Machine learning e análise de sentimento: Projetando o risco de insolvência bancária

Resumo

A principal motivação deste artigo é utilizar técnicas de *machine learning* para construir uma nova métrica de classificação de risco de insolvência para os bancos negociados na B3. Em seguida, será utilizado um conjunto de modelos de predição para projetar a classificação de risco destas instituições. Convencionalmente, a literatura analisa o risco de insolvência bancária a partir dos dados contábeis e variáveis macroeconômicas. Além dessas variáveis, esse artigo irá construir uma série de sentimento do gestor da instituição bancária, via relatórios trimestrais (ITR), e essa será utilizada para melhorar a acurácia das previsões do risco bancário. Os resultados indicam que a classificação de risco bancário, via algoritmo *k-means*, foi capaz de classificar 17% da amostra no grupo de maior risco (1), enquanto 83% da amostra ficou no grupo de menor risco de falência (0). Utilizando a métrica do Z-score verificamos que 65% da amostra faz parte do grupo de baixo risco e 35% da amostra no grupo de risco elevado. Desse modo, o algoritmo *k-means* é mais rigoroso em classificar um banco na categoria de maior risco. Na sequência utilizamos os dados já descritos para projetar o risco de insolvência bancária. Os resultados desta etapa mostraram que o modelo de árvore de decisão apresentou o melhor desempenho para a amostra de teste. Além disso, constatou-se que a inclusão da variável de sentimento bancário foi capaz de melhorar o desempenho dos modelos de previsão, principalmente, quando o sentimento bancário é construído a partir de um dicionário variante no tempo.

Palavras-Chave: Insolvência bancária, *Machine learning*, Cluster, Sentimento bancário.

1 Introdução

A fragilidade financeira de uma instituição bancária pode gerar efeitos devastadores em uma economia e no sistema financeiro de um país. Atualmente, vários tipos de modelos de previsão são empregados com o objetivo de prever o risco bancário, fornecendo informações aos reguladores para que estes sejam capazes de tomar alguma decisão de forma antecipada, evitando ou minimizando os efeitos negativos sobre o resto do sistema financeiro.

Dentro deste contexto, este trabalho pretende construir uma nova métrica de classificação de risco de fragilidade financeira dos principais bancos de capital aberto negociados na B3. Vão ser criados instrumentos que ajudam a dirimir o risco bancário aumentando a sustentabilidade do sistema financeiro que podem ajudar a previsibilidade e com isso antecipar eventuais situações de vulnerabilidade. Essa nova medida de risco bancário será comparada com o Z-score que é uma métrica tradicional e amplamente utilizada na literatura. Após a construção da variável de risco, via técnicas de clusterização, esta será projetada por um conjunto de modelos de predição com o intuito de saber qual deles oferece a melhor acurácia. Além disso, o trabalho também busca verificar se o sentimento do gestor da instituição bancária é uma variável relevante na predição da nova métrica construída.

A importância deste trabalho está ligada ao fato que crises financeiras sempre têm consequências catastróficas, em especial, a crise *Subprime*, iniciada em 2008 com o estouro da bolha do mercado imobiliário. Essa crise teve múltiplas consequências na economia global, mostrando, entre outras questões, que os problemas financeiros das instituições bancárias vão além dos problemas sociais e econômicos e são capazes de afetar agentes em todo o mundo.

Diante da forte repercussão financeira adversa gerada, comportamentos de fragilidade bancária passaram a ganhar cada vez mais destaque na literatura, uma vez que tanto os investidores como os donos de depósitos tendem a perder a confiança nessas instituições inadimplentes, o que pode contaminar os demais bancos presentes no mercado e, no longo prazo, resultar em uma crise bancária. Essa última denota em consequências ainda mais severas que vão desde a paralisação da oferta de crédito para empresas e famílias até a fuga de capitais daquele país (Barbosa, 2017).

De acordo com Lepetit e Strobel (2013) a insolvência de uma instituição bancária ocorre quando as perdas incorridas por essa instituição não podem ser cobertas pelos seus recursos próprios. A partir desse fato a literatura tem desenvolvido medidas que buscam mensurar o risco de insolvência, entre essas, destacam-se, sistema CAMELS e o Z-score, em que o último indica a distância em que o banco se encontra de um comportamento de insolvência. Para mais detalhes, ver: Suss e Treitel (2019), Vieira, Silva e Florêncio (2020), Viswanathan, Srinivasan e Hariharan (2020).

Além de mensurar o risco bancário, a literatura também passou a se preocupar em prever e antecipar a falência destas instituições. Numerosas técnicas foram desenvolvidas ao longo dos anos na tentativa de fornecer aos analistas e tomadores de decisão métodos eficazes de previsão do risco de falência bancário com base em vários índices financeiros e modelos matemáticos, com esses modelos incluindo regressões lineares e logísticas, *splines* de regressão adaptativa multivariada, análise de sobrevivência, programação linear e quadrática e programação de critérios múltiplos, conforme visto em Karels e Prakash (1987), Ezzamel, Mar-Molinero e Beech (1987) e Ravi et al. (2008). Segundo Huang e Yen (2019) muitas dessas técnicas são tipicamente baseadas nas suposições de separabilidade linear e normalidade multivariada e, de fato,

na independência das variáveis explicativas. No entanto, essas condições são frequentemente violadas em situações da vida real.

Com o aumento expressivo no número de dados e informações alguns autores começaram a empregar técnicas de *machine learning* (ML) para a predição do risco bancário[1]. De acordo com Huang e Yen (2019) as técnicas de ML têm a capacidade de extrair informações significativas de dados não estruturados, ao mesmo tempo que lidam com a não linearidade de maneira eficaz. No entanto, a aplicação de técnicas avançadas de ML à previsão financeira ainda é uma área relativamente nova para os pesquisadores explorarem.

Paule-Vianez, Gutiérrez-Fernández e Coca-Pérez (2019) afirmam que as variáveis utilizadas na maioria dos estudos para prever o risco de falência das intituições financeiras têm sido os índices financeiros, especialmente os índices classificados em capital, ativos, gestão, resultados, liquidez e sensibilidade (Sistema CAMELS) e algumas variáveis econômicas[2]. Outros autores buscaram incluir novas variáveis na explicação do risco de insolvência bancário. Uma delas é o sentimento que o gestor da instituição bancária transmite por meio dos relatórios trimestrais e comunicações ao mercado. A ideia é captar, por meio do sentimento textual, uma relação direta entre o risco de insolvência e o tom pessimista, buscando aumentar a capacidade preditiva dos modelos.

Essa discussão pode ser encontrada em Gupta, Simaan e Zaki (2016), em que esses ressaltam que o sentimento textual é capaz de predizer a falência dos bancos de capital aberto, considerando ainda que o tom textual otimista na comunicação dos gestores tem um maior poder preditivo para essas instituições, identificando que os bancos insolventes apresentam sentimentos mais positivos do que os seus pares não falidos.

A literatura de previsão estabeleceu o valor da análise textual, bem como uma metodologia geral para converter texto em *scores* quantitativos que avaliam principalmente as polaridades dos textos. De acordo com Gentzkow, Kelly e Taddy (2019), as informações codificadas no texto são um complemento rico para os tipos de dados mais estruturados tradicionalmente usados na pesquisa empírica. De fato, nos últimos anos, ocorreu um uso intenso de dados textuais em diferentes áreas de pesquisa.

Assim, este trabalho busca contribuir com a literatura descrita acima em alguns pontos, sendo eles: primeiro é a construção de uma nova métrica de risco de insolvência bancária a partir do agrupamento de *cluster* por meio da técnica *k-means* que consiste em um método de ML não supervisionado e que permite classificar uma base de dados através de agrupamentos que minimizam o erro quadrado. Isso nos permite modelar o risco de insolvência em vez de falência total. Essa abordagem tem uma série de vantagens principais, sendo a mais importante o alinhamento às necessidades práticas dos órgãos reguladores que procuram intervir muito antes do fracasso, ou seja, da falência total do banco.

Em segundo lugar, o trabalho contribui com a literatura sobre o risco de falência bancária, indo além das técnicas de modelagem convencionais, utilizando métodos da literatura de ML ao lado de abordagens mais tradicionais. De acordo com Suss e Treitel (2019) abordagens convencionais, como modelos de regressão logística, são incapazes de levar em conta interações complexas e não linearidades, tendendo a ter um desempenho pior do que suas contrapartes de aprendizado de máquina mais flexíveis. Neste artigo,

[1] Ver os trabalhos de Sun e Li (2012), Erdogan (2013), Kim, Baik e Cho (2016), Chou, Hsieh e Qiu (2017), Xia et al. (2017), Hsu (2019), Suss e Treitel (2019), dentre outros.

[2] Ver Cole e Gunther (1998), González-Hermosillo (1999), Kumar e Ravi (2007), Curry, Elmer e Fissel (2007), Rosa e Gartner (2017), Constantin, Peltonen e Sarlin (2018), dentre outros.

comparamos a regressão logística, com cinco modelos de ML bayesiano: *Naive Bayes (NB)*, *Random Forest (RF)*, AdaBoost, *Support Vector Machines* (SVM) e *Decision Trees* (DT).

A terceira é verificar se o tom textual dos relatórios trimestrais dos bancos é capaz de melhorar a acurácia na previsão do risco de falência bancária. Além disso, o presente trabalho utiliza-se de um dicionário variante no tempo na construção da variável de sentimento bancário. Até o momento na literatura de previsão de risco bancário, os escassos trabalhos que aplicam alguma variável de sentimento bancário utilizam um dicionário fixo. Portanto, o uso de um dicionário variante no tempo é algo inédito nesta discussão.

Os resultados indicam que a classificação de risco bancário, via algoritmo *k-means*, foi capaz de classificar 17% da amostra no grupo de maior risco (1), enquanto 83% da amostra ficou no grupo de menor risco de falência (0). Utilizando a métrica do Z-score verificamos que 65% da amostra faz parte do grupo de baixo risco e 35% da amostra no grupo de elevado risco. Desse modo, o algoritmo *k-means* é mais rigoroso em classificar um banco na categoria de alto risco. Na sequência utilizamos os dados já descritos para projetar o risco de insolvência bancária. Os resultados desta etapa mostraram que o modelo de árvore de decisão apresentou o melhor desempenho para a amostra de teste. Além disso, constatou-se que a inclusão de variáveis de sentimento bancário é capaz de melhorar o desempenho dos modelos de previsão, principalmente, quando o sentimento bancário é construído a partir de um dicionário variante no tempo.

O presente artigo é dividido em quatro seções. A primeira é a introdução apresentada acima. A segunda apresenta a metodologia aplicada para a construção da variável de risco de insolvência bancário e das variáveis de sentimento bancário. Nessa seção também são mostrados os modelos de previsão. A terceira ilustra os principais resultados obtidos. Por fim, a quarta indica as conclusões finais, discutindo as principais contribuições e limitações do atigo.

2 Metodologia

A estratégia empírica da pesquisa foi realizada em quatro etapas: a primeira foi usar o algoritmo *k-means* para realizar os agrupamentos (clusters) das instituições bancárias utilizadas na amostra e assim criar uma nova forma de classificação de risco de insolvência bancária. Após a construção da série de cluster, esta passou a ser considerada a variável dependente.

O segundo passo foi a construção das métricas de sentimento bancário. Uma variável foi criada a partir de um dicionário fixo e a outra por meio de um dicionário variante no tempo, com o intuito de verificar se o sentimento do gestor da instituição financeira contido nos relatórios trimestrais é capaz de melhorar a predição do risco de insolvência. A terceira etapa foi empregar técnicas estatísticas de ML supervisionada para prever a classificação das instituições bancárias. A ideia é identificar o modelo preditivo mais robusto para projetar a variável de risco de insolvência construída. Por fim, foram calculadas as acurácias e as taxas de falsos negativos e falsos positivos dos modelos.

2.1 Risco de insolvência bancária

A construção da nova variável usada como *proxy* para o risco de insolvência bancária foi realizada por meio da clusterização dos dados dos bancos escolhidos para o estudo. Para compor a amostra, foram escolhidos 12 bancos de capital aberto com ações negociadas na B3 que podem ser visualizados na Tabela 1.

Tabela 1 – Lista dos bancos

Banco	Definição
Banco ABC	ABC
Banrisul	BANRS
Banco do Brasil	BB
Bradesco	BRA
Banco BRB	BRB
BTG Pactual	BTG
Banco Indusval	IND
Itáu	ITA
Banco Mercantil	MER
Banco Panamericano	PAN
Banco Pine	PIN
Santander	SAN

A janela temporal tem início no quarto trimestre de 2012 e termina no primeiro trimestre de 2021. A frequência dos dados é trimestral. Assim como no trabalho de Damasceno et al. (2021), a variável escolhida para a criação do cluster foi o desvio padrão móvel de doze trimestres do retorno sobre os ativos da empresa (σROA) que é usado no cálculo do Z-score. Além disso, o novo indicador de risco será comparado com o Z-score.

Similarmente a pesquisa de Vieira, Silva e Florêncio (2020), o Z-score foi mensurado por meio das médias móveis e do desvio padrão móvel considerando doze trimestres (três anos), em que os valores dos onze trimestres anteriores e o do trimestre contemporâneo foram utilizados para essa mensuração, conforme apresentado na equação abaixo:

$$Z_{score\{i,t\}} = \frac{\mu ROA_{i,t} + \mu ETS_{i,t}}{\sigma ROA_{i,t}} \tag{1}$$

Em que:

$Z_{score\{i,t\}} =$ risco de insolvência do banco i, no período t;

$\mu ROA_{i,t} =$ média móvel de doze trimestres do retorno sobre os ativos do banco i, no periodo t;

$\mu ETS_{i,t} =$ média móvel de doze trimestres da razão entre o patrimônio líquido e o ativo total para a firma i, no periodo t;

$\sigma ROA_{i,t} =$ desvio padrão móvel de doze trimestres do retorno sobre os ativos da empresa i, no periodo t.

Com o intuito de categorizar os bancos que apresentam um maior risco de insolvência foi utilizado como ponto de corte o primeiro quartil da variável Z-score, o qual foi calculado para cada trimestre analisado. A escolha desse limiar se deu uma vez que os valores menores do Z-score denotam bancos que apresentam maior probabilidade de insolvência, assim os bancos identificados no primeiro quartil receberam o valor 1, já os demais foram categorizados com o valor 0.

Para a criação da nova medida de risco, primeiramente, as variáveis financeiras dos bancos sofreram um processo de normalização com o intuito de melhorar a clusterização. Em seguida, foi utilizado o algoritmo *k-means*, em que foi definida a existência de dois clusters, um para os bancos com alta probabilidade de falência (1) e outro para os bancos com baixa probabilidade de falência (0). Além disso, diferentemente do

trabalho de Damasceno et al. (2021), na presente pesquisa optou-se por clusterizar os bancos um trimestre por vez e e não calcular os clusters com todos os trimestres de uma só vez. Esse caminho foi escolhido, pois caso contrário a clusterização pelo *k-means* ficaria distorcida.

Após a obtenção dessa nova variável, ela foi comparada com a categorização do Z-score mencionada no parágrafo anterior e, posteriormente, utilizada como variável dependente dos modelos de predição.

2.2 Agrupamento por cluster - K-means

O algoritmo *k-means* usa uma forma simples e fácil de classificar um conjunto de dados por meio de um número de cluster, que deve ser fixado antes da execução do algoritmo (Varella e Quadrelli, 2017).

Segundo Fortuna e Maturo (2019) o conceito principal é definir k centroides, um para cada cluster pré-definido, esses centroides devem ser definidos por conta da separação entre os resultados. Desta forma a escolha deve ser feita para colocá-los distantes um dos outros. A seguir deve-se fazer com que cada ponto pertença a um conjunto de dados e vinculá-lo ao centroide mais próximo. Quando da conclusão desta etapa todos os pontos devem pertencer a um grupo.

O algoritmo *k-means* tem como objetivo minimizar uma função objetivo, neste caso específico uma função quadrática de erro que pode ser vista conforme mostra Fortuna e Maturo (2019) na equação abaixo.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{x} \left\| {x^{(j)}}_i - c_j \right\|^2 \tag{2}$$

Este procedimento termina sempre com um resultado de separação, embora não necessariamente ótimo. O algoritmo *k-means* é sensível aos centros de agrupamento selecionados aleatoriamente, por conta disso, é necessária sua definição assertiva e que deve ser tomada como uma premissa.

2.3 Logit

A regressão logística é uma das técnicas mais utilizadas para a área de análise de risco de crédito. Apresenta como característica que a difere da regressão linear discriminante a possibilidade de identificação de crescimento não linear do *default* sob um formato de uma função sigmoide, com crescimento acelerado, gerando maior acurácia preditiva em muitos casos, conforme descrito em Provencher, Baerenklau e Bishop (2002). Para esses autores, este método é bastante utilizado em situações em que a variável dependente assume valores dicotômicos, como é o caso dos problemas de classificação de risco de uma empresa. Sua execução consiste em estimar a probabilidade de ocorrência de um evento com base em um conjunto de variáveis (Silva, Ribeiro e Matias, 2016). Em sua forma funcional a regressão logística pode ser representada por:

$$P_a = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}} \tag{3}$$

Em que P_a representa a probabilidade de uma entidade assumir um dado valor (normalmente expresso em termos de uma variável dicotômica), $\beta_1 \ldots \beta_n$, representam os coeficientes das variáveis (ou características) e $\mathrm{x}_1 \ldots \mathrm{x}_n$ são as variáveis explicativas ou características.

2.4 Naive Bayes

Dentre os modelos supervisionados de ML, as redes bayesianas consistem em um dos métodos mais utilizados e que se configuram como uma classe de modelos estatísticos que apresentam resultados significativos para lidar com eventos de elevada incerteza, sendo aplicado o Teorema de Bayes para identificar comportamentos relacionados à dependência probabilística condicional (Silva, Ribeiro e Matias, 2016).

Especificamente esse modelo utiliza um grupo de variáveis aleatórias (atributos) retratadas em um grafo estatístico por meio de nós e arcos, os quais são definidos em função de uma relação de precedência (condicional), ou seja, refletem a probabilidade de ocorrência de um evento de interesse em estudo, em virtude da ocorrência de um outro evento tido como condicional, obtendo dessa forma a tabela de probabilidade condicional. Por meio dessa abordagem, tem-se o algoritmo de classificação *naive bayes*. Assim, durante a etapa de treinamento, o algoritmo faz uso dos dados dessa subamostra para compreender quais os valores condicionais das variáveis independentes (atributos) que estão associadas às classes do modelo, para que em uma segunda etapa denominada de validação, utilizando os dados da base de teste, seja possível realizar predições sobre a classe de cada observação na amostra, fazendo uso dos valores das variáveis preditoras identificados pelo modelo na etapa de treinamento (Silva, Ribeiro e Matias, 2016).

2.5 Decision Trees e Random Forest

O processo de construção de uma árvore de decisão pode ser resumido em duas etapas: em primeiro lugar, dividimos o espaço do preditor em várias regiões não sobrepostas (por exemplo, regiões J) e, em segundo lugar, a previsão para uma nova observação é dada pela média de os valores de resposta dos dados de treinamento pertencentes à mesma região da nova observação (Nie et al., 2011)

O critério para construir as regiões ou "caixas" é minimizar a soma dos quadrados residuais (RSS), mas não considerando todas as partições possíveis do espaço de feições em J caixas porque seria computacionalmente inviável. Em vez disso, uma divisão binária recursiva é usada: em cada etapa, o algoritmo escolhe o preditor e o ponto de corte, de forma que a árvore resultante tenha o RSS mais baixo. O processo é repetido até que um critério de parada seja alcançado (Nie et al., 2011).

De acordo com Choubin et al. (2018), uma árvore de decisão pode ser considerada um aprendiz básico no campo de ML. A principal vantagem das árvores de decisão em relação aos modelos de regressão linear é que, no caso de um relacionamento altamente não linear e complexo entre os recursos e a resposta, as árvores de decisão podem superar as abordagens clássicas. Embora as árvores de decisão possam não ser muito robustas e geralmente possam fornecer menos precisão preditiva do que alguns dos outros métodos de regressão, essas desvantagens podem ser facilmente melhoradas agregando muitas árvores de decisão, usando métodos, como *bagging* e *random forest*. Esses métodos têm em comum que podem ser considerados métodos de aprendizagem por conjunto.

A metodologia de *random forest* (RF) foi proposta inicialmente por Breiman (2001) como uma maneira de reduzir a variação de árvores de decisão e baseia-se na agregação de *bootstrap* (*bagging*) de árvores de decisão construídas aleatoriamente.

2.6 Support Vector Machine (SVM)

Desde que o SVM foi introduzido a partir da teoria de aprendizagem estatística por Vapnik (1995), uma série de estudos foi anunciada sobre sua teoria e aplicações. Comparado com a maioria das outras técnicas de ML, o SVM aumenta o desempenho em reconhecimento de padrões, estimativa de regressão, previsão de séries temporais financeiras, dentre outras aplicações. Ressalta-se que a breve descrição de SVM se concentra inteiramente no problema de reconhecimento de padrões no campo de classificação. A explicação detalhada e as provas de SVM podem ser verificadas nos livros de Vapnik (1995) e Vapnik (1999).

De acordo com Shin, Lee e Kim (2005), o SVM produz um classificador binário, os chamados hiperplanos de separação ótimos, por meio do mapeamento não linear dos vetores de entrada no espaço de recursos de alta dimensão. O SVM constrói um modelo linear para estimar a função de decisão usando limites de classes não lineares com base em vetores de suporte. Se os dados forem separados linearmente, o SVM treina máquinas lineares para um hiperplano ideal que separa os dados sem erro e na distância máxima entre o hiperplano e os pontos de treinamento mais próximos. Os pontos de treinamento mais próximos do hiperplano de separação ideal são chamados de vetores de suporte. Todos os outros exemplos de treinamento são irrelevantes para determinar os limites das classes binárias. Em casos gerais em que os dados não são separados linearmente, o SVM usa máquinas não lineares para encontrar um hiperplano que minimiza o número de erros do conjunto de treinamento.

2.7 Adaptive boosting (AdaBoost)

Boosting adaptativo (AdaBoost) é um método de ML desenvolvido por Freund, Schapire e Abe (1999). O Adaboost combina todos os classificadores fracos para criar um classificador forte. Ele pode ser aplicado em combinação com vários outros algoritmos de aprendizagem para melhorar o desempenho. Os resultados dos aprendizes fracos são associados a uma acumulação ponderada que representa os resultados finais do classificador ponderado. O Adaboost é adaptativo, uma vez que os aprendizes fracos posteriores são ajustados para favorecer instâncias que foram classificadas erroneamente pelos classificadores anteriores (Taherkhani, Cosma e McGinnity, 2020).

Quando o AdaBoost está em processo de treinamento, ele escolhe as funções ideais para aumentar o desempenho de predição do modelo, reduzir a dimensão e possivelmente encurtar o tempo de execução, já que não há necessidade de calcular recursos irrelevantes (Taherkhani, Cosma e McGinnity, 2020).

2.8 Mensurando os índices de sentimento

Nesta seção será apresentada a metodologia de construção dos índices de sentimento bancário a partir dos relatórios trimestrais (ITR) e Demonstrações Financeiras Padronizadas (DFP) dessas organizações exigidos pela CVM. Dessa maneira, para construção das variáveis de sentimento textual, inicialmente foi obtida a classificação setorial disponibilizada no site da Brasil, Bolsa, Balcão (B3), a qual apresenta a lista dos bancos de capital aberto no país. Posteriormente, foram capturados os ITRs e as DFPs dessas companhias por meio do endereço eletrônico da CVM, coletando dessa forma os relatórios dos três primeiros trimestres de cada ano analisado, mediante os documentos ITR e o último trimestre por sua vez, correspondente aos relatórios DFP.

Assim como no trabalho de Damasceno et al. (2021), após a obtenção dos documentos ITR e DFP, que fazem parte da base de dados da pesquisa, iniciou-se a etapa de transformação dos arquivos do formato original, em *Portable Document Format* (PDF), para texto separado por tabulações (TXT). Posteriormente, estes arquivos gerados na etapa anterior passaram por um processo de pré-ajuste da amostra, sendo removidos os espaços duplos, pontuações, números, bem como os stopwords, que se configuram como uma lista de preposições, pronomes, conjunções e formas verbais que não apresentam relevância explicativa em documentos textuais.

Para explorar as informações textuais presentes nos relatórios ITR e DFP foi empregada a Análise de Processamento Natural, mediante algoritmos de leitura automatizada escritos em linguagem R. Além disso, foi aplicada a técnica do *vector space model*, que considera as palavras presentes nos textos como vetores, os quais serão utilizados para a estimação do peso de cada palavra em um determinado documento de acordo com a frequência das mesmas, conforme apresentado na equação a seguir.

$$P_{i,j} \begin{cases} \frac{(1+\log(Tf_{i,j}))}{(1+\log(a_j))} \times \log \frac{N}{df_i}, & \text{se} \quad Tf_{i,j} \geq 1 \\ 0, & \text{se} \quad Tf_{i,j} = 0 \end{cases} \tag{4}$$

Em que $P_{i,j}$ consiste no peso da palavra i no documento j, por sua vez, $Tf_{i,j}$ compreende a totalidade de ocorrências de uma determinada palavra i em um relatório j, a_j diz respeito à média de frequências das palavras de um documento financeiro, N é o total de relatórios da amostra, e, df_i é o total de relatórios administrativos com ao menos uma ocorrência da palavra i.

A aplicação de ponderações com logaritmos foi realizada objetivando minimizar a atuação de palavras de alta frequência (*outliers*) nos documentos da base de dados, evitando que aqueles apresentem um peso maior na estimação. Acerca desse assunto, Loughran e McDonald (2011) argumentam que tal prática reduz a interferência de *outliers*, comprovando a eficácia desse método após examinarem relatórios 10-K, produzidos por firmas norte-americanas.

De acordo com Shapiro, Sudhof e Wilson (2020) existem duas metodologias gerais para quantificar o sentimento no texto. A primeira é conhecida como metodologia lexical. Esta abordagem se baseia em listas predefinidas de palavras, chamadas de léxicos ou dicionários, com cada palavra atribuída uma pontuação para a emoção de interesse. Geralmente, essas pontuações são simplesmente -1, 0 e 1 para negativo, neutro e positivo, mas alguns léxicos têm mais de três categorias. As aplicações típicas desta abordagem medem o conteúdo emocional de um determinado corpus de texto com base na prevalência de palavras negativas versus positivas no corpus. Esses métodos de correspondência de palavras são chamados de métodos de *bag-of-words* (BOW) devido as características contextuais de cada palavra, como sua ordem no texto, classe gramatical, coocorrência com outras palavras e outras características contextuais específicas ao texto em que a palavra aparece são ignorados. Shapiro, Sudhof e Wilson (2020) afirma que a segunda abordagem, mais incipiente, emprega técnicas de ML para construir modelos complexos para prever probabilisticamente o sentimento de um determinado conjunto de texto.

Nesse ponto precisamos enfatizar que vamos utilizar duas abordagens distintas para a construção do sentimento do gestor. Em uma delas será adotada a estimação do sentimento textual com dicionário fixo, tendo como base o algoritmo de leitura desenvolvido por Machado et al. (2019). Nele são desconsiderados os termos que não estavam associados, no referido dicionário, a nenhuma das duas tipologias de sentimento.

Em um segundo momento será utilizada a abordagem com dicionários variantes no tempo. Essa

discussão está disponível em Lima, Godeiro e Mohsin (2019) e, conforme destacado por esses, a suposição de um dicionário invariável no tempo não parece ser realista em documentos que introduzem novas palavras ao longo do tempo ou se o vocabulário usado em períodos de recessão difere do usado em períodos de expansões econômicas. Os autores ressaltam que mesmo se o vocabulário fosse constante ao longo do tempo, o poder preditivo de algumas palavras pode variar, ou seja, a relevância das palavras se alteram, mas a literatura existente não explica esse efeito e, portanto, os preditores resultantes não refletem as informações textuais mais preditivas encontradas nos documentos em um determinado momento. Com base nisso, aplicamos para a construção do sentimento bancário de dicionário variante no tempo a abordagem desenvolvida por Lima, Godeiro e Mohsin (2019).

Assim, utilizando a metodologia proposta pelos autores para construir o dicionário variante no tempo, primeiramente criamos um vetor, $X_{i,t}$, em que cada elemento do vetor mostra observações em série temporal da frequência em que cada palavra (ou combinação de palavras) aparece nos ITRs de cada banco *i* até o tempo *t*. Portanto, esta etapa transforma as palavras em valores numéricos sem usar um dicionário pré-especificado (fixo). Essa representação numérica é de alta dimensão e esparsa; portanto, a redução da dimensionalidade deve ser empregada na próxima etapa. Na segunda etapa, usamos o ML para selecionar as palavras mais preditivas $X^*_{i,t} \subset X_{i,t}$.

O modelo de *elastic net* foi escolhido para realizar a segunda etapa:

$$y_{i,t+h} = W'_{i,t}\boldsymbol{\beta}_h + X'_{i,t}\phi_h + \epsilon_{i,t+h} \tag{5}$$

em que $h \geq 0$ é o horizonte de previsão, $\beta\widehat{\beta}_h$ e $\phi\widehat{\phi}_h$ são estimadas minimizando a seguinte função objetivo:

$$\min_{\beta_h,\phi_h} \sum_t \left(y_{i,t+h} - W'_{i,t}\beta_h - X'_t\phi_h\right)^2 + \lambda_1 \left\|\phi_h\right\|_{\ell_1} + \lambda_2 \left\|\phi_h\right\|_{\ell_2} \tag{6}$$

em que W_t é um vetor *k x 1* de preditores pré-determinados, como defasagens de y_t bem como preditores tradicionais de dados estruturados e $\|\cdot\|_{\ell_1}$ e $\|\cdot\|_{\ell_2}$ são a norma ℓ_1 e ℓ_2, respectivamente. Então, a partir da seleção das palavras com maior poder preditivo, temos para cada período *t* um conjunto de palavras que servem como dicionário léxico para a obtenção da série de sentimento bancário. No presente artigo, vamos aplicar esse método para a nova variável de risco bancário. Assim, o algoritmo vai selecionar em cada relatório ITR no período *t* de cada banco *i* as palavras que mais explicam mudanças no risco de insolvência e a partir desse conjunto de palavras, ou seja, dicionário variante no tempo, são geradas as séries de sentimento bancário para cada instituição.

Por fim, ambas abordagens de dicionário, calculam o índice de sentimento pela diferença entre palavras positivas e negativas, dividida pela soma de palavras positivas e negativas, como foi proposto por Hubert e Labondance (2018):

$$SB_t = \frac{\text{PositiveWords}_{\text{t}} - \text{NegativeWords}_{\text{t}}}{\text{PositiveWords}_{\text{t}} + \text{NegativeWords}_{\text{t}}} \tag{7}$$

Portanto, obtemos a medida de sentimento bancário, *SB*, que varia entre -1 e 1.

2.9 Preditores

Para prever os risco de insolvência bancária usamos dados macroeconômicos e financeiros dos bancos como preditores. A janela temporal e frequência dos dados é idêntica aos da Seção 2.1. A Tabela 2 apresenta as variáveis utilizadas como preditores.

Tabela 2 – Preditores

Variável	Operacionalização	Fonte
Sentimento Textual Bancário de Dicionário Fixo (SBF)	Tom do sentimento textual de cada relatório da amostra com dicionário fixo	Resultados da pesquisa
Sentimento Textual Bancário de Dicionário Variante (SBV)	Tom do sentimento textual de cada relatório da amostra com dicionário variante	Resultados da pesquisa
Tamanho (TAM)	Razão entre depósitos totais e ativo total	Economática e Resultados da pesquisa
Retorno sobre os ativos (ROA)	Razão entre o lucro operacional e ativo total	Economática e Resultados da pesquisa
Capitalização (ETS)	Razão entre patrimônio líquido e ativo total	Economática e Resultados da pesquisa
Produto Interno Bruto (PIB)	Variação percentual do PIB real em relação ao trimestre anterior	IBGE
Índice Nacional de Preços ao Consumidor Amplo (IPCA)	Variação percentual do IPCA em relação ao trimestre anterior	IBGE
Ibovespa (IBOV)	Variação percentual do Ibovespa em relação ao trimestre anterior	Anbima
Ciclo Econômico (CICLO)	Ciclo do PIB real trimestral obtido pelo filtro HP	Resultados da pesquisa

A utilização de uma variável que mede o tom do sentimento textual dos relatórios trimestrais já existe na literatura. Damasceno et al. (2021) construiu uma variável de sentimento bancário por meio do dicionário fixo de Machado et al. (2019) para os bancos brasileiros de capital negociados na B3. Entretanto, como já foi mencionado na seção anterior, um dicionário fixo ao longo do tempo não consegue captar a alteração da importância das palavras ao longo do tempo e, principalmente, não é capaz de captar o surgimento de novos termos importantes, por exemplo, o termo "Covid-19"surgiu em 2020 nos relatórios dos bancos e ele não é captado pelo dicionário de Machado et al. (2019). Então, na presente pesquisa também construímos uma variavel de sentimento bancário com um dicionário variante no tempo a partir da abordagem de Lima, Godeiro e Mohsin (2019).

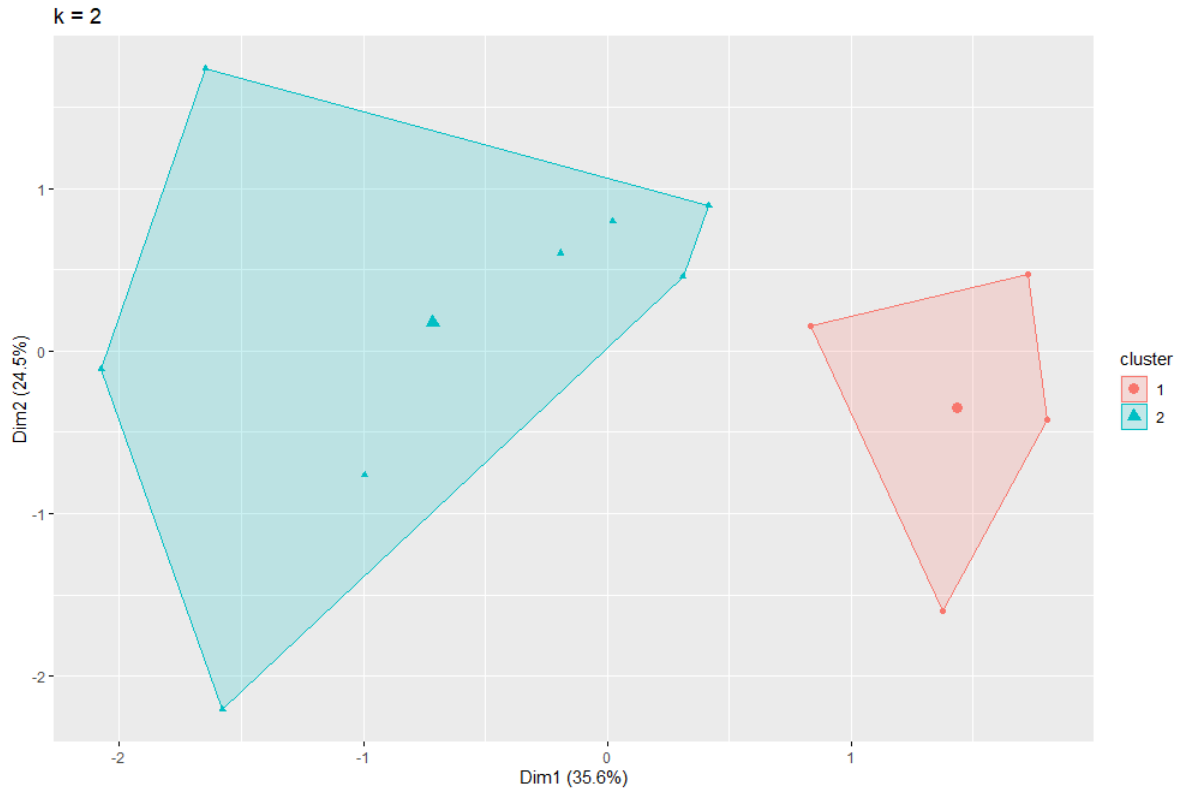
Com relação as variáveis financeiras, essas foram escolhidas de acordo com a literatura já consolidada, entre esses, destaca-se: Rosa e Gartner (2017), Vieira, Silva e Florêncio (2020). No caso das variáveis macroeconômicas, o PIB já é amplamente utilizado na literatura. Assim como Suss e Treitel (2019), adicionamos a inflação (IPCA) e um índice acionário (IBOV). Suss e Treitel (2019) também recomendaram que trabalhos futuros sobre o tema incorporassem ao conjunto de preditores uma variável *proxy* para captar os efeitos do ciclo econômico. Para atender esta sugestão incluímos o ciclo econômico (Ciclo) e esse foi obtido a partir da aplicação do filtro Hodrick-Prescott na série do PIB.

3 Resultados

3.1 Clusterização e risco bancário

Para a realização do agrupamento dos clusters foi considerado o desvio padrão móvel do retorno sobre os ativos (σROA), empregado para a execução do algoritmo *k-means*. O Z-score foi usado como parâmetro de comparação com os resultados obtidos pelo método de clusterização. Para tal, foi criada uma variável *dummy* que classificou o grupo de maior risco (1), definido pelo primeiro quartil do Z-score, e o grupo de menor risco (0) é representado pelos demais valores. A Figura 1 mostra um exemplo do processo de clusterização pelo *k-means*, neste caso, do primeiro trimestre de 2021.

Figura 1 – Clusterização formada a partir do desvio padrão móvel do retorno sobre os ativos (σROA)



Fonte: Resultados da pesquisa.

Nota: O software classifica dois clusters entre 1 e 2, mas é equivalente aos valores 0 e 1 usados no presente artigo.

Ao compararmos os resultados da classificação de risco bancário, definida pelo *k-means*, com o Z-score é possível perceber que há uma distinção na separação dos grupos quanto a métrica adotada, conforme descrito na Tabela 3. Esse achado também foi reportado por Damasceno et al. (2021), em que esses encontraram similaridade entre Z-score e a clusterização apenas para o grupo de menor risco.

Tabela 3 – Comparação do agrupamento dos clusters com a classificação do Z-score - 1T21

	Total de observações com o Z-score	Total de observações com σROA	Observações comuns quanto ao risco	Percentual de similaridade
Grupo de maior risco (1)	139	69	25	17.99%
Grupo de menor risco (0)	257	327	163	63.42%
Total de observações	396	396	188	-

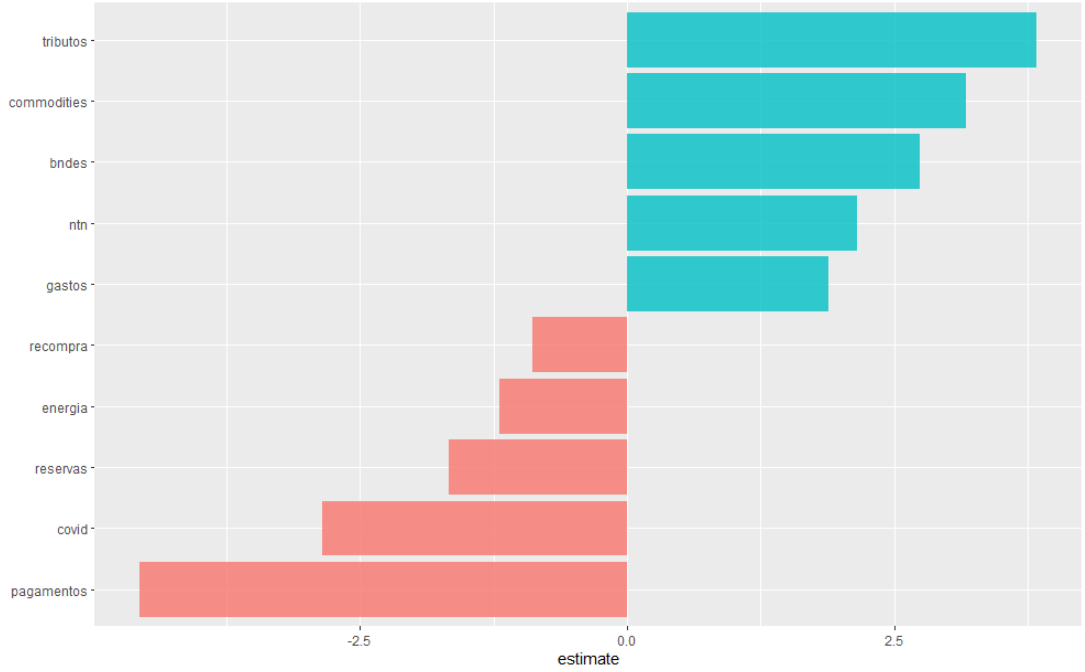
Como pode ser visto na Tabela 3, o método *k-means* é mais rigoroso quanto a classificação dos bancos em uma situação de maior risco. Além disso, verificamos que o método de agrupamento de clusters classificou 63.42% das observações de menor risco rotuladas pelo Z-score. A maior diferença acontece quando é considerado do grupo de elevado risco. O Z-score categorizou 139 observações no grupo de maior risco. No entanto, pelo método de clusters, houve *matching* com apenas 17.99% das observações. Ressalta-se ainda, que ambos os percentuais foram obtidos após a comparação das mesmas observações pelas duas medidas de classificação empregadas. Com a categorização da *dummy* do Z-score 65% da amostra faz parte do grupo de baixo risco e 35% da amostra no grupo de elevado risco. Com a variável de cluster 83% da amostra faz parte do grupo de menor risco, enquanto, 17% do grupo de maior risco. Esse resultado sugere que o método de classificação pelo agrupamento é mais rigoroso, quanto a categorização do grupo de maior risco, do que o Z-score.

3.2 Análise do sentimento bancário

Nesta seção serão apresentados os índices de sentimento bancário, construídos com o dicionário variante no tempo, dos doze bancos utilizados na amostra. Como foi mencionado na seção 2.8, foi usado para a criação das séries de sentimento variante no tempo a abordagem proposta por (Lima, Godeiro e Mohsin, 2019). Assim, para cada instituição bancária o algoritmo de *Elastic Net* seleciona as palavras mais preditivas para a nova variável de risco em cada ITR. A partir disso a série de tom textual é gerada. As palavras consideradas mais preditivas para o risco de falência varia de acordo com os relátórios de cada instituição bancária.

A Figura 2, ilustrativamente, apresenta os 5 principais termos positivos e negativos mais preditivos para o risco de insolvência encontrados nos relatórios do Banco ABC. Neste caso, o dicionário variante no tempo encontrou que termos como "títulos", "gastos", "covid"e "pagamentos"contidos nos relatórios trimestrais do Banco ABC são as palavras que possuem um maior efeito no risco bancário.

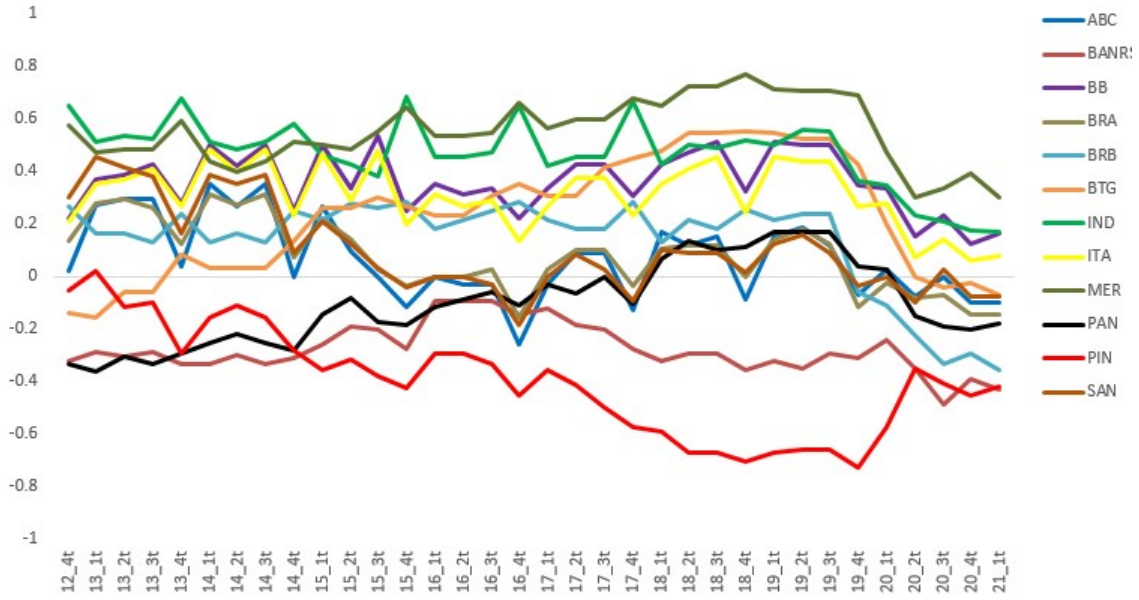
Figura 2 – Coeficientes mais positivos e negativos na determinação do sentimento bancário - Banco ABC - 1T21



Fonte: Resultados da pesquisa.

Após o algoritmo *Elastic Net* encontrar as palavras mais relevantes na explicação do risco bancário no trimestre t, o valor do tom do sentimento é obtido. No exemplo da Figura 2, dado esse conjunto de palavras, o valor do tom do Banco ABC no primeiro trimestre de 2021 foi de -0,10, ou seja, um tom pessimista.

Figura 3 – Séries de sentimento bancário com dicionário variante no tempo



Fonte: Resultados da pesquisa.

Já a Figura 3 ilustra as séries de sentimento bancário de cada instituição bancária. É possivel notar que ao longo do tempo o tom dos ITRs do Banco Mercantil foi o mais otimista ao longo do tempo, enquanto

o Banco Pine apresentou o tom mais pessimista durante o período de análise. A Figura 3 também mostra que a partir do segundo trimestre de 2020 o tom do sentimento textual de praticamente todos os bancos obteve uma forte queda, ou seja, o tom dos relatórios passou a ser mais pessimista após o início da pandemia da Covid-19, iniciada durante esse período.

3.3 Análise dos modelos de previsão

Assim como no trabalho de Damasceno et al. (2021) para o cálculo da acurácia em um primeiro momento é necessário fazer uso da matriz de confusão, sendo essa considerada como uma ferramenta útil durante a etapa de avaliação de modelos de classificação, ao fornecer os dados necessários para a mensuração das medidas de especificidade e sensibilidade, as quais posteriormente são utilizadas para o cálculo da acurácia dos modelos. A sensibilidade pode ser definida como:

$$\text{sensibilidade } = \frac{TP}{TP + FN} \tag{8}$$

Em que: sensibilidade = Taxa de verdadeiros positivos, indicando o percentual de instituições bancárias que possuem elevado risco de falência futura, e que foram classificadas corretamente; TP = Verdadeiro positivo, indica o número de casos positivos (maior risco de insolvência) que foram corretamente identificados; FN = Falso negativo, indica o número de casos positivos (maior risco de insolvência) que foram classificados de forma incorreta como casos negativos (menor risco de insolvência).

Já a especificidade é definida pela seguinte expressão:

$$\text{especificidade } = \frac{TN}{TN + FP} \tag{9}$$

especificidade = Taxa de verdadeiros negativos, indica a proporção de bancos que apresentam menor risco de insolvência e que foram classificados corretamente; TN = Verdadeiro negativo, indica o número de casos negativos (menor risco de insolvência) que foram corretamente identificados; FP = Falso positivo, indica o número de casos negativos (menor risco de insolvência) que foram classificados de forma incorreta como casos positivos (maior risco de insolvência).

Finalmente, a acurácia é mensurada como:

$$\text{Acurácia } = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

Foram escolhidos seis modelos de previsão, conforme descrito na Tabela 4. Como mencionado anteriormente, além de encontrar o modelo com maior poder de predição do risco bancário, também vamos verificar se as variáveis de sentimento bancário são capazes de melhorar a acurácia dos modelos, ou seja, se o tom contido nos relatórios trimestrais dos bancos é uma informação relevante no monitoramento do risco de falência de tais instituições.

Tabela 4 – Definição dos modelos de previsão

Sigla	Definição	Sigla	Definição
NB1	Modelo Naive Bayes s/ sentimento bancário	NB2	Modelo Naive Bayes com SBF
NB3	Modelo Naive Bayes com SBV	NB4	Modelo Naive Bayes com SBF e SBV
LOGIT1	Modelo Logit s/ sentimento bancário	LOGIT2	Modelo Logit com SBF
LOGIT3	Modelo Logit com SBV	LOGIT4	Modelo Logit com SBF e SBV
SVM1	Modelo SVM s/ sentimento bancário	SVM2	Modelo SVM com SBF
SVM3	Modelo SVM com SBV	SVM4	Modelo SVM com SBF e SBV
RF1	Modelo Random Forest s/ sentimento bancário	RF2	Modelo Random Forest com SBF
RF3	Modelo Random Forest com SBV	RF4	Modelo Random Forest com SBF e SBV
AB1	Modelo AdaBoost s/ sentimento bancário	AB2	Modelo AdaBoost com SBF
AB3	Modelo AdaBoost com SBV	AB4	Modelo AdaBoost com SBF e SBV
DT1	Modelo Árvore de Decisão s/ sentimento bancário	DT2	Modelo Árvore de Decisão com SBF
DT3	Modelo Árvore de Decisão com SBV	DT4	Modelo Árvore de Decisão com SBF e SBV

Portanto, como é possível ver na Tabela 5, para cada modelo foram testadas quatro especificações. A primeira é a estimação com todos os preditores menos as duas vaariáveis de sentimento bancário (SBF e SBV). A segunda é a estimação com todos os preditores menos SBV. A terceira é a estimação com todos os preditores menos SBF. A quarta é a estimação com todos os preditores.

Assim como nos trabalhos de Rosa e Gartner (2017) e Damasceno et al. (2021), foi aplicada a etapa de *cross-validation* para evitar a ocorrência de *over-fitting*. Com relação a divisão da amostra entre treino e teste, optou-se por colocar 80% da amostra para o treino dos modelos e 20% da amostra para o teste e mensuração da acurácia.

As acurácias das previsões dos modelos pode ser vista na Tabela 5. De forma geral, o modelo que atingiu a maior taxa de acurácia foi o DT (especificação 3), com acurácia de 95% para a amostra de teste. Já os modelos SVM e LOGIT apresentaram a pior acurácia, com valores de 85%. Vale destacar que esses dois últimos modelos, assim como o modelo NB, não apresentaram alteração em suas acurácias com as modificações nos preditores. Esse fato ocorreu porque no SVM e LOGIT em todas as 4 especificações só previram valores 0, ou seja, eles não foram capazes de prever valores 1. No caso do modelo NB, nas quatro especificações ele só conseguiu prever apenas um valor para alto risco.

Tabela 5 – Acurácia das previsões dos modelos

Modelo/Especificação	1	2	3	4
NB	0,8625	0,8625	0,8625	0,8625
RF	0,9125	0,8875	0,9125	0,925
SVM	0,8500	0,8500	0,8500	0,8500
LOGIT	0,8500	0,8500	0,8500	0,8500
AB	0,9250	0,9125	0,9375	0,9125
DT	0,9250	0,9250	0,9500	0,9375

Também destacamos que com excessão dos três modelos citados no parágrafo anterior, a inclusão do sentimento bancário proporciona ganhos de desempenho dos modelos, pois em nenhum deles a especificação do tipo 1 (modelos sem variáveis de sentimento bancário) conseguiu superar a acurácia das especificações com variáveis de sentimento bancário. Pode-se também perceber que modelos com a especificação 3 (modelos com SBV e sem SBF), com excessão do RF, obtiveram os melhores desempenhos.

Quando comparado com os resultados obtidos pelos estudos com temática similiar, a presente pesquisa conseguiu um desempenho superior aos trabalhos de Climent, Momparler e Carmona (2019), Suss e Treitel (2019), Huang e Yen (2019) Damasceno et al. (2021). Contudo, Viswanathan, Srinivasan e Hariharan (2020) obtiveram um desempenho superior a 95%, neste caso, os autores reportam uma acurácia de 95,93% com o modelo *random forest*.

Os resultados obtidos mostram alguns pontos que merecem destaque. O primeiro deles é que os modelos de ML do tipo *emsemble* obtiveram acurácias superiores ao modelo tradicional (modelo logit). Esses resultados convergem ao encontrado pela literatura, com excessão do trabalho de Damasceno et al. (2021) que encontrou um melhor desempenho com o modelo logit, pois o autor reporta que o modelo *random forest* foi descartado por apresentar *over-fitting*.

O segundo é que a inclusão do sentimento bancário como preditor é capaz de melhorar o desempenho dos modelos de predição. O terceiro é que entre as variáveis de sentimento bancário, a que aplicou um dicionário variante no tempo conseguiu um resultado melhor em termos de acurácia. Cabe ressaltar que esses resultados ainda não foram reportados na literatura por se tratar de uma discussão inédita.

3.4 Taxas de falsos positivos e falsos negativos

A questão da identificação do melhor modelo para prever o risco de insolvência bancária, exposta na seção anterior, é de grande relevância para os tomadores de decisão de um sistema de alerta preventivo. Contudo, isso não é suficiente para um sistema de acompanhamento do risco bancário. De acordo com Suss e Treitel (2019) duas métricas de desempenho diferentes são relevantes para os tomadores de decisão de um sistema de alerta preventivo: as taxas de erro de falso negativo (FN) e falso positivo (FP).

A taxa FN representa a proporção de bancos que efetivamente possuem alto risco, no entanto, o algoritmo classifica como de baixo risco, é provavelmente a mais importante das duas métricas do ponto de vista regulatório. Um sistema de alerta preventivo que não aciona o alarme quando deveria, especialmente para instituições grandes e sistemicamente importantes, pode ter consequências negativas para toda a economia. Obviamente, para minimizar a taxa de FN, podemos apenas diminuir o limite de probabilidade prevista pelo qual classificamos os bancos de baixo e alto risco. Já a taxa FP é a proporção de bancos de baixo risco e que são classificados erroneamente como de alto risco. A taxa FN é calculada como $\frac{FN}{FN+TP}$ e taxa FP é obtida por $\frac{FP}{FP+TN}$.

Tabela 6 – Taxas de falsos positivos e falsos negativos

	Taxa de FN	Taxa de FP
NB	14,10%	16,45%
RF	5,71%	4,35%
SVM	15,00%	16,45%
LOGIT	15,00%	16,45%
AB	4,35%	2,94%
DT	2,94%	2,94%

É possivel ver na Tabela 6 que o modelo DT obteve a menor taxa de FN, mostrando que apenas 2,94% das previsões foram classificadas de baixo risco quando possuim risco alto. Comparando esse resultado com

o mesmo limiar de 25% do trabalho de Suss e Treitel (2019), o presente artigo conseguiu obter taxas de FN próximas ao encontrado pelos autores. O menor valor que os autores encontraram foi de 2,2% com o modelo *random forest*. Com relação as taxas de FP, os modelos do presente trabalho conseguiram taxas mais baixas quando comparada ao encontrado por Suss e Treitel (2019).

4 Conclusão

Esse artigo tem três contribuições para a literatura que trata de risco de insolvência bancária, sendo elas: utilização da técnica não-supervisionada de cluster para classificar, conforme o risco de insolvência, os bancos negociados na B3; construção dos índices de sentimento bancário a partir dos relatórios trimestrais (ITR) e Demonstrações Financeiras Padronizadas (DFP) dessas organizaçõesexigidos pela CVM, tendo como base a abordagem com dicionários variantes no tempo; por fim, comparamos várias técnicas de aprendizado de máquina e estatísticas clássicas, implementando um procedimento rigoroso de validação cruzada aleatória de bloco duplo para avaliar o desempenho da previsão fora da amostra.

Na análise de sentimento verificamos que, ao longo do tempo, o tom dos ITRs do Banco Mercantil foi o mais otimista, enquanto o Banco Pine apresentou o tom mais pessimista durante o período de análise. Outro dado interessante é que o índice gerado foi capaz de captar os efeitos negativos da crise sanitária da Covid-19, presente nos relatórios dos bancos. Em outras palavras, a partir do segundo trimestre de 2020 o tom do sentimento textual de praticamente todos os bancos obteve uma forte queda, ou seja, o tom dos relatórios passou a ser mais pessimista.

Os resultados obtidos mostraram que o algoritmo de árvore de decisão é superior na previsão dos dados da amostra de teste de classificação, embora os algoritmos do tipo *ensemble* (*random forest* e AdaBoost) tenham se aproximado em termos de acurácia. Também constatamos que a inclusão do sentimento bancário como preditor promove ganhos de acurácia, principalmente, se for utilizado o dicionário variante no tempo. Além disso, constatamos que o modelo de árvore de regressão possui o menor custo em termos de falsos negativos e falsos positivos.

Do ponto de vista prático, apresentamos uma medida alternativa para classificação de risco, desenvolvemos um indicador capaz de captar o sentimento do gestor do banco quanto as condições de mercado e, por fim, projetamos o risco de insolvência bancária por meio de diferentes técnicas de aprendizagem de máquina. A fusão destas etapas auxilia os gestores, reguladores e supervisores a antecipar comportamentos atípicos ou que envolvem maior risco de insolvência das instituições financeiras.

Cabe ressaltar que este artigo tratou, exclusivamente, os bancos de capital aberto negociados na B3 no Brasil, devido o nível de regulação e a quantidade de informações disponíveis para estes. Portanto, estudos futuros podem expandir o escopo para os demais bancos do país. Também podem ser construídas variáveis de sentimento bancário com o uso de outros dicionários. Dada a janela amostral escolhida em que houve a pandemia da Covid-19, em 2020, então, trabalhos futuros podem verificar os resultados antes e depois da pandemia e mensurar os resultados nesses dois ambientes.

Referências

BARBOSA, J. H. d. F. Early warning system para distress bancário no brasil. 2017.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

CHOU, C.-H.; HSIEH, S.-C.; QIU, C.-J. Hybrid genetic algorithm and fuzzy clustering for bankruptcy prediction. *Applied Soft Computing*, Elsevier, v. 56, p. 298–316, 2017.

CHOUBIN, B. et al. Precipitation forecasting using classification and regression trees (cart) model: a comparative study of different approaches. *Environmental earth sciences*, Springer, v. 77, n. 8, p. 1–13, 2018.

CLIMENT, F.; MOMPARLER, A.; CARMONA, P. Anticipating bank distress in the eurozone: An extreme gradient boosting approach. *Journal of Business Research*, Elsevier, v. 101, p. 885–896, 2019.

COLE, R. A.; GUNTHER, J. W. Predicting bank failures: A comparison of on-and off-site monitoring systems. *Journal of Financial Services Research*, Springer, v. 13, n. 2, p. 103–117, 1998.

CONSTANTIN, A.; PELTONEN, T. A.; SARLIN, P. Network linkages to predict bank distress. *Journal of Financial Stability*, Elsevier, v. 35, p. 226–241, 2018.

CURRY, T. J.; ELMER, P. J.; FISSEL, G. S. Equity market data, bank failures and market efficiency. *Journal of Economics and Business*, Elsevier, v. 59, n. 6, p. 536–559, 2007.

DAMASCENO, P. I. d. S. et al. Risco de insolvência e sentimento textual bancário: uma análise dos bancos de capital aberto no brasil. Universidade Federal da Paraíba, 2021.

ERDOGAN, B. E. Prediction of bankruptcy using support vector machines: an application to bank bankruptcy. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 83, n. 8, p. 1543–1555, 2013.

EZZAMEL, M.; MAR-MOLINERO, C.; BEECH, A. On the distributional properties of financial ratios. *Journal of Business Finance & Accounting*, Wiley Online Library, v. 14, n. 4, p. 463–481, 1987.

FORTUNA, F.; MATURO, F. K-means clustering of item characteristic curves and item information curves via functional principal component analysis. *Quality & Quantity*, Springer, v. 53, n. 5, p. 2291–2304, 2019.

FREUND, Y.; SCHAPIRE, R.; ABE, N. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, JAPANESE SOC ARTIFICIAL INTELL, v. 14, n. 771-780, p. 1612, 1999.

GENTZKOW, M.; KELLY, B.; TADDY, M. Text as data. *Journal of Economic Literature*, v. 57, n. 3, p. 535–74, 2019.

GONZÁLEZ-HERMOSILLO, M. B. *Determinants of ex-ante banking system distress: A macro-micro empirical exploration of some recent episodes*. [S.l.]: International Monetary Fund, 1999.

GUPTA, A.; SIMAAN, M.; ZAKI, M. When positive sentiment is not so positive: Textual analytics and bank failures. *Available at SSRN 2773939*, 2016.

HSU, M.-F. A fusion mechanism for management decision and risk analysis. *Cybernetics and Systems*, Taylor & Francis, v. 50, n. 6, p. 497–515, 2019.

HUANG, Y.-P.; YEN, M.-F. A new perspective of performance comparison among machine learning algorithms for financial distress prediction. *Applied Soft Computing*, Elsevier, v. 83, p. 105663, 2019.

HUBERT, P.; LABONDANCE, F. Central bank sentiment. *URL: https://www. nbp. pl/badania/seminaria/14xi2018. pdf. Working paper*, 2018.

KARELS, G. V.; PRAKASH, A. J. Multivariate normality and forecasting of business bankruptcy. *Journal of Business Finance & Accounting*, Wiley Online Library, v. 14, n. 4, p. 573–593, 1987.

KIM, Y. J.; BAIK, B.; CHO, S. Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Systems with Applications*, Elsevier, v. 62, p. 32–43, 2016.

KUMAR, P. R.; RAVI, V. Bankruptcy prediction in banks and firms via statistical and intelligent techniques–a review. *European journal of operational research*, Elsevier, v. 180, n. 1, p. 1–28, 2007.

LEPETIT, L.; STROBEL, F. Bank insolvency risk and time-varying z-score measures. *Journal of International Financial Markets, Institutions and Money*, Elsevier, v. 25, p. 73–87, 2013.

LIMA, L. R.; GODEIRO, L.; MOHSIN, M. Time-varying dictionary and the predictive power of fed minutes. *Available at SSRN 3312483*, 2019.

LOUGHRAN, T.; MCDONALD, B. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, Wiley Online Library, v. 66, n. 1, p. 35–65, 2011.

MACHADO, M. A. V. et al. Índice de sentimento textual: uma análise empírica do impacto das notícias sobre risco sistemático. *Revista Contemporânea de Contabilidade*, v. 16, n. 40, p. 24–42, 2019.

NIE, G. et al. Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, Elsevier, v. 38, n. 12, p. 15273–15285, 2011.

PAULE-VIANEZ, J.; GUTIÉRREZ-FERNÁNDEZ, M.; COCA-PÉREZ, J. L. Prediction of financial distress in the spanish banking system: An application using artificial neural networks. *Applied Economic Analysis*, Emerald Publishing Limited, 2019.

PROVENCHER, B.; BAERENKLAU, K. A.; BISHOP, R. C. A finite mixture logit model of recreational angling with serially correlated random utility. *American Journal of Agricultural Economics*, Wiley Online Library, v. 84, n. 4, p. 1066–1075, 2002.

RAVI, V. et al. Soft computing system for bank performance prediction. *Applied soft computing*, Elsevier, v. 8, n. 1, p. 305–315, 2008.

ROSA, P. S.; GARTNER, I. R. Financial distress in brazilian banks: an early warning model. *Revista Contabilidade & Finanças*, SciELO Brasil, v. 29, p. 312–331, 2017.

SHAPIRO, A. H.; SUDHOF, M.; WILSON, D. J. Measuring news sentiment. *Journal of Econometrics*, Elsevier, 2020.

SHIN, K.-S.; LEE, T. S.; KIM, H.-j. An application of support vector machines in bankruptcy prediction model. *Expert systems with applications*, Elsevier, v. 28, n. 1, p. 127–135, 2005.

SILVA, R. A.; RIBEIRO, E. S.; MATIAS, A. B. Aprendizagem estatística aplicada à previsão de default de crédito. *Revista de Finanças Aplicadas*, v. 7, n. 2, p. 1–19, 2016.

SUN, J.; LI, H. Financial distress prediction using support vector machines: Ensemble vs. individual. *Applied Soft Computing*, Elsevier, v. 12, n. 8, p. 2254–2265, 2012.

SUSS, J.; TREITEL, H. Predicting bank distress in the uk with machine learning. Bank of England Working Paper, 2019.

TAHERKHANI, A.; COSMA, G.; MCGINNITY, T. M. Adaboost-cnn: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning. *Neurocomputing*, Elsevier, v. 404, p. 351–366, 2020.

VAPNIK, V. *The nature of statistical learning theory*. [S.l.]: Springer science & business media, 1995.

VAPNIK, V. N. An overview of statistical learning theory. *IEEE transactions on neural networks*, IEEE, v. 10, n. 5, p. 988–999, 1999.

VARELLA, J. L.; QUADRELLI, G. Redes neurais e análise de potência. *Revista de Tecnologia Aplicada*, v. 6, n. 3, 2017.

VIEIRA, C. A. M.; SILVA, R. R. da; FLORÊNCIO, D. B. Complexidade e risco dos conglomerados financeiros operantes no brasil complexity and risk in brazilian banking holding companies. *Revista BASE–v*, v. 17, n. 2, 2020.

VISWANATHAN, P.; SRINIVASAN, S.; HARIHARAN, N. Predicting financial health of banks for investor guidance using machine learning algorithms. *Journal of Emerging Market Finance*, SAGE Publications Sage India: New Delhi, India, v. 19, n. 2, p. 226–261, 2020.

XIA, Y. et al. A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, Elsevier, v. 78, p. 225–241, 2017.