Product 3 - -Product 3 - Technical document containing evaluations and analyzes of Credit Risk and Banking Spread.

Avaliação do Risco de Crédito das Operações do BNDES

Patrick Alves

**Resumo**:

Utilizando o cadastro de informações da Receita Federal, propomos uma metodologia de mensuração da inadimplência e consequentemente do risco de crédito nas operações de financiamento do Banco Nacional de Desenvolvimento Econômico e Social (BNDES). Utilizamos modelos de aprendizado de máquina para mensurar a probabilidade de inadimplência e categorizamos as probabilidades enquadrar as empresas nos respectivos riscos de crédito.

**Introdução**

O crédito é necessário para financiar diversos iniciativas e negócios. Recentemente, credores e empresas de tecnologia financeira (“fintech”) estão procurando usar fontes alternativas de dados e formas inovativas de análise de risco de crédito. Essas inovações podem ampliar o acesso ao crédito, especialmente para pessoas com históricos de crédito escassos. No mercado financeiro do consumidor, dados alternativos referem-se a informações usadas para avaliar a qualidade de crédito que geralmente não fazem parte de um relatório de crédito. Alguns exemplos incluem:

* Pagamentos de aluguel.
* Pagamentos por celular.
* Pagamentos de TV a cabo.
* Sites de reputação on-line.
* Métricas de análise de sentimento de mídias sociais.

O uso de dados alternativos pode expandir o acesso ao crédito, onde mesmo os clientes sem histórico de pagamentos podem ter uma avaliação de risco de crédito com precisão, desde que as fontes alternativas apresentem variabilidade. Outra vantagem é disponibilizar escores alternativos de risco de crédito, que podem ser combinados com escores de bureaus de crédito tradicionais, adicionando fontes de informação na gestão de risco.

Atualmente alguns credores adotam políticas contratuais que exigem escores de crédito acima de determinado limiar, mas sem restringir que sejam escores de *bureaus* de crédito tradicionais. Alguns desses credores podem estar dispostos a conceder o empréstimo, se encontrarem critérios objetivos confiáveis para determinar quais pessoas têm menos probabilidade de inadimplência no empréstimo, com base em fontes de dados alternativas ().

Os dados tradicionalmente usados pelos credores e *bureaus* de crédito, podem não refletir todas as atividades que a empresa se envolve. Dados alternativos podem fornecer informações mais atualizadas e em tempo real. Por exemplo, podem existir cadastros desatualizados ou informações incompletas mesmo em dados tradicionais.

Os escores de crédito tradicionais são fortemente influenciados pela própria conduta financeira da pessoa física ou jurídica. E alguns dados alternativos não são tão intimamente relacionados a conduta financeira propriamente dita, fazendo tais fontes de informação interessantes para concessão de crédito com objetivos sociais ou para financiar projetos de maior risco, tais como linhas voltadas para inovação tecnologia.

Entretanto, o uso de fontes de dados alternativas pode penalizar ou recompensar certos grupos ou comportamentos de maneiras difíceis de prever. Por exemplo, empresas de cujos processos de produção utilizam solução em nuvem e que adotaram de trabalho remoto universal. Essas empresas podem mudar o endereço de seu estabelecimento sem impactar suas atividades produtivas, e isso pode dar uma falsa impressão de instabilidade que pode afetar seu acesso ao crédito.

**Fontes de Dados**

Apresentamos a organização em formato SAS dos dados abertos da Receita Federal (RF) e do Banco Nacional de Desenvolvimento Econômico e Social (BNDES). Essas informações são microdados de empresas disponibilizados, respectivamente, pelo portal de dados abertos[[1]](#footnote-1) do Governo Federal e pelo portal da Transparência[[2]](#footnote-2) do BNDES. A organização e documentação desses microdados é uma etapa importante em avaliações de financiamentos do BNDES, performance econômica de empresas e probabilidade de inadimplência.

Neste documento apresentamos em cada uma das tabelas existentes, os campos disponíveis e os relacionamentos existente entre as tabelas. Será detalhado também o processo de recuperação dos CNPJs que estão censurados na base de dados de operações indiretas do BNDES. Por fim, apresentados algumas estatísticas descritas advindas do cruzamento de informações entre o BNDES e Receita Federal.

**Introdução**

A necessidade de maior *accountability* tem estimulado as iniciativas de bases de dados abertos contendo microdados de interesse público. Entre essas iniciativas, podemos citar portal o Cadastro Nacional de Pessoas Jurídicas da Receita Federal e o Portal Transparência do BNDES, contendo microdados de empresas. A unidade observacional dos dados disponibilizados é a operação de crédito, contendo o CNPJ censurado das empresas e as condições da operação de crédito (taxa de juros, prazo, amortização, valor do contrato). A unidade observacional das informações da Receita Federal é a matriz empresa, a empresa final e o sócio da empresa, dependendo da tabela relacional. Nosso na concatenação desses dados é desenvolver uma proxy para o risco de crédito dessas operações, relacionando esta medida de risco com as operações do BNDES.

A Receita Federal disponibiliza 03 tabelas da microdados: Empresa, Filiais e Sócios (Tabela 1). A tabela de empresas (RF\_EMPRESA) possui aproximadamente 40 milhões registros correspondendo a matriz das empresas, os dados de filiais (RF\_FILIAIS) possuem aproximadamente 42 milhões de observações correspondendo as plantas locais, e a tabela de sócios (RF\_SOCIOS) possuem aproximadamente 18 milhões de observações, correspondendo a estrutura societárias das empresas. As informações de empresas (RF\_EMPRESA), filiais (RF\_FILIAIS) e sócios (RF\_SOCIOS) possuem como chave comum o CNPJ raiz da empresa (08 primeiros dígitos do CNPJ).

A tabela de operações indiretas (BNDES\_INDIR\_AUTO) do BNDES possui em torno de 269 mil observações, correspondendo a todas as operações de crédito indiretas no período de 2017 a 2022 (Tabela 1). A tabela de operações não-automáticas (BNDES\_NAOAUTO) possuem aproximadamente 20 mil observações, onde o CNPJ não se encontra censurado. A tabela de operações da administração pública (BNDES\_ADM\_PUB), tais como prefeituras, e autarquias, possuindo aproximadamente 4 mil observações. As informações do BNDES possuem como chave comum o CNPJ raiz da empresa (08 primeiros dígitos do CNPJ), o qual pode ser utilizado também para concatenar informações da Receita Federal.

Com a introdução das diretrizes de conformidade de Basel II e Basel III, e a necessidade de mais acurácia nas mensurações de risco de crédito, os modelos de análise de sobrevivência ganharam maior importância ao longo dos anos (Dirick, 2012). Historicamente, a análise de sobrevivência tem sido utilizada principalmente em medicina e engenharia, onde o tempo de duração até ocorrência de um evento é analisada.

Narain (1992) foi a primeira pessoa que sugeriu a utilização de análise de sobrevivência no contexto de risco de crédito. A vantagem da utilização

Time to Default in Credit Scoring Using Survival Analysis: A benchmark study. Journal of The Operational Research Society, v. 68, p. 652-665.

SHAP e LIME: Uma Avaliação do Poder Discriminativo no Risco de Crédito

Esses modelos são geralmente chamados de “caixas-pretas”, o que significa que você conhece as entradas e saídas, mas há poucas maneiras de entender o que está acontecendo sob o capô. Como resposta a isso, vimos vários modelos diferentes de IA explicada florescerem nos últimos anos, com o objetivo de permitir que o usuário veja por que a caixa preta deu uma determinada saída. Nesse contexto, avaliamos dois modelos eXplainable AI (XAI) muito populares em sua capacidade de discriminar observações em grupos, por meio da aplicação de modelagem não supervisionada e preditiva aos pesos que esses modelos XAI atribuem a recursos localmente. A avaliação é realizada em dados reais de Pequenas e Médias Empresas, obtidos de repositórios oficiais italianos, e pode servir de base para o emprego de tais modelos XAI para extração de características pós-processamento.

1. Introdução

A estimativa da probabilidade de inadimplência (PD) é uma questão que os bancos e outras instituições financeiras têm enfrentado desde o surgimento do crédito. Sistemas e metodologias evoluíram junto com o conhecimento e a tecnologia, mas não foi até recentemente que os incríveis avanços dados em TI deram uma verdadeira sacudida na forma como ela era executada pela indústria. A princípio, as instituições incumbentes resistiram à aplicação de novos paradigmas, o que favoreceu o surgimento de um número crescente de startups de fintech cujo objetivo é fornecer uma estimativa da solvabilidade de pessoas e empresas e fazer com que essa estimativa seja a mais alta fidelidade possível.

Para poder fornecer essa estimativa, é claro que essas empresas alavancam novas e diversas fontes de dados, aproveitam as inovações na estrutura regulatória relativa a dados financeiros (por exemplo, PSD2 europeu (Comissão Europeia, (2015)) e exploram o poder preditivo muito maior que alguns dos algoritmos recém-implementados oferecem respeito aos métodos tradicionais. No entanto, o aumento no poder de previsão dos novos algoritmos prejudica a explicabilidade, uma vez que os modelos agora são tão complexos que é quase impossível estabelecer vínculos claros entre os funcionamento do modelo e do output dado, o que seguramente representa um problema e dificulta a sua difusão, para além de levantar uma série de questões éticas e regulamentares, que começam a ser abordadas (ver, por exemplo, Comissão Europeia (2020)).

Para resolver esse trade-off, o conceito de eXplainable AI (XAI) surgiu introduzindo um conjunto de técnicas de aprendizado de máquina (ML) que produzem modelos que oferecem um trade-off aceitável entre explicabilidade e utilidade preditiva e permite que os humanos entendam, confiem e gerenciar as gerações emergentes de modelos de IA. Entre as técnicas emergentes, dois frameworks têm sido amplamente reconhecidos como o estado da arte em eXplainable AI e são eles:

• o framework Lime, introduzido por (Ribeiro et al., 2016)

• Valores SHAP, introduzidos por (Lundberg e Lee, 2017).

Em finanças, a interpretabilidade é especialmente importante porque a confiança do modelo nas características corretas deve ser garantida; ainda, não existem muitos estudos com foco na aplicação do XAI neste contexto específico (Bussmann, 2020). propõem um modelo XAI baseado em valores de Shapley aplicado no contexto de decisões de empréstimo para PME que buscam financiamento por meio de plataformas P2P, enquanto a pesquisa de (Ariza-Garzón et al., 2020) visa avaliar a capacidade preditiva de vários modelos de ML no contexto de pontuação de crédito das plataformas de empréstimo P2P, em seguida aplicando o método Shapley para fornecer explicabilidade à previsão. O precedente mais interessante talvez seja a pesquisa de (HadjiMisheva et al., 2021), onde os autores exploram a utilidade das estruturas SHAP e Lime no contexto do gerenciamento de risco de crédito, delineando os obstáculos práticos na aplicação dessas técnicas a vários tipos diferentes de algoritmos de ML, bem como propor soluções para os desafios enfrentados.

Nosso estudo visa comparar as estruturas SHAP e LIME avaliando sua capacidade de definir grupos distintos de observações, empregando os pesos atribuídos às feições por meio de seu algoritmo de interpretabilidade local como espaço de entrada para abordagens não supervisionadas e supervisionadas. Fazemos isso construindo nossa abordagem em um dos algoritmos de aprendizado supervisionado de melhor desempenho, porém complexo, XGBoost (Chen e Guestrin, 2016), empregado para prever a probabilidade de inadimplência de pequenas e médias empresas italianas.

1. https://dadosabertos.bcb.gov.br/ [↑](#footnote-ref-1)
2. <http://www.bndes.gov.br/wps/portal/site/home/transparencia/centraldedownloads> [↑](#footnote-ref-2)