RESUMO EXECUTIVO

Atributos dos donos de negócios e importância na formalização CNPJ

Sistema SEBRAE

Brasília - DF, 18 de Maio de 2022

Serviço Brasileiro de Apoio às Micro e Pequenas Empresas – SEBRAE
Unidade de Gestão Estratégica e Inteligência
SGAS 605 – Conjunto A – Asa Sul – Brasília/DF – CEP 70200-904
Tel.: 55 61 3348-7180
Site: `https://www.sebrae.com.br/`

CONSELHO DELIBERATIVO NACIONAL
Presidente
José Roberto Tadros

DIRETORIA EXECUTIVA

Diretor-Presidente
Carlos do Carmo Andrade Melles

Diretor Técnico
Bruno Quick Lourenço de Lima

Diretor de Administração e Finanças
Eduardo Diogo

Gerente da Unidade de Gestão Estratégica e Inteligência
Adriane Ricieri Brito
Gerente Adjunto da Unidade de Gestão Estratégica e Inteligência
Fausto Ricardo Keske Cassemiro

Coordenador do Núcleo de Pesquisa e Gestão do Conhecimento
Kennyston Costa Lago

Equipe Técnica
Tomaz Back Carrijo
Felipe Marcel Neves
Juliana Borges Vaz

A Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) é realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e visa evidenciar as mudanças na força de trabalho e outras informações relevantes para o desenvolvimento socioeconômico do país. O presente trabalho possui como objetivo analisar os dados da PNAD Contínua do 4° trimestre de 2021, verificando as melhores possibilidades para gerar um modelo de aprendizado de máquina que compreenda características que influenciam a formalização dos estabelecimentos. Na análise utilizou-se o peso amostral ajustado por pós-estratificação fornecido pelo IBGE para que as estimativas totais correspondam com as estimativas oficiais disponibilizadas. A população considerada no estudo são os indivíduos classificados como Donos de Negócios (Empregadores e Conta Própria), com a variável resposta sendo a presença de registro do negócio/empresa no Cadastro Nacional da Pessoa Jurídica - CNPJ.

Para compreender as informações da PNAD Contínua e usá-las para modelagem, ocorreu uma análise exploratória dos dados, incluindo análises estatísticas de associação e correlação entre as variáveis. Na análise das principais características que aumentam a probabilidade dos Donos de Negócios estarem formalizados, a pesquisa foi realizada com variáveis disponíveis relacionadas ao negócio/empresa, e algumas variáveis pessoais, entre elas escolaridade, raça-cor, faixa-etária e gênero. A Figura 1 mostra o comportamento da variável que representa a posição na ocupação do indivíduo no negócio/empresa em que é associado, e que ajuda na caracterização dos Donos de Negócios, que englobam o indivíduo na posição de ”Empregador”ou ”Conta Própria”. Esse público alvo representa um percentual de 31,29% nessa classificação, um total estimado de 29.700.877 indivíduos no trimestre de análise.

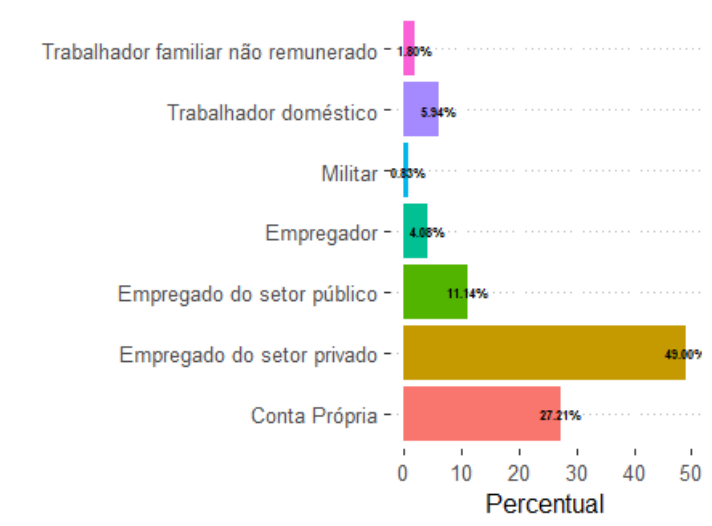


Figura 1 – **Distribuição da variável - Posição na Ocupação no negócio/empresa.**
Fonte: Elaborado pelos autores.

A distribuição da variável de interesse é representada na Figura **??**. Observa-se que 67,92% (total estimado de 20.171.397) dos indivíduos pertencentes a classificação de Donos de Negócios não possuem o registro CNPJ, podendo ser potenciais empresários que não estão formalizados. A análise de dados possibilitou um filtro de variáveis para se utilizar na modelagem, conforme alguns fatores, tais

como a proporção da presença ou ausência de nulos, valor de correlação, entre outros. A base final gerada, consistiu-se em 12 variáveis preditoras (13 no caso da variação com UFs), 1 variável resposta (se o empreendimento tem CNPJ ou não) e 1 variável acessória (peso amostral).

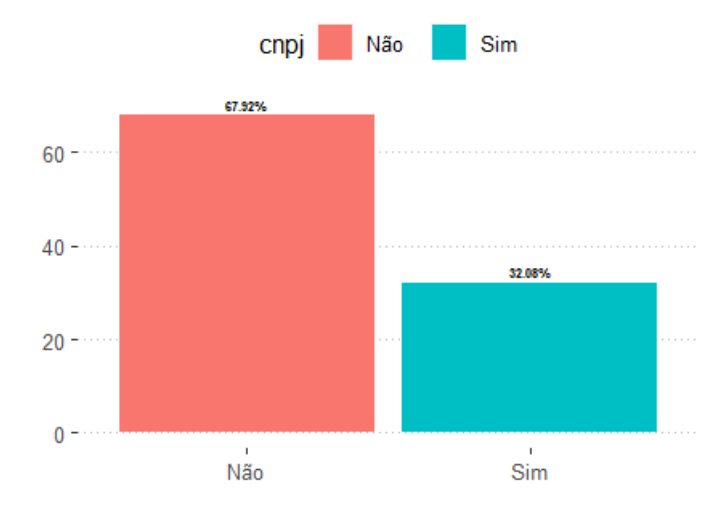
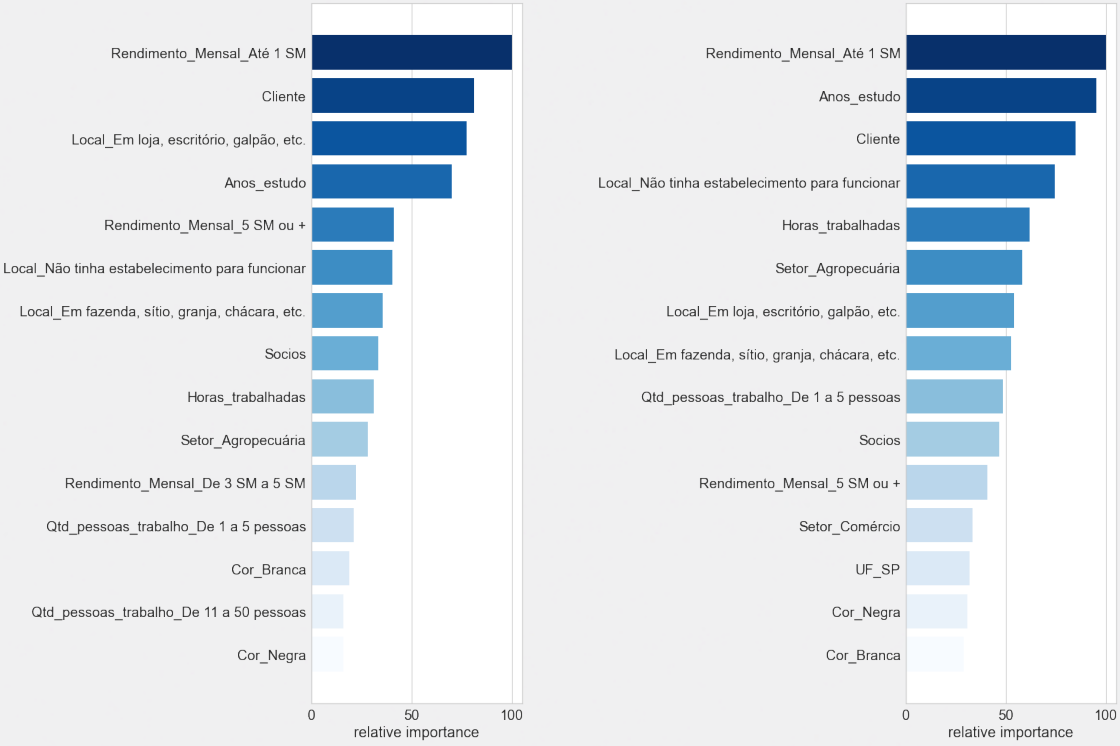


Figura 2 – **Distribuição da variável resposta - Negócio/Empresa com registro CNPJ?**
Fonte: Elaborado pelos autores.

Uma vez que as variáveis finais foram definidas com o auxílio da análise exploratória de dados, ocorreu a preparação dos dados para a modelagem, utilizando validações clássicas de modelos de aprendizagem de máquina. Os modelos foram desenvolvidos considerando duas variações, com a ausência ou presença da variável Unidade Federativa. A preparação dos dados para a modelagem resultou em um aumento do número das variáveis preditoras de 12 para 33, no caso da variação sem UFs, e 13 para 60 para a variação com UFs. De modo a escolher o modelo final com melhor desempenho foram usados quatro modelos-base iniciais: Árvore de Decisão, Regressão Logística, SVM (*Support Vector Machine*) e Catboost (*Gradient Boosting*). Dentre as diversas métricas de avaliação do desempenho do modelo, utiliza-se como principal a AUC (*Area Under the ROC Curve*), que mensura a probabilidade de previsão da variável resposta para avaliar e comparar os modelos com mais precisão, variando de 0 a 1. Um modelo cujas previsões estão 100% erradas tem uma AUC de 0; aquele cujas previsões estão 100% corretas tem uma AUC de 1. Após a escolha do algoritmo conforme o melhor desempenho, ocorreu a otimização de seus parâmetros, e isto foi usado para criar o modelo final onde os resultados foram extraídos.

O algoritmo Catboost obteve melhor desempenho dentre os modelos-base, sendo o método escolhido para criação do modelo final. O modelo obteve uma AUC em média de 0.88 para ambas as variações (sem UFs e com UFs). Como observado, os resultados ficaram bem similares, com um pequeno aumento no valor das métricas para quando é considerada a variável de UFs. Verifica-se também os valores de *Feature importance*, um método para se obter a importância da contribuição das variáveis no modelo, com as importâncias relativas de cada variável plotadas graficamente. Conforme a análise de *feature importance* (Figura 3), estão entre as principais variáveis mais importantes para a predição:



(a) Variação sem inclusão de UFs

(b) Variação com inclusão de UFs

Figura 3 – ***Feature importance* do modelo final das 15 principais variáveis.**
Fonte: Elaborado pelos autores.

rendimento mensal de até um salário mínimo (SM), anos de estudo (escolaridade), cliente (empregador ou conta própria), local de funcionamento do negócio/empresa (principalmente em loja, escritório e galpão, fazenda, chácara ou mesmo na ausência de um estabelecimento para funcionar), número de horas trabalhadas por semana, presença ou ausência de sócios, e rendimento mensal de 5 salários mínimos (SM) ou mais.

De modo a interpretar as importâncias das variáveis nos resultados do modelo de um modo mais abrangente, foi utilizado os valores de SHAP (*SHAP values*). SHAP é um método baseado na teoria dos jogos para explicar previsões individuais, usado para aumentar a transparência e a interpretabilidade dos modelos de aprendizado de máquina. SHAP além de evidenciar a importância das variáveis tal como *feature importance*, também indica se determinado valor de uma variável teve um impacto positivo ou negativo na probabilidade do *output* do modelo, em nosso caso, a probabilidade do empreendimento ter CNPJ ou não. A Tabela 1 traz os fatores mais importantes para a formalização do negócio conforme as análises de SHAP, apontando um resumos da influência das 8 variáveis mais importantes.

A visão do modelo por UF (Figura 4) apontou em destaque que se o empreendimento do indivíduo tiver residência no Maranhão, Amazonas, Pará e Pernambuco, a probabilidade do negócio/empresa possuir CNPJ é menor, sinalizando uma maior necessidade de suporte para esses estados no quesito do empreendedorismo. Diferentemente, a origem no estado de São Paulo, Paraná e Santa Catarina au-

Tabela 1 – **Resumo de influência das 8 variáveis mais importantes (em ordem decrescente) através da análise de valores de SHAP - variação sem inclusão de UFs. Influência pode ser positiva (+) ou negativa (-).**

Variável	Influência	Descrição
Rendimento mensal de até 1 SM	-	Se o rendimento é de até um salário mínimo, menor é a probabilidade do negócio ter CNPJ
Anos de estudo	+	Quanto maior o número de anos de estudo do indivíduo, maior é a probabilidade do negócio ter CNPJ
Local em loja escritório, galpão, etc	+	Quando o local é em loja, escritório e galpão, maior é a probabilidade do negócio ser CNPJ
Cliente empregador	+	Se o cliente é empregador, maior é a probabilidade do negócio ter CNPJ
Horas trabalhadas	+	Quanto maior o número de horas trabalhadas na semana, maior é a probabilidade do negócio ter CNPJ
Setor agropecuária	-	Quando o setor é de agropecuária, menor é a probabilidade do negócio ter CNPJ
Sócios	+	Quando existe a presença de sócios, maior é a probabilidade do negócio ser CNPJ

Fonte: Resultados originais da pesquisa.

menta a probabilidade do negócio/empresa possuir CNPJ. A influência pode ser existente com maior probabilidade de CNPJ (1), menor (-1), e não existente (0). O rank de valores de SHAP normalizado, é o valor da influência de SHAP dividido pela sua ordem de importância.

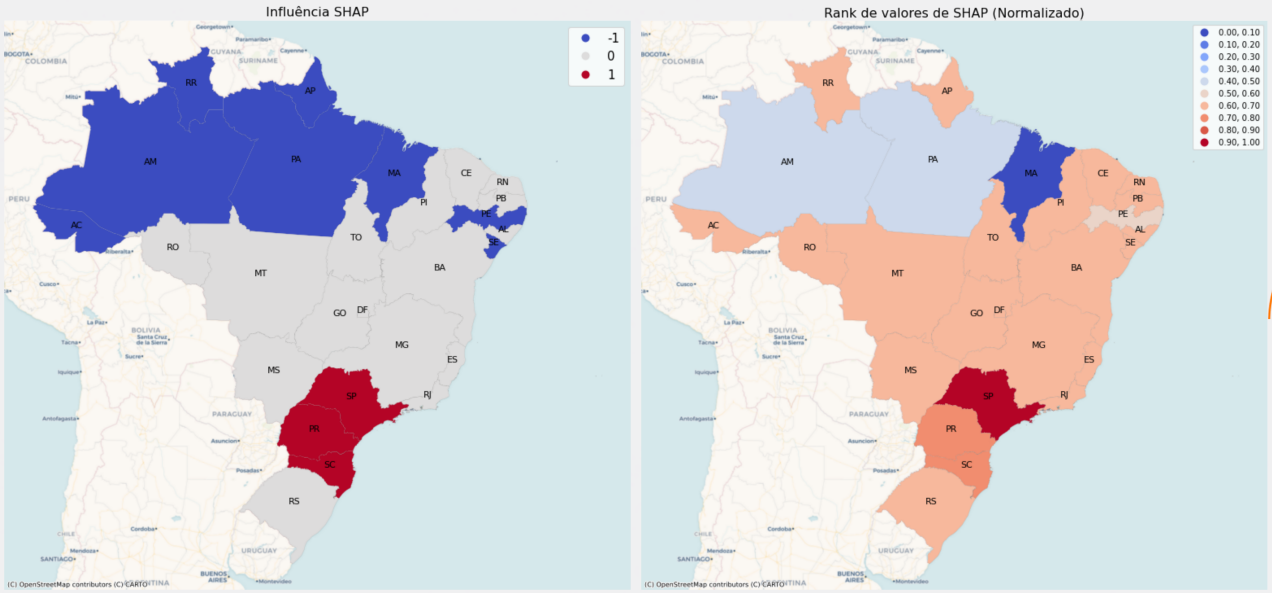


Figura 4 – **Influência do valor de SHAP e rank de valores normalizados.**
Fonte: Elaborado pelos autores.

O bom desempenho do modelo conforme as métricas consideradas, apenas reforçou que os resultados conseguem ser robustos o suficiente para descrever a importância que determinadas variáveis

possuem para caracterizar os fatores que levam um empreendimento estar formalizado. Resumindo: dentre os fatores mais importantes para a formalização do negócio estão o rendimento do indivíduo, sua escolaridade (anos de estudo), o local onde o negócio funciona (escritório, galpão, fazenda, sítio, ...), se o cliente era empregador ou conta própria, seu segmento de atuação, presença ou ausência de sócios, horas efetivas de trabalho na semana, e sua UF de origem.


SEBRAE
50+50