser corretamente classificado como inadimplente é zero (0%). Nesse caso, pode se concluir que o algoritmo está *overfitting* a classe dos adimplentes e *underfitting* a classe dos inadimplentes. Nesse caso, pode se concluir que quanto maior a renda, menor a chance de ser inadimplente.

3.4 Todos x Inadimplente

Resultados com `balance`, `student` e `income`

- **Acurácia:** 0.9737

		Predito	
		Adimplente	**Inadimplente**
Real	**Adimplente**	2896	10
	Inadimplente	69	25

Tabela 9: Matriz de Confusão

Classe	**Precisão**	**Revocação**	**F1-Score**
Adimplente	0.98	1.00	0.99
Inadimplente	0.71	0.27	0.39
Acurácia		0.97	
Macro avg	0.85	0.63	0.69
Weighted avg	0.97	0.97	0.97

Tabela 10: Relatório de Classificação

Coeficientes do Modelo

- **Coeficientes ($\beta_1, \beta_2, \beta_3$):** 2.7678, 0.0774, -0.4549
- **Intercepto (β_0):** -6.0106

Análise

Podemos avaliar pela junção dos 3 dados numa mesma análise que o *balance* é mais importante na detecção de inadimplência que *student* e *income*, e que isso reflete num

coeficiente maior. No caso de *student*, o coeficiente se aproxima de 0. Podemos perceber que a análise rasa dos dados nos permite concluir correlações equivocadas. Mesmo assim, essa análise feita não nos permite inferir conclusões significativas sobre a classe de interesse (os inadimplentes), pois a chance de um inadimplente ser corretamente classificado como inadimplente é baixa (apenas 27%). Nesse caso, pode se concluir novamente que o algoritmo está *overfitting* a classe dos adimplentes e *underfitting* a classe dos inadimplentes.

4 Ajuste do Limiar de Decisão

Para contornar os sucessivos erros de *underfitting* em inadimplentes, alteraremos o Limiar de Decisão e analisaremos as conclusões.

- **Acurácia:** 0.899

		Predito	
		Adimplente	**Inadimplente**
Real	**Adimplente**	2616	290
	Inadimplente	13	81

Tabela 11: Matriz de Confusão

Classe	**Precisão**	**Revocação**	**F1-Score**
Adimplente	1.00	0.90	0.95
Inadimplente	0.22	0.86	0.35
Acurácia		0.90	
Macro avg	0.61	0.88	0.65
Weighted avg	0.97	0.90	0.93

Tabela 12: Relatório de Classificação

Coeficientes do Modelo

- **Coeficientes ($\beta_1, \beta_2, \beta_3$):** 2.7678, 0.0774, -0.4549
- **Intercepto (β_0):** -6.0106

Análise

Podemos avaliar que somente pela mudança do limiar de decisão já obtivemos um resultado muito superior para inferir conclusões significativas sobre a classe de interesse (os inadimplentes), pois a chance de um inadimplente ser corretamente classificado como inadimplente é alta (86%), ainda que apenas 22% dos classificados como inadimplentes sejam inadimplentes. Nesse caso, apesar do modelo dar *overfit* na classe dos adimplentes e *underfit* na classe dos inadimplentes, o resultado final é mais aceitável.

5 Conclusão

Ao utilizar a Regressão Logística, observamos como podemos modelar relações entre variáveis e identificar padrões. As análises feitas nos permitem avaliar o peso de cada variável sobre o resultado final e concluir que **balance** é a mais importante para maior precisão. Também é possível concluir que é necessário conhecer seu problema e que a mudança do limiar de decisão pode ser de suma importância para obter uma solução adequada.