Universidade Federal do Ceará
Centro de Ciências
Departamento de Estatística e Matemática Aplicada

Leonardo Medeiros Pereira
Vitor Salomão Aguiar de Araújo Guimarães

Análise de risco de crédito com o uso de modelos de regressão logística, redes neurais e algoritmos genéticos

FORTALEZA
2022

Resumo

A análise de risco de crédito é uma das formas de se modelar, avaliar perfis e consequentemente reduzir o risco de inadimplência, mantendo o equilíbrio financeiro da credora. Dentre as principais instituições que realizam a análise e distribuição de crédito, destacam-se bancos e seguradoras, mas se estendem desde governos a supermercados. Essencialmente, a análise é efetuada através de um estudo do histórico financeiro de um consumidor ou empresa, assim elaborando uma nota de risco (risk scoring/credit scoring) e classificando o solicitante em bom ou mau pagador.

A mensuração do risco auxilia na tomada de decisão de liberação ou não de crédito ao requisitante, sendo assim, o procedimento é de fundamental importância para definir a melhor estratégia para a instituição credora em relação à quantidade de crédito disponível e valor da taxa de juros repassada ao solicitante.

Neste trabalho temos por objetivo apresentar e comparar modelos de crédito, utilizando modelos estocásticos e métodos computacionais específicos, tais como: Redes Neurais (Neural Networks); Regressão Logística (Logistic Regression), sendo este um modelo estatístico que permite a predição/estimação da probabilidade do solicitante de fato pagar ou não o empréstimo; Algoritmos Genéticos (Genetic Algorithms), que essencialmente são algoritmos de otimização e busca, que constituem uma classe particular de algoritmos evolutivos, que consideram mecanismos de evolução natural e recombinação genética(crossover).

Para ilustrar as técnicas, consideramos uma base de dados financeiros e comparamos os resultados obtidos pelas três metodologias em questão, tentando enfatizar vantagens e limitações de cada uma delas.

Palavras-chave: Regressão Logística, Redes Neurais, Algoritmos Genéticos.

Sumário

1 INTRODUÇÃO **4**
1.1 OBJETIVO 4
1.1.1 ATIVIDADES 4
1.2 MOTIVAÇÃO DO TRABALHO 4

2 METODOLOGIA **5**

3 APLICAÇÃO E RESULTADOS **7**
3.1 Regressão Logística 7
3.2 Redes Neurais 8
3.3 Algoritmos Genéticos 8

4 CONSIDERAÇÕES FINAIS **9**

5 REFERÊNCIAS BIBLIOGRÁFICAS **10**

1 INTRODUÇÃO

Quando uma empresa ou pessoa solicita um empréstimo financeiro junto a uma entidade monetária, como bancos ou empresas de crédito, essas instituições a partir dos dados históricos do solicitante realizam uma análise prévia, classificando o cliente em bom ou mau pagador, essa classificação é chamada de pontuação de crédito (credit scoring), baseando-se em métodos estatísticos e computacionais.

Os credores, usam o credit score para avaliar o risco potencial representado pelo empréstimo de dinheiro aos consumidores e para atenuar perdas devido a dívidas não pagas. Os credores usam pontuações de crédito para determinar a que taxa de juros e quais limites de crédito, também usam pontuações de crédito para determinar quais clientes provavelmente gerarão mais receita.

Para a aplicação do método pode ser utilizado o modelo de regressão logística, algoritmos genéticos, redes neurais e outros, esse trabalho irá focar nesses três.

1.1 OBJETIVO

O objetivo desse trabalho é comparar o funcionamento do método scoring utilizando a regressão logística em confronto com o método de redes neurais e o de algoritmos genéticos.

1.1.1 ATIVIDADES

Foi feita uma pesquisa sobre o método scoring em vários sites e a procura de um banco de dados para colocar os métodos em prática. O banco de dados escolhido foi de um banco alemão, além de um estudo dos modelos.

1.2 MOTIVAÇÃO DO TRABALHO

A motivação desse trabalho é mostrar como a regressão logística é utilizada na prática, também fazer com que o discente procure por outros métodos usados no mundo e que não são abordados em sala de aula.

2 METODOLOGIA

Os modelos utilizados foram:

- Regressão Logística: A técnica mais utilizada no mercado para o desenvolvimento de modelos de credit scoring. Apresenta vantagem em relação à Análise Discriminante, por não pressupor dados de entrada com distribuição normal, embora seja desejável que as variáveis tenham essa distribuição.

- Redes Neurais: Inspirado na estrutura de um neurônio que adquire conhecimento a partir de experiências anteriores.
  Um modelo de rede neural artificial processa certas características e produz respostas similarmente ao cérebro humano. Com as seguintes suposições:
  1. Processamento das informações ocorre dentro dos chamados neurônios;
  2. Os estímulos são transmitidos pelos neurônios por meio de conexões;
  3. Cada conexão tem associada a si um peso, que, numa rede neural padrão, multiplica-se ao estímulo recebido;
  4. Cada neurônio contribui para a função de ativação para determinar o estímulo de saída.

- Algoritmo Genético(AG): baseado na Teoria da Evolução de Darwin, os indivíduos são codificados em genótipos e são representados por estruturas de dados chamadas de cromossomos. São comumente usados para gerar soluções de alta qualidade para problemas de otimização e pesquisa, contando com operadores inspirados biologicamente, como mutação , cruzamento e seleção.

- Variáveis: Idade, Sexo, Estado Civil, Emprego, Habitação, Contas de poupança, Conta corrente, Valor do crédito, Prazo, Finalidade do empréstimo, Bens Disponíveis, etc.

- Para a análise de dados fizemos uso dos softwares Python e R.

Os dados utilizados nessa análise foram disponibilzados por um banco alemão, contendo 1000 clientes, A divisão de credibilidade entre os clientes é 700 bons pagadores e 300 maus pagadores. Para uma padronização do experimento utilizamos 20% dos dados para treinamento e 80% para teste em todos os modelos.

Após a aplicacão foi utiliada a metodologia comparativa entre os modelos, avaliando a acurácia e visualizando a matriz de confusão de seus respectivos algoritmos.

Neste caso, os dados seguem o formato da Tabela 1.

Tabela 1: Exemplo Tabela.

Exemplos	Exemplo1	Exemplo2	Exemplo3
Var1	var11	var12	var13
Var2	var21	var22	var23
Var2	var21	var22	var23

Tabela 2: Exemplo 2 Tabela.

	Estatística Computacional		
Exemplos	Exemplo1	Exemplo2	Exemplo3
Var	Exemplo1	Exemplo2	Exemplo3
Var	ExemploX1	ExemploX2	ExemploX3

- Mariz de Confusão: É um layout de tabela específico que permite a visualização do desempenho de um algoritmo, normalmente de aprendizado supervisionado (em aprendizado não supervisionado é geralmente chamado de matriz de correspondência). Cada linha da matriz representa as instâncias de uma classe real enquanto cada coluna representa as instâncias de uma classe prevista, ou vice-versa, ("real"e "prevista"), e conjuntos idênticos de classes em ambas as dimensões.

Tabela 3: Exemplo Matriz de Confusão.

		Valor Predito	
Real	Sim	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

3 APLICAÇÃO E RESULTADOS

A seguir serão apresentados os modelos, suas aplicações e resultados.

3.1 Regressão Logística

Regressão Logística: nesse trabalho ela é usada para estimar a probabilidade associada a ocorrência de fatores que classificam o cliente como bom ou mau pagador. Temos fatores como: profissão, pagamento do mesmo anterior, atraso nos pagamentos, entre outros. Lembrando que os bancos podem adotar as variáveis que preferirem para decidir se o cliente é um bom ou mau pagador.

A regressão logística ao invés de ajustar uma reta como a linear, ela ajusta um gráfico na forma de um "S". Ela trabalha com a variável destino de forma binária, no caso do código o 0 é o mau pagador e 1 o bom pagador, além disso esse método só deve ser aplicado com amostras muito grandes, o que é o caso desse trabalho.

A regressão Logística tem como vantagem a sua fácil interpretação, eficiência e necessita de poucos recursos computacionais. Sua desvantagem é não conseguir resolver problemas não lineares, o que é um problema pois os sistemas atuais são não lineares.

A regressão logística segue a seguinte fórmula $h_\Theta(X) = \frac{1}{1+e^{-\Theta x}}$. Θ é o parâmetro usado para treinar o modelo e X são os dados de entrada.A saída dessa fórmula é o valor de previsão, esse valor tem que está mais próximo de 1 para ser uma amostra positiva, caso contrário é mais provável que a instância seja um amostra negativa.

Os dados para treinamento foram distribuídos de maneira aleatória, ou seja, com uma proporção de bons e maus pagadores desconhecida. Após o modelo ter sido implementado foi realizado um cálculo de probabilidade dos valores preditos, aqueles que apresentavam probabilidade acima de 0,5 foram classificados como bons pagadores e por consequência os abaixo como mau pagadores. O modelo, com os parâmetros definidos, apresentou uma acurácia de 76%, em comparação aos valores reais.

Tabela 4 - Matriz de confusão relacionada ao modelo de Regressão Logística

Regressão Logística				
		Valor Predito		
		Bom	Mau	Acerto (%)
Valor Real	Bom	503	137	78,6%
	Mau	55	105	65,6%
	Total	558	242	

3.2 Redes Neurais

Redes Neurais: Através das classes do sklearn foram escolhidos 10 classificadores, como "Random Forest", "Ada Boost", "Gradient Boost", "Decision Tree"entre outros . Como cada modelo é distinto foi necessário realizar ajustes nos parâmetros, porém mantendo a porcentagem de 20% dos dados para teste. Dentre todos os classificadores, aquele que se mostrou mais eficente foi o Ridge Regression Cross Validation, com uma acurácia de 75,62%, em comparação aos valores reais.

Tabela 5 - Matriz de confusão relacionada ao modelo de Redes Neurais.

Rede Neural (Ridge Regression Cross Validation)				
		Valor Predito		
		Bom	Mau	Acerto (%)
Valor Real	Bom	510	48	91,3%
	Mau	147	95	39,2%
	Total	657	143	

3.3 Algoritmos Genéticos

Algoritmo Genético: Um dos exemplos de aplicações do AG incluem a otimização de árvores de decisão para melhor desempenho, resolvendo quebra- cabeças sudoku , otimização de hiperparâmetros; classificação, como por exemplo se folhas ou cogumelos são comestíveis. Para o estudo foi usadada principalmente a abordagem de classificação.

Cada variável recebeu um peso aleatório, inicialmente variando de -0.1 a 0.1, de forma que o modelo selecione o vetor de pesos que apresente o melhor critério de classificação do cliente. Após os valores dos pesos para cada variável terem sido definidos, o modelo apresentou uma acurácia de 71%, em comparação com os valores reais.

Tabela 6 - Matriz de confusão relacionada ao modelo de Algoritmo Genético

Algoritmo Genético				
		Valor Predito		
		Bom	Mau	Acerto (%)
Valor Real	Bom	492	68	87,8%
	Mau	164	76	31,6%
	Total	656	144	

4 CONSIDERAÇÕES FINAIS

Os três modelos apresentaram resultados relativamente similares e satisfatórios para o conjunto de dados considerado. O modelo de regressão logística obteve resultados levemente superiores aos demais, explicando o motivo dele ser o mais comum no mercado para a criação de modelos de credit scoring.

O objetivo deste estudo não foi uma abordagem complexa das técnicas, e sim salientar que a tendência é que com o avanço dos algoritmos de classificação as empresas de crédito se tornem ainda mais criteriosas ao realizarem o empréstimo, encontrando formas mais eficientes na tomada de decisão, utilizando estratégias de data driven.

5 REFERÊNCIAS BIBLIOGRÁFICAS

ARTIFICIAL neural network. [S. l.], 13 nov. 2022. Disponível em: https://en.wikipedia.org/wiki/Artificial_ neural_ network. Acesso em: 18 nov. 2022.

CREDITABILITY - German Credit Data. [S. l.], 12 mar. 2021. Disponível em: https://www.kaggle.com/datasets/mpwolke/cusersmarildownloadsgermancsv. Acesso em: 18 nov. 2022.

GOUVÊA, Maria Aparecida; GONÇALVES, Eric Bacconi; NASSIF MANTOVANI, Daielly Melina. APLICAÇÃO DE REGRESSÃO LOGÍSTICA E ALGORITMOS GENÉTICOS NA ANÁLISE DE RISCO DE CRÉDITO. Revista Universo Contábil, [S.l.], v. 8, n. 2, p. 84-102, abr. 2012. ISSN 1809-3337. Disponível em: ¡encurtador.com.br/dhlyP¿. Acesso em: 8 nov. 2022.

HORN, D.M, (2016), Credit Scoring Using Genetic Programming. Disponível em: ¡encurtador.com.br/jAFPQ¿.Acesso em: 7 nov. 2022.

LaTex. Disponível em: https://pt.wikipedia.org/wiki/LaTeX. Acesso em: 14 de outubro de 2021.

PEDREGOSA, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

PYTHON SOFTWARE FOUNDATION. Python Language Site: Documentation, 2022. Página de documentação. Disponível em: <https://www.python.org/doc/>. Acesso em: 04 de nov. de 2022.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

SCIKIT-LEARN: Machine Learning in Python. [S. l.], 2013. Disponível em: https://scikit-learn.org/stable/modules/classes.html. Acesso em: 18 nov. 2022.