Predição de falência em empresas

Lucas Bryan Treuke

Escola de Matemática Aplicada, FGV EMAp - Rio de Janeiro, RJ
Junho 2023

1. Resumo

Atualmente, a sobrevivência e o sucesso das empresas estão intrinsecamente ligados à sua capacidade de gerar lucro e evitar a falência. Prever a falência de uma empresa pode ser crucial para tomar medidas preventivas e garantir a sustentabilidade dos negócios. Neste artigo, exploramos a possibilidade de usar indicadores financeiros para prever a falência de empresas. Nosso estudo se concentra em empresas localizadas em Taiwan, com dados coletados entre 1999 e 2009. O uso desses indicadores pode fornecer insights valiosos sobre a saúde financeira das empresas e ajudar na tomada de decisões estratégicas.

2. Introdução

2.1. Relevância

Atualmente, o dinheiro é o principal motivador para pesquisa, inovação e resolução de problemas. Empresas competem diariamente por uma fatia do mercado e buscam gerar lucro, crescer e se estabelecer. A falência de um negócio é uma das piores situações para um empreendedor, pois significa ter uma empresa que gera mais gastos do que lucro.

É crucial se preocupar com os indicadores que apontam risco de falência. Embora o lucro esteja diretamente relacionado à sobrevivência de uma empresa, existem outros indicadores que podem ser avaliados com mais detalhes. Alguns negócios conseguem se recuperar mesmo registrando prejuízo. Será que essas empresas não tinham dívidas acumuladas? Ou será que fazer empréstimos pode ser a chave para revitalizar a empresa? Quais são os indicadores mais relevantes e alarmantes para a falência de empresas? Podemos usar esses indicadores para prever a falência de empresas? Essas são algumas das perguntas que exploraremos neste artigo, que buscará responder a essas questões.

Para o artigo em questão, nos restringiremos a um escopo reduzido quanto à escala global, e analisaremos empresas localizadas em Taiwan, com dados que variam entre 1999 e 2009.

Taiwan, apesar dos atritos com a China, possui indicadores impressionantes e que podem acarretar em uma análise que possa informar sobre outros países. Seu Produto Interno Bruto (PIB) é de US$ 589 bilhões, e seu PIB per capita é o dobro da média mundial, alcançando aproximadamente US$ 25.000,00. Taiwan é dependente da importação de recursos naturais, sendo o setor primário o menor em sua economia. No entanto, os setores secundário e terciário são muito desenvolvidos, especialmente nas indústrias de tecnologia da informação e comunicação. Vale destacar a fabricação de semicondutores e outros dispositivos eletrônicos, já que Taiwan é responsável por mais de 80% da produção mundial desses componentes.

2.2. Dados

Levando em consideração esse cenário favorável para o setor empresarial, bem como a disponibilidade de dados sobre empresas taiwanesas, utilizaremos o conjunto de dados sobre falência do Kaggle, disponível em https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction. Esse conjunto de dados foi coletado a partir do jornal econômico de Taiwan, abrangendo o período de 1999 a 2009. A definição de falência das empresas foi baseada nas regulamentações comerciais da Bolsa de Valores de Taiwan.

O conjunto de dados possui 6.819 instâncias e 96 atributos. Para nossa análise, vamos trabalhar com um conjunto reduzido de colunas, que são as seguintes:

- 'Net Income to Total Assets'
- 'ROA(A) before interest and % after tax'
- 'ROA(B) before interest and depreciation after tax'
- 'ROA(C) before interest and depreciation before interest'
- 'Debt ratio %'
- 'Net worth/Assets'

- 'Retained Earnings to Total Assets'
- 'Net profit before tax/Paid-in capital'
- 'Per Share Net profit before tax (Yuan ¥)'
- 'Current Liability to Assets'
- 'Working Capital to Total Assets'
- 'Bankrupt?'

Essa seleção foi feito observando o gráfico da correlação de cada coluna com a target (conforme a figura abaixo), e do significado de cada variável.

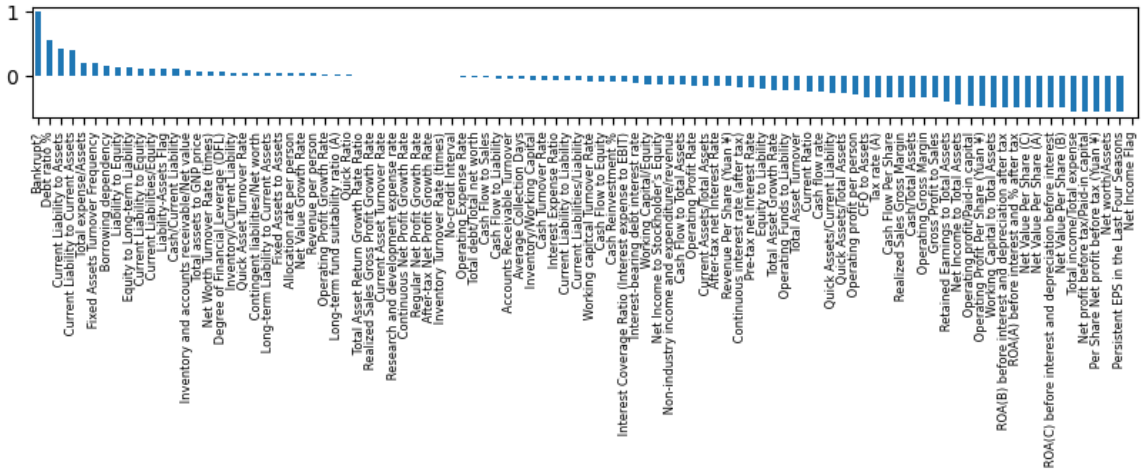

Figura 1: Gráfico de barras indicando correlação das variáveis com a target

1. 'Net Income to Total Assets': Essa coluna representa a proporção do lucro líquido em relação ao total de ativos da empresa. É uma medida da rentabilidade da empresa em relação aos ativos que possui.
2. 'ROA(A) before interest and % after tax': Essa coluna representa o Retorno sobre Ativos (ROA) antes dos juros e após os impostos. O ROA é uma medida de eficiência que indica a capacidade da empresa em gerar lucro a partir de seus ativos.
3. 'ROA(B) before interest and depreciation after tax': Essa coluna representa o ROA antes dos juros e da depreciação, mas após os impostos. É outra medida de eficiência que leva em consideração os efeitos da depreciação nos resultados da empresa.
4. 'ROA(C) before interest and depreciation before interest': Essa coluna representa o ROA antes dos juros e da depreciação, sem considerar os impostos. É mais uma medida de eficiência que exclui os efeitos dos impostos sobre o desempenho da empresa.
5. 'Debt ratio %': Essa coluna representa a proporção da dívida em relação ao total de ativos da empresa. É uma medida da alavancagem financeira da empresa e indica sua capacidade de pagamento das obrigações.
6. 'Net worth/Assets': Essa coluna representa a proporção do patrimônio líquido em relação aos ativos totais da empresa. O patrimônio líquido é a diferença entre os ativos e as obrigações da empresa e essa medida indica a solidez financeira da empresa.
7. 'Retained Earnings to Total Assets': Essa coluna representa a proporção dos lucros retidos em relação aos ativos totais da empresa. Os lucros retidos são os lucros acumulados que não foram distribuídos aos acionistas como dividendos.
8. 'Net profit before tax/Paid-in capital': Essa coluna representa o lucro líquido antes dos impostos em relação ao capital integralizado. É uma medida da rentabilidade em relação aos recursos financeiros investidos pelos acionistas.
9. 'Per Share Net profit before tax (Yuan ¥)': Essa coluna representa o lucro líquido antes dos impostos por ação em moeda Yuan. É uma medida de rentabilidade por ação e permite comparar o desempenho da empresa em termos de lucro por ação.
10. 'Current Liability to Assets': Essa coluna representa a proporção das obrigações de curto prazo em relação aos ativos totais da empresa. As obrigações de curto prazo incluem dívidas e outras obrigações que devem ser pagas em um curto período de tempo.

11. 'Working Capital to Total Assets': Essa coluna representa a proporção do capital de giro em relação aos ativos totais da empresa. O capital de giro é a diferença entre os ativos circulantes (como caixa, contas a receber) e as obrigações circulantes (como contas a pagar) e essa medida indica a capacidade da empresa em financiar suas operações diárias.
12. 'Bankrupt?': Essa coluna indica se a empresa faliu ou não. É a variável de resposta ou variável dependente que será usada para prever a falência com base nas outras colunas explicativas.

3. Metodologia

Inicialmente, selecionamos um conjunto reduzido de colunas com base na análise de correlação e significado das variáveis. Essas colunas representam indicadores financeiros importantes, como a rentabilidade, a alavancagem financeira e a saúde financeira da empresa. Em seguida, realizamos o pré-processamento dos dados, normalizando-os e removendo colunas altamente correlacionadas para evitar multicolinearidade.

3.1. Pré-processamento de dados

Antes de construir o modelo de previsão, é necessário pré-processar os dados. Isso pode incluir tratamento de valores ausentes, normalização de dados, codificação de variáveis categóricas e divisão do conjunto de dados em treinamento e teste.

Quanto aos dados, minima limpeza foi necessária pois eles já estavam prontos pra uso.

O primeiro passo foi normalizar os dados, e em seguida analisar para ver se eles estão com dados em bom estado para utilizar em um modelo de predição.
Após analise das colunas selecionadas em relação ao target de falência, é importante verificar a correlação entre essas colunas. Com o gráfico de heatmap, é possível identificar as correlações existentes entre as variáveis.

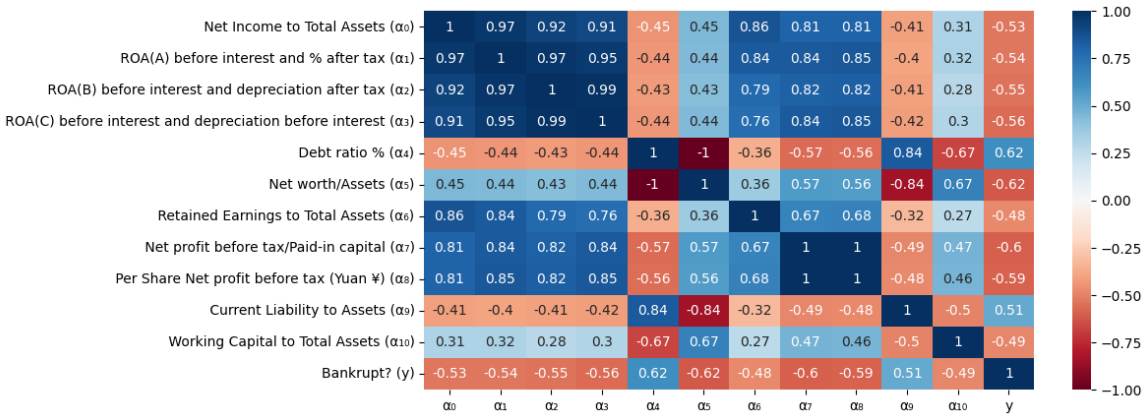

Figura 2: Heatmap de correlação das variáveis

Note que algumas colunas apresentam uma correlação muito alta, chegando a valores de -1 e 1, o que indica que elas possuem a mesma informação. Essas correlações perfeitas indicam uma redundância nos dados e podem causar problemas na modelagem.

É recomendado remover as colunas altamente correlacionadas, pois elas não adicionam informações adicionais e podem introduzir multicolinearidade, que pode prejudicar a interpretação do modelo e sua capacidade de generalização. A multicolinearidade ocorre quando duas ou mais variáveis independentes estão altamente correlacionadas entre si, o que pode levar a coeficientes de regressão instáveis ou de difícil interpretação.

Além disso, as correlações altas, que variam de 0.9 a 0.98, entre as variáveis selecionadas indicam relações lineares ou quadráticas entre elas. Por exemplo, podemos observar uma relação quadrática entre as variáveis α_0 (Net Income to Total Assets) e α_1 (ROA(A) before interest and % after tax), o que pode ser verificado por meio de um scatterplot entre essas colunas.

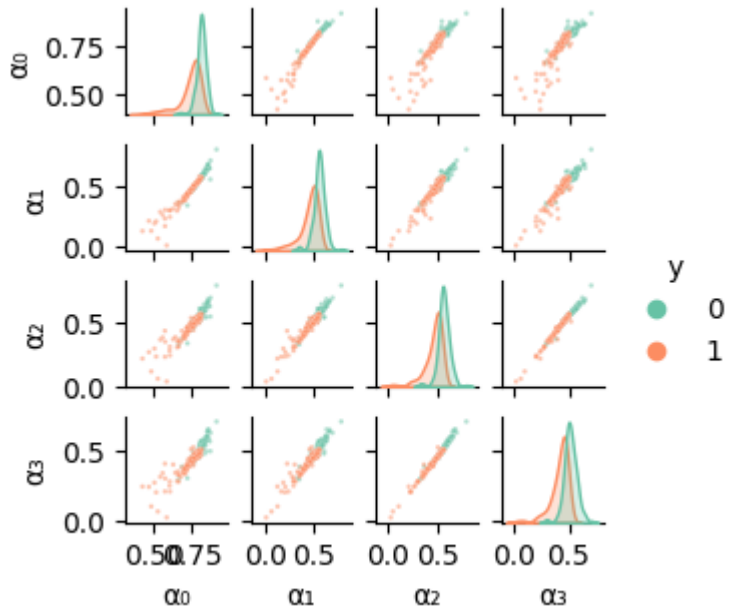

Figura 3: Pairplot entre α_0, α_1, α_2 e α_3

Essas informações sobre as relações entre as variáveis são relevantes para a modelagem, pois podem influenciar a escolha do algoritmo de classificação e as transformações necessárias nos dados. Por exemplo, ao identificar uma relação linear entre ,
podemos considerar o uso de um modelo linear ou um algoritmo que capture essa linearidade, como regressão logística.

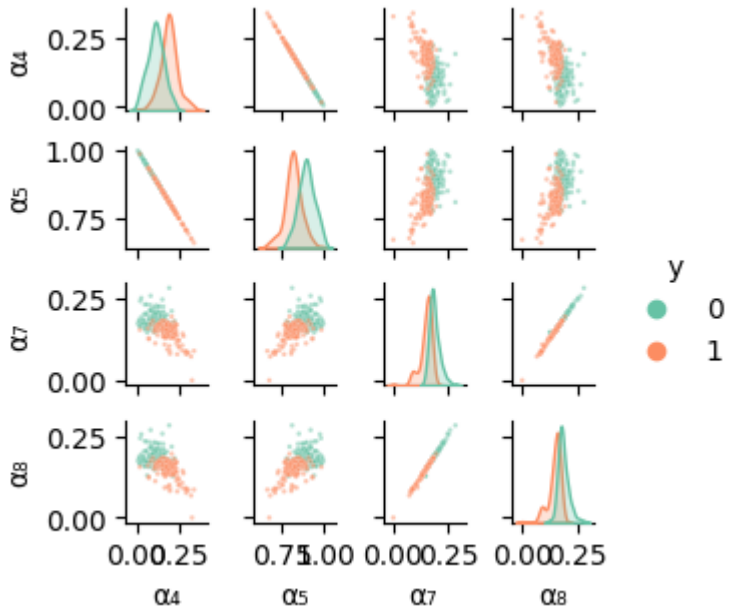

Figura 4: Pairplot entre α_4, α_5, α_7 e α_8

Já a relação inversamente proporcional entre α_4 (Debt ratio %) e α_5 (Net worth/Assets) pode indicar a presença de uma possível variável latente subjacente, como a capacidade de pagamento da empresa ou sua saúde financeira. Além disso a relação linear entre α_7 e α_8 é clara.

Dado essas interações, tais variáveis serão guardadas para futuramente avaliar se essas sutilezas nas suas observações divergindo da linha de base dada pelo α_0 (ou pela α_5 ou α_8, dependendo de qual exemplar de grupo mantivemos); afeta o resultado final por meio da interação.

Assim, vamos remover α_1, α_2, α_3, α_4, e α_7 para as primeira análises.
A relação entre as variáveis passa a ser a seguinte:

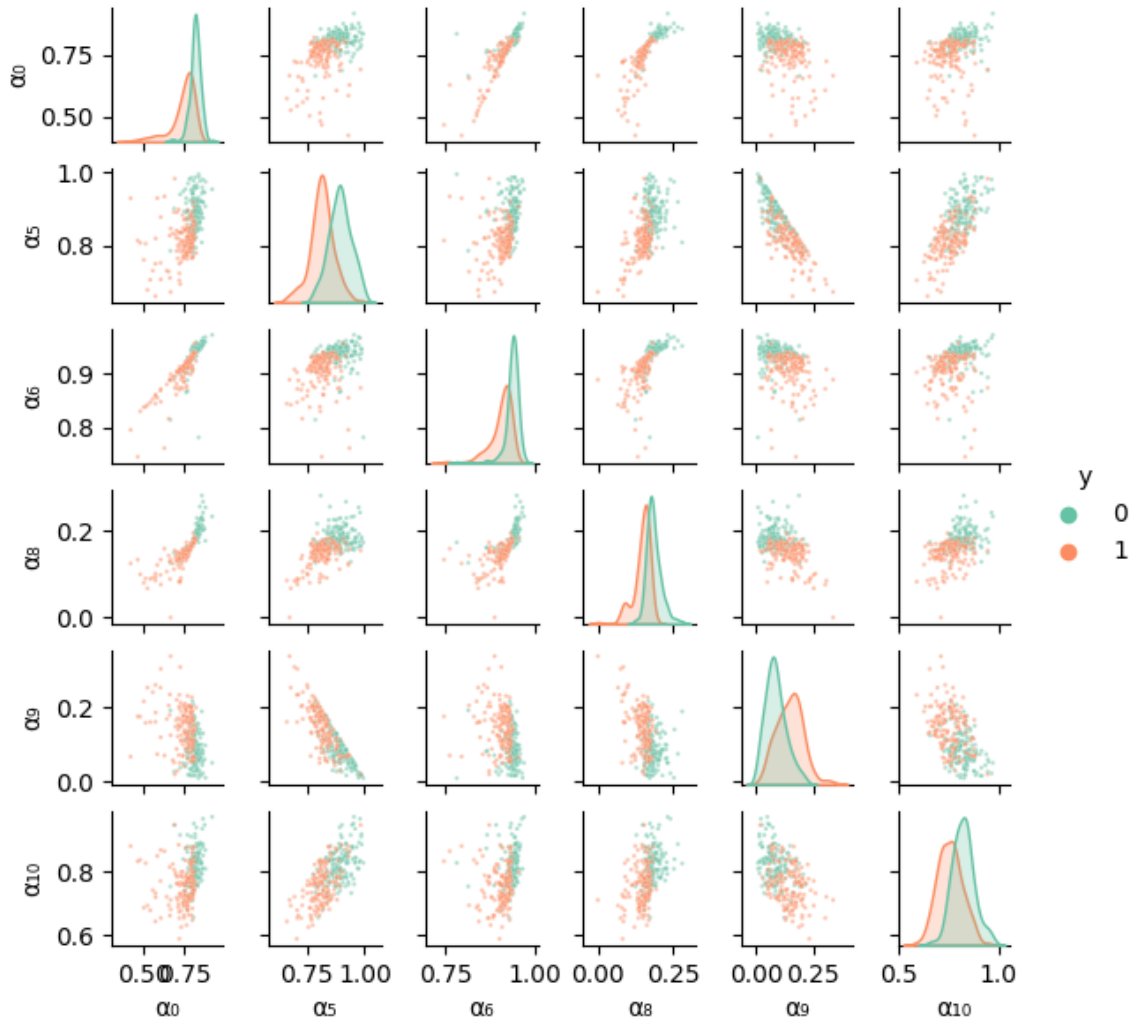

Figura 5: Pairplot entre os preditores restantes

Para construir um modelo de previsão de falência, podemos utilizar a notação "y ~ x_0 + ... + x_n" para indicar a relação entre o target (y) e as variáveis selecionadas (x_i).
Dessa forma, podemos usar programas estatísticos ou bibliotecas de modelagem, como o R ou o Python com bibliotecas como o StatsModels (utilizada no trabalho), para realizar a regressão logística ou outros modelos adequados.

Assim, a partir de formulas podemos ajustar os coeficientes do modelo para estimar o efeito de cada variável nas chances de falência da empresa.

Para avaliação dos modelos treinados, foram utilizadas as métricas de Acurácia, AUC e AIC.

O dataset conta com 220 entradas de empresas que faliram, as demais
Durante o ajuste dos modelos, foi utilizado um grupo de 150 entradas correspondentes a empresas que faliram, e 150 das que não vieram a falência, para contornar o desbalanceamento do dataset.

3.2. Resultados

Para facilitar as notações depois da remoção de preditoras, considere:
α_0: x_0, α_5: x_1, α_6: x_2, α_8: x_3, α_9: x_4, α_{10}: x_5

Para iniciar a análise, vamos visualizar a influência de cada coluna selecionada em relação ao target (falência da empresa) por meio de boxplots.

Não se atente a valores das variáveis, interprete apenas a posição das médias em verde ou laranja uma em relação a outra para a mesma variável

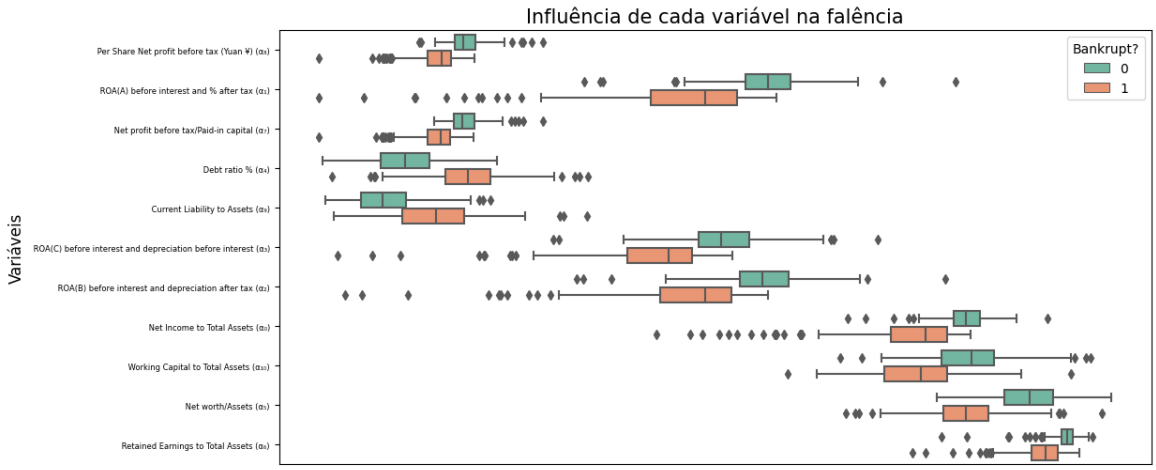

Figura 6: Boxplot avaliando cada preditor

A análise dos boxplots das variáveis em relação ao target mostrou que a média dos indicadores se desloca de forma distinta para empresas que faliram em comparação com as que não faliram, evidenciando a relação entre os indicadores financeiros e o risco de falência.

Utilizando a regressão logística como modelo de previsão, ajustamos diversos modelos com diferentes combinações de variáveis. Avaliamos a performance dos modelos utilizando métricas como acurácia, área sob a curva (AUC) e critério de informação de Akaike (AIC).

Os resultados obtidos mostraram que os modelos construídos apresentaram um desempenho satisfatório na previsão de falência das empresas. O modelo que utilizou todas as variáveis selecionadas obteve uma acurácia de 84.2% e uma AUC de 0.922, indicando uma boa capacidade de classificação. Além disso, a inclusão de interações entre as variáveis não trouxe melhorias significativas nos resultados, sugerindo que as variáveis selecionadas não apresentam uma interação que enriquece o modelo, e pela curva ROC vista elas parecem significar a mesma coisa.

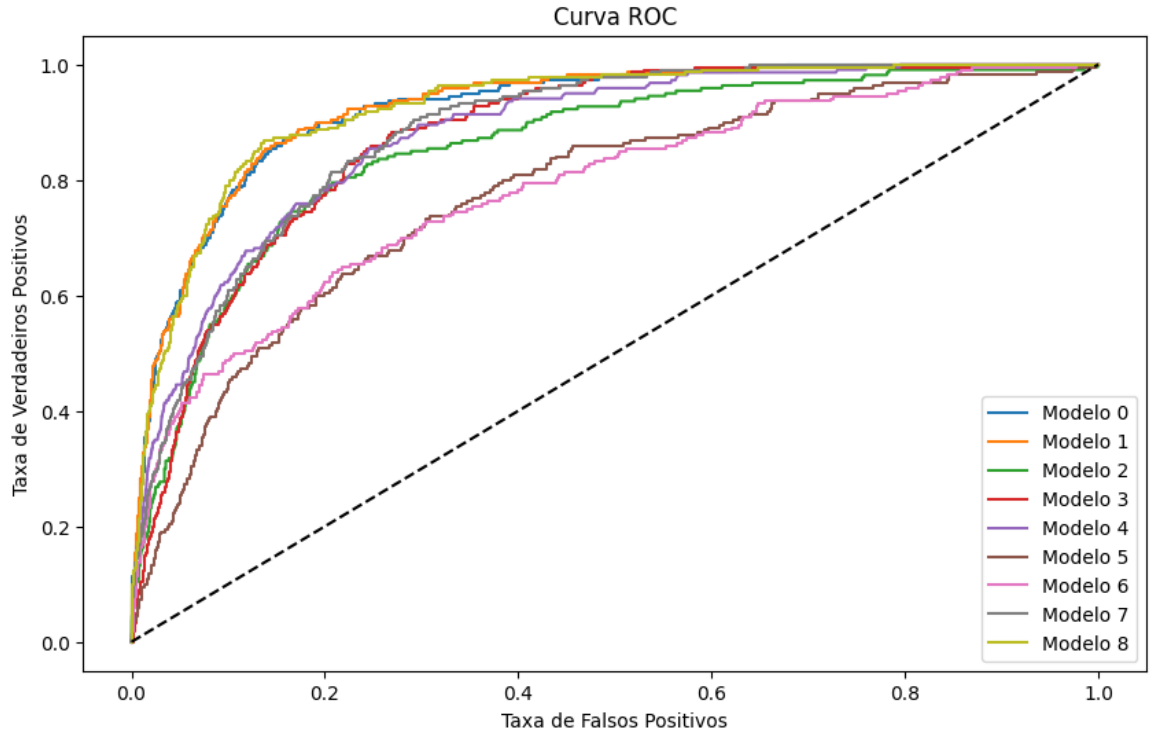

Figura 7: Curva ROC para os modelos

Modelo	Descrição	Acurácia	AUC	AIC
0	Todas as variáveis	0.842	0.922	173.520
1	Todas as variáveis + interações	0.838	0.925	174.555
2	x_1	0.751	0.855	222.272
3	x_2	0.833	0.875	235.987
4	x_3	0.808	0.880	204.661
5	x_4	0.709	0.775	251.169
6	x_5	0.653	0.783	222.697
7	x_0	0.838	0.883	218.203
8	x_0 + x_1	0.838	0.924	180.851

Em suma, este estudo demonstrou a viabilidade de prever a falência de empresas com base em indicadores financeiros selecionados. Essas previsões podem ser úteis para auxiliar empreendedores, investidores e instituições financeiras a tomar decisões informadas sobre investimentos, concessão de crédito e gerenciamento de riscos. No entanto, é importante ressaltar que a previsão de falência é um desafio complexo e que outros fatores além dos indicadores financeiros podem influenciar o resultado. Portanto, é recomendado o uso desses modelos como uma ferramenta complementar, em conjunto com outras análises e informações relevantes, para uma tomada de decisão mais robusta.

Por fim podemos avaliar de uma maneira diferente o melhor modelo

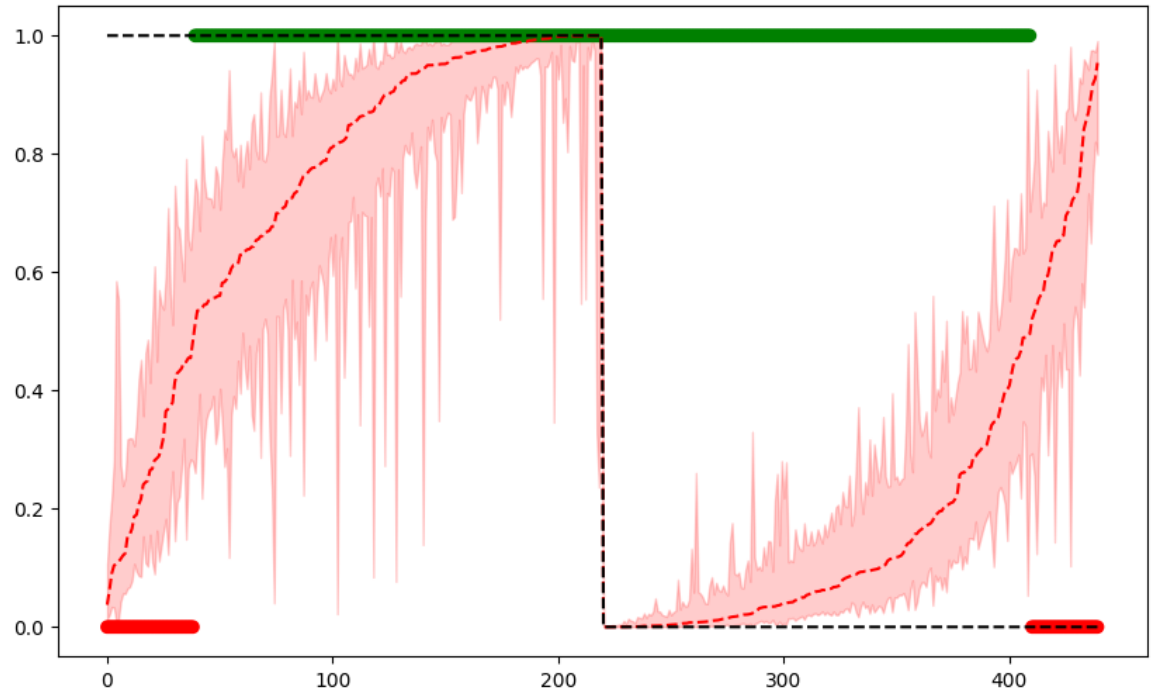

Figura 8: Intervalo de confiança, quanto mais ao centro melhores as predições

A linha pontilhada preta indica o verdadeiro valor de y, enquanto a vermelha é a predição.
O gráfico foi ordenado de modo que as predições mais certeiras estão mais ao centro, e contam com uma marcação verde no topo.
As predições erroneas estão com uma marcação vermelha embaixo, nas extremidades.
Podemos perceber que as previsões estão boas mas o modelo não está muito certo para os dados de empresas que faliram (os intervalos de confiança são maiores e menos suaves).

3.3. Limitações:

3.3.1. Tempo:

O baixo tempo disponível que tive para dedicar ao presente artigo inviabilizou uma abordagem mais minuciosa.

3.3.2. Extrapolação:

As previsões do modelo são baseadas nos dados de treinamento disponíveis. Se os dados não forem representativos o suficiente ou se houver vieses nos dados, o modelo pode não generalizar bem para novos dados.
Devido aos poucos dados e ao nicho deles, o modelo provavelmente não pode ser utilizado para dados fora do dataset usado.

3.3.3. Limitações dos dados disponíveis:

Mesmo com o grande volume de dados, eles estavam muito desbalanceados, deixando uma quantidade baixa de dado aproveitável.

3.3.4. Estabilidade temporal:

As condições e relações entre as variáveis podem mudar ao longo do tempo, o que pode afetar a precisão das previsões do modelo. Os dados existentes também são desatualizados, e podem não refletir mais a realidade.

3.4. Possíveis extensões:

3.4.1. Melhoria do modelo:

Pode-se explorar diferentes técnicas de modelagem, como algoritmos mais avançados, levando em considerações mais fatores e interações.

3.4.2. Exploração de variáveis adicionais:

Investigar a inclusão de novas variáveis relevantes ou a transformação de variáveis existentes pode fornecer mais insights e melhorar as previsões.

3.4.3. Validação externa:

Realizar validação externa do modelo, usando conjuntos de dados independentes, pode fornecer uma avaliação mais robusta de seu desempenho e capacidade de generalização.

4. Referências

- Kaggle. (s.d.). Company Bankruptcy Prediction. Disponível em: https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction
- Brasil Escola. "Taiwan." Disponível em: https://brasilescola.uol.com.br/geografia/taiwan.htm.
- Liu, T.-Y. et al. "Taiwan as a key strategic hub for the global economy: A critical overview." World Development Perspectives, vol. 1, pp. 10-14, 2016. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0377221716000412.
- Faz Capital. "Por que Taiwan é tão estratégico para a economia global?" Disponível em: https://fazcapital.com.br/por-que-taiwan-e-tao-estrategico-para-a-economia-global/#:~:text=Taiwan%20%C3%A9%20um%20pa%C3%ADs%20estrat%C3%A9gico,de%20produtos%20de%20alta%20tecnologia.
- Smolski, V. "Regressão logística." Livro Avançado de Estatística. Disponível em: https://smolski.github.io/livroavancado/reglog.html#o-modelo.