Atributos dos donos de negócios e importância na formalização CNPJ

Sistema SEBRAE

Brasília - DF, 18 de Maio de 2022

SEBRAE 50+50

Todos os direitos reservados.

A reprodução não autorizada desta publicação, no todo ou em parte, constitui violação aos direitos autorais (Lei nº 9.610).

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

RESUMO

Indivíduos classificados como Donos de Negócios, ”Empregador”e ”Conta Própria”, são um grupo de empreendedores que podem ou não estarem devidamente formalizados, ou seja, podem ter ou não CNPJ registrado na Receita Federal do Brasil. A partir dos microdados da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD/IBGE), utilizando modelo prescritivo de classificação de aprendizagem de máquina, o presente estudo realizou análise de variáveis relacionadas ao negócio/empresa visando conhecer as principais características que aumentem as probabilidades dos donos de negócio estarem formalizados. No último trimestre de 2021, apenas 32,08% dos donos de negócio possuíam registro CNPJ. Estima-se que existam 20 milhões de empresários não formalizados no Brasil. Esta informação demonstra a importância do presente estudo. Em geral, 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 Unidade da Federação.

1 INTRODUÇÃO

A Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) é realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) visando acompanhar as flutuações da força de trabalho e outras informações relevantes para o desenvolvimento socioeconômico do país. O presente trabalho possui o objetivo de analisar os dados da PNAD Contínua do 4° trimestre de 2021, verificando as melhores possibilidades para gerar um modelo de aprendizagem de máquina visando compreender características que influenciam a formalização dos estabelecimentos. O quadro 1 apresenta as variáveis iniciais consideradas no estudo, as que estão em destaque foram utilizadas no modelo após todos os ajustes realizados e explicados na metodologia.

A população considerada no estudo são os indivíduos classificados como Donos de Negócios, com a variável resposta sendo a presença de CNPJ. Donos de Negócios, compostos por indivíduos classificados como ”Empregador”e ”Conta Própria”, correspondem a cerca de 31% dos entrevistados na PNAD Contínua, totalizando um número estimado de cerca de 29.700.877 indivíduos (Figura 1).

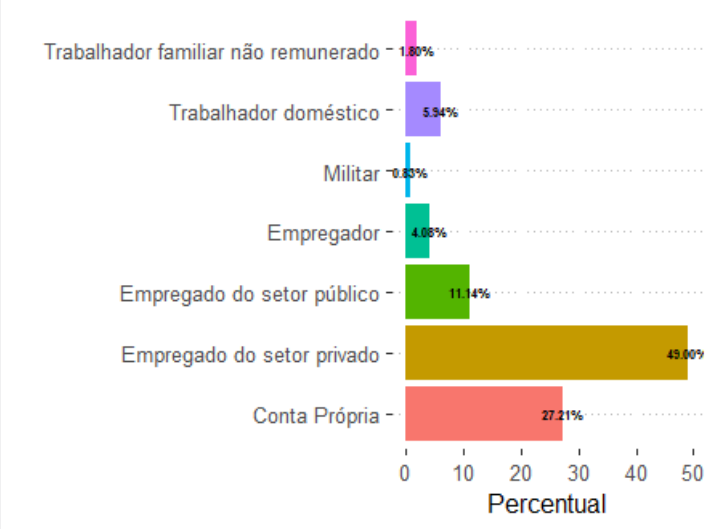

Figura 1 – **Distribuição da variável - Posição na Ocupação no negócio/empresa.**
Fonte: Elaborado pelos autores.

Considerando os donos de negócio e sua formalização, observa-se que 67,92% (total estimado de 20.171.397) dos indivíduos não possuem o registro CNPJ (Figura 2). Estas informações demonstram que o tema deste estudo aborda uma parcela considerável da população.

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.

2 METODOLOGIA

A análise de dados e os modelos criados foram desenvolvidos em ambiente R v4.1.1 (R CORE TEAM, 2021) e Python v3.9.12 (VAN ROSSUM; DRAKE JR, 1995). Os principais pacotes usados estão

Quadro 1 – **Variáveis iniciais consideradas no estudo.**

Código - PNADC	Nome da Variável	Descrição
Ano e Trimestre	ano_trimestre	Ano/Trimestre
UF	regiao	Região
UF	**UF**	**Unidade Federativa**
V1028	**peso_amostral**	**Peso Amostral**
V2007	**gênero**	**Homem ou Mulher**
V2009	**faixa etária**	**faixa etária dos indivíduos**
V2010	**raça**	**raça negra, branca, outras**
VD3005	**anos de estudo (esco-laridade)**	**Anos de estudo**
V4009	qtde_trabalhos	N° de trabalhos que o indivíduo tinha na semana de referência
VD4010	**setor**	**Setores de atuação**
V4012	**cliente**	**Empregador (1) ou conta própria (0)**
V4013	Nome_secao	Seção - CNAE
V4013	Nome_divisao	Divisão - CNAE
V4013	Nome_classe	Classe - CNAE
V4016	empregados	N° de empregados trabalhavam nesse negócio/em-presa
V4017	**sócios**	**Na semana de referência, tinha pelo menos um sócio que trabalhava nesse negócio/empresa?**
V40171	qtde_sócios	N° e sócios do negócio/empresa
V4018	**pessoas**	**N° de pessoas trabalhavam no negócio/empresa**
V4019	**cnpj**	**Negócio/empresa com registro CNPJ**
V4020	**local_1**	**Tipo de local funcionava a negócio/empresa**
V4021	local_2	Exercia normalmente o trabalho em estabelecimento desse negócio/empresa ?
V4022	local_3	Então, onde exercia normalmente esse trabalho?
V40331	recb_din_mensal	Recebimento/retirada em dinheiro no mês
V403311	**rendimento**	**Faixa do rendimento/retirada em dinheiro**
V403312	rend_bruto	Rendimento bruto em reais
V403311	rendimento_mes_ref	Número da faixa do rendimento/retirada em dinheiro
V403412	rend_bruto_mes_ref	Rendimento bruto em reais no mês de referência
V4039C	**horas_trabalhadas**	**Horas efetivas de trabalho na semana**
V4040	**tempo_trabalho**	**Tempo de trabalho**

descritos nas sessões especificas onde foram utilizados.

2.1 Fontes de dados

A importação dos microdados da PNAD Contínua foi realizada diretamente pelo site do IBGE (IBGE, 2021), e a granularidade temporal dos dados foi trimestral, mais especificamente, foi considerado o 4° trimestre de 2021. Os dados sobre os níveis hierárquicos (seção, divisão e classe) da Classificação Nacional de Atividades Econômicas (CNAE) foram obtidos também através do site do IBGE.

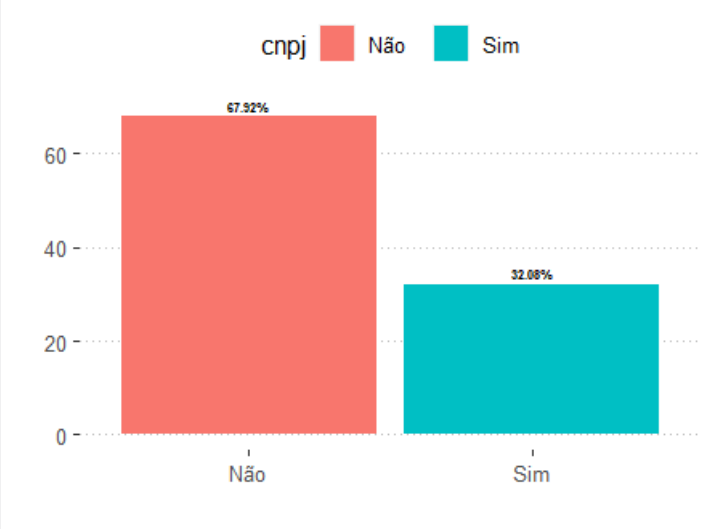

Figura 2 – **Distribuição da variável resposta - Negócio/Empresa com registro CNPJ?**
Fonte: Elaborado pelos autores.

2.2 Tratamento e Análise dos Dados

O público-alvo do estudo são os Donos de Negócios, caracterizados pela seleção das categorias ”Conta Própria”e ”Empregador”da variável V4012, que representa a posição de ocupação daquele indivíduo no negócio/empresa. A variável de resposta corresponde se o negócio/empresa era registrado no Cadastro Nacional da Pessoa Jurídica — CNPJ. A seleção de variáveis preditoras foi realizada visando aquelas variáveis relacionadas ao negócio, dentre outras pessoais como gênero, raça, faixa-etária e anos de estudo. Também foi incluído três variáveis derivadas do código principal da atividade do negócio/empresa (CNAE). Existe ainda uma variável apenas para identificação da base, a variável temporal ”Ano_trimestre”, que não foi considerada no modelo.

Alguns ajustes foram realizados antes da análise descritiva dos dados, são eles:

- Filtro da idade do indivíduo compatível a definição de População Economicamente Ativa (PEA),
- Retirada de casos de não resposta da variável utilizada na caracterização do Dono de Negócio,
- Retirada de variáveis de mesma informação, mudando apenas a granularidade,
- Retirada de variáveis de preditoras com a porcentagem do número de nulos maior que 25% e
- Investigação sobre variáveis muito relacionadas entre si e com baixa variância com o intuito de retirar as mesmas do modelo.

Após o ajuste da base de dados, houve uma análise gráfica do comportamento das variáveis com o apoio do pacote *ggplot2* (WICKHAM, 2016), e em seguida uma análise de correlação entre as variáveis. Para a construção dos gráficos e da investigação, houve a incorporação do instrumento de

plano amostral empregado para estimar as quantidades populacionais disponibilizado na própria PNAD Contínua — variável V1028. O pacote *survey* (LUMLEY, 2020) dispõe de funções características para elaborar gráficos que absorvam os pesos amostrais das observações. Para a modelagem, a variável peso amostral foi incluída de forma acessória na hora do ajuste dos modelos, e não como variável preditora, detalhes de como isto foi feito estão descritos na seção "Modelos base".

No caso das variáveis categóricas, realizou-se primeiramente o teste de hipótese Qui-quadrado de independência com a variável resposta e todas as variáveis independentes. O preceito básico fundamental do teste de independência de variáveis que recorre à tabela de contingência é a comparação de proporções, ou seja, possíveis discordâncias entre as frequências observadas e esperadas para um certo episódio. O nível de significância adotado foi de 5% e as hipóteses do teste são:

H_0) As variáveis são independentes.
H_1) As variáveis não são independentes.

Para medir a intensidade dessas associações, utilizamos as seguintes medidas de associação: Coeficiente de Contingência e o V de Crámer. Ambas variam de 0 (ausência de associação) a 1 (associação muito forte). Dentre as variáveis numéricas, aplicou-se o procedimento estatístico de correlação de Spearman que é uma medida não paramétrica de correlação de postos. Para mais informações sobre os procedimentos adotados nessa parte do estudo, pode se consultar o conteúdo do livro "Estatística não-Paramétrica Para Ciências do Comportamento"(SIEGEL; CASTELLAN, s.d.).

2.3 Preparação dos dados para modelagem

A base de dados final foi separada em variáveis preditoras (X) e variável resposta (y). Foram utilizadas duas formas clássicas de validação de modelos de *machine learning*, a divisão dos dados entre treino (80%) e teste (20%), e a validação cruzada k-fold (5 *folds*). Na divisão dos dados entre treino e teste (também conhecida como *hold-out*), o conjunto de treino é usado para construir o algoritmo de previsão. No conjunto de teste, é aplicado o algoritmo construído pelos dados de treino para podermos estimar seu desempenho em dados não vistos. Diferentemente, o método de validação cruzada k-fold consiste em dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho (para a nossa modelagem, foi escolhido 5) e, a partir daí, um subconjunto é utilizado para teste e os k-1 restantes são utilizados para treino. Este processo é realizado k vezes alternando de forma circular o subconjunto de teste, por fim, é calculado a média dos resultados para validação.

Antes do uso dos modelos, ocorreu imputação de valores faltantes e em seguida *encoding* de variáveis categóricas (i.e. transformação de variáveis categóricas em numéricas). A imputação de variáveis numéricas foi baseado na mediana, e na moda para variáveis categóricas. Foram consideradas diversas formas de *encoding* de variáveis categóricas, tais como *One-hot encoding*, *James Stein encoder* e WOE (*Weight of Evidence*) *encoder* (MCGINNIS et al., 2018). Devido a seu desempenho superior nos modelos base, além de facilidade na interpretação dos resultados, o *One-hot encoding* foi o método escolhido para o modelo final. *One hot encoding* consiste em substituir a variável categórica por uma

combinação de variáveis binárias que assumem o valor 0 ou 1 (variáveis *dummy*), para indicar se uma determinada categoria está presente em uma observação.

Para modelos que exigem escalonamento de variáveis *per se* para melhor funcionamento (tais como o *Support Vector Machine*), isto foi introduzido na *pipeline* do modelo através das funções *StandardScaler* e *RobustScaler* presentes no pacote scikit-learn v1.0 (BUITINCK et al., 2013), as demais funções usadas nesta seção também foram criadas ou usaram recursos deste pacote. Em relação à base de dados final, os modelos foram desenvolvidos considerando duas variações, com a ausência ou presença da variável unidade federativa (com UFs e sem UFs).

2.4 Modelos base

De modo a escolher o modelo final (com melhor desempenho), para aplicar a otimização de hiperparâmetros, foram usados quatro modelos: Árvore de Decisão, Regressão Logística, SVM (*Support Vector Machine* ou Máquina de Vetores de Suporte), e Catboost (*Gradient Boosting*).

A Árvore de Decisão é um modelo composto por um conjunto de nós de decisão (quando um nó é dividido em sub-nós adicionais) e nós de término (nó final do caminho inferencial, nós não divididos), organizados hierarquicamente. A Árvore de Decisão foi escolhida por ser um "*glass model*", isto significa que a capacidade de interpretação, principalmente através das visualizações em árvore, características deste modelo, ajudam bastante a entender o peso das variáveis e como elas se relacionam para gerar a predição. Porém, apesar da fácil interpretação, modelos de árvores de decisão sofrem por não escalarem muito bem para dados não vistos, além de seu desempenho não tão alto comparado a outros modelos de classificação.

A regressão logística é um modelo que usa uma função sigmóide para converter a previsão bruta de um modelo linear em um valor entre 0 e 1. É um modelo simples, tradicional e eficaz, considerado o modelo padrão de classificação para a maioria das ciências, por isso sua escolha. A principal limitação da regressão logística é a suposição de linearidade entre a variável dependente e as variáveis independentes.

Diferentemente, SVM é um algoritmo que visa maximizar a margem entre classes positivas e negativas mapeando vetores de dados de entrada para um espaço dimensional superior. Dessa maneira, ela tenta encontrar uma separação robusta entre as classes. É um modelo que apresenta desempenho alto em geral comparado a diversos outros algoritmos de classificação, porém possui dificuldade para a atuação em *datasets* maiores.

Por fim, também foi testado um algoritmo de *gradient boosting*, o Catboost. A ideia principal por trás desse algoritmo é construção de modelos (no caso de Catboost, árvores de decisão) sequencialmente, onde os modelos subsequentes tentam reduzir os erros do modelo anterior, construindo um novo modelo sobre os erros ou resíduos das iterações anteriores. O Catboost *per se* é um framework de *gradient boosting*, e tem como grandes vantagens um ótimo desempenho comparado a outros algoritmos, e o fato de ter capacidade inata para lidar com variáveis categóricas. Uma das únicas desvantagens consideráveis do Catboost é uma quantidade grande de parâmetros para otimização.

Os pacotes scikit-learn v1.0 (BUITINCK et al., 2013) e catboost v1.0.5 (DOROGUSH; ERSHOV; GULIN, 2018) foram utilizados para aplicação dos modelos. Outra característica deste trabalho foi o uso de pesos amostrais, o scikit-learn durante o *fit* dos modelos possui uma propriedade chamada *sample weight*, usada para adicionar pesos amostrais, dessa forma os pesos existentes nos dados foram incorporados pelos modelos. Para mais informações sobre os algoritmos, consulte (GÉRON, 2019) e (DOROGUSH; ERSHOV; GULIN, 2018).

2.5 Avaliações dos modelos-base e escolha de hiperparâmetros

Escolhemos algumas métricas apropriadas para classificar o desempenho dos modelos, dentre elas AUC (*Area Under the ROC Curve*) e *Balanced Accuracy* como as principais, *Precision* e *Recall* como complementares. AUC utiliza probabilidades de previsão da variável resposta, com base nisso, podemos avaliar e comparar os modelos com mais precisão, pode ser considerada a métrica mais importante em nosso estudo. *Balanced accuracy* também consegue avaliar o desempenho dos modelos, mas apenas se o modelo classifica corretamente ou não a variável resposta, sem levar em contas as probabilidades, o mesmo ocorre com *Precision* e *Recall*, com a diferença que estas duas últimas analisam características de como o modelo está classificando os dados e não seu desempenho de modo unificado. Todas as métricas consideradas no estudo foram aplicadas através do pacote scikit-learn (BUITINCK et al., 2013) e considerando pesos amostrais. Para mais informações sobre as métricas, consulte (GÉRON, 2019) As métricas usadas e o *framework* de otimização de hiperparâmetros estão descritas a seguir.

2.5.1 AUC (*Area Under the ROC Curve*)

Para uma melhor compreensão do AUC, deve-se descrever brevemente o que seria uma curva ROC. ROC é um gráfico que mostra o desempenho de um modelo de classificação em todos os limites de classificação. Esta curva traça dois parâmetros:

- Taxa de Verdadeiros Positivos (*True Positive Rate* - TPR)
- Taxa de Falsos Positivos (*False Positive Rate* - FPR)

TPR é o sinônimo de *Recall* sendo definido como:

$$\text{TPR} = \frac{TP}{TP + FN}$$

FPR é definido como:

$$\text{TPR} = \frac{FP}{FP + FN}$$

Onde TP (*True positives*), significa verdadeiros positivos, FP (*False positives*), falsos positivos, e FN (*False negatives*), falsos negativos. Uma curva ROC plota TPR vs. FPR em diferentes limites de classificação. A redução no limiar de classificação, acaba por classificar mais itens como positivos, aumentando assim tanto os falsos positivos quanto verdadeiros positivos. A figura (3) mostra uma curva ROC típica.

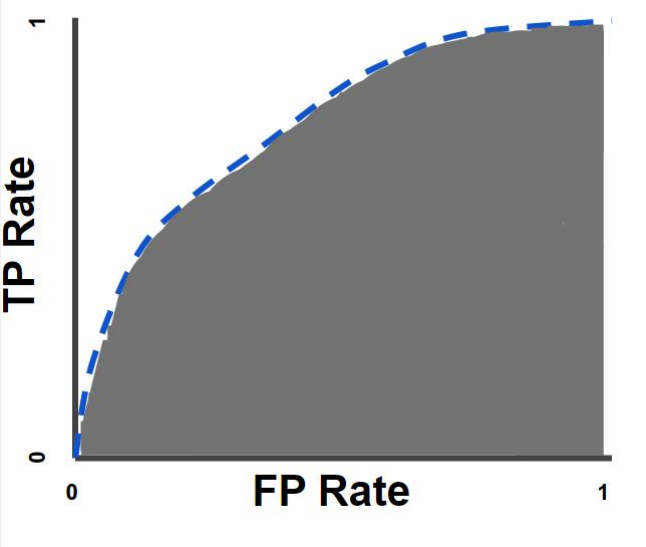

Figura 3 – **Taxa (*Rate*) de TP vs. FP em diferentes limiares de classificação.**
Fonte: Elaborado pelos autores.

Para computar os pontos em uma curva ROC, existe um algoritmo eficiente chamado AUC. AUC significa em português "Área sob a Curva ROC", ou seja, mede toda a área bidimensional abaixo de toda a curva ROC (considere o cálculo integral) de (0,0) a (1,1). AUC fornece uma medida agregada de desempenho em todos os limites de classificação possíveis. Uma maneira de interpretar AUC é como a probabilidade de que o modelo classifique um exemplo positivo aleatório mais alto do que um exemplo negativo aleatório. Na figura (4), AUC representa a probabilidade de que um exemplo aleatório positivo (verde) esteja posicionado à direita de um exemplo aleatório negativo (vermelho). AUC como métrica varia em valor 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. AUC é desejável, pois é invariante em escala. Ela mede quão bem as previsões são classificadas, em vez de seus valores absolutos.

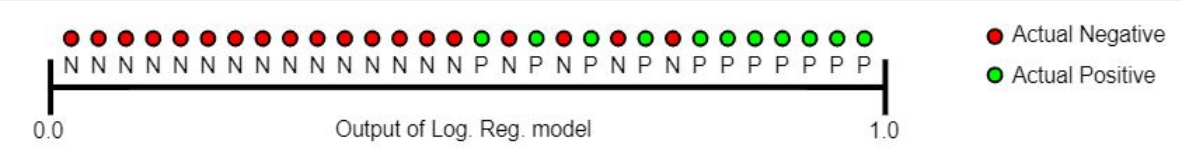

Figura 4 – **Previsões classificadas em ordem crescente de pontuação de uma regressão logística.**
Fonte: Elaborado pelos autores.

2.5.2 *Balanced accuracy*

Balanced accuracy calcula a métrica acurácia de forma balanceada, evitando estimativas de desempenho infladas em conjuntos de dados desequilibrados. É a média-macro dos scores de *recall* por classe ou, de modo equivalente, a acurácia bruta onde cada amostra é ponderada conforme a prevalên-

cia inversa de sua verdadeira classe. Assim, para conjuntos de dados balanceados, a pontuação é igual à acurácia. No caso binário, *balanced accuracy* é igual à média aritmética entre a taxa de verdadeiros positivos e taxa de verdadeiros negativos, ou a AUC com previsões binárias em vez de score, pode ser calculada como:

$$\texttt{Balanced-accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

Se o classificador tiver um desempenho igualmente bom em qualquer uma das classes, esse termo será reduzido à acurácia convencional (ou seja, o número de previsões corretas dividido pelo número total de previsões). Em contraste, se a acurácia convencional estiver acima do acaso apenas porque o classificador tira vantagem de um conjunto de teste desbalanceado, então *Balanced accuracy*, conforme apropriado, cairá para $\frac{1}{n\_classes}$. Os valores da métrica variam de 0 a 1.

2.5.3 *Precision*

Precision é a razão entre os verdadeiros positivos e todos os positivos (verdadeiros + falsos positivos), tenta responder a seguinte pergunta, que proporção de identificações positivas estava realmente correta? Os valores da métrica variam de 0 a 1, *Precision* é definido da seguinte forma:

$$\texttt{Precision} = \frac{TP}{TP+FP}$$

2.5.4 *Recall*

Recall é a razão entre os verdadeiros positivos por qualquer coisa que deveria ter sido prevista como positiva (verdadeiros positivos + falsos negativos), tenta responder a seguinte pergunta, que proporção de verdadeiros positivos foi identificada corretamente? Os valores da métrica variam de 0 a 1, *recall* é definido da seguinte forma:

$$\texttt{Recall} = \frac{TP}{TP+FN}$$

Tanto *precision* quanto o *recall* devem ser examinados em conjunto para melhor compreensão. *Precision* e *recall* estão frequentemente em antagonismo. Ou seja, melhorar *precision* normalmente reduz o *recall* e vice-versa.

2.5.5 Otimização de hiperparâmetros

Uma das tarefas determinantes na construção de modelos de aprendizado de máquina é a otimização de hiperparâmetros, isto reflete-se diretamente no desempenho do modelo. Para a realização deste feito foi utilizado o pacote Optuna v3.0 (AKIBA et al., 2019). Optuna é uma estrutura de otimização de hiperparâmetros de código aberto que automatiza a sua pesquisa. Por ser uma ferramenta que se concentra na simplicidade, flexibilidade e escalabilidade da implementação e que também possui integração com o scikit-learn, o pacote utilizado na criação dos nossos modelos, Optuna foi escolhida

como base para otimização. Após a escolha de um “campo” de busca de hiperparâmetros, o Optuna foi rodado para procurar a melhor combinação de hiperparâmetros em 25 tentativas. Os hiperparâmetros testados para o modelo final serão descritos nos resultados.

2.6 Construção teórica da interpretação dos resultados do modelo

Para a interpretação dos resultados do modelo, foram utilizados a *feature importance* das variáveis e os valores de SHAP (*SHAP values*). Utilizamos o pacote Yellowbrick v1.4 (BENGFORT et al., 2018) para visualização da *feature importance*, e o pacote SHAP v0.4 (LUNDBERG; LEE, 2017) para calcular os valores de SHAP e suas visualizações. *Feature importance* mede o aumento no erro de previsão global do modelo depois que os valores das variáveis são permutados, quebrando a relação delas com o resultado real, a *feature importance* pode ser ranqueada para cada variável e as importâncias relativas plotadas. SHAP (*SHapley Additive exPlanations*) por Lundberg and Lee (LUNDBERG; LEE, 2017) 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.

Através do pacote SHAP podem ser feitas diferentes visualizações acerca dos resultados, de modo a extrair o máximo de informações relevantes. Pode-se observar o impacto global dos valores de shap no modelo, sendo possível fazer inferências sobre o impacto médio das variáveis no modelo, e o quanto seus valores o impactaram. Para averiguar estas informações, dois gráficos interessantes são usados neste estudo, o SHAP *feature importance* e o SHAP *summary plot*. A ideia por trás do SHAP *feature importance* é de certo modo simples: variáveis com grandes valores absolutos de shap são importantes. Como queremos a importância global, é calculado a média dos valores absolutos de shap por variável. Em seguida, as variáveis são ordenadas de modo decrescente e o plot é criado. Deste modo, as variáveis no topo contribuem mais para o modelo do que as de baixo, portanto, têm alto poder preditivo.

Um plot que permite uma análise mais rica é SHAP *summary plot*, este gráfico combina a importância da variável com seus efeitos (figura 5). Cada ponto no gráfico de resumo é um valor shap para uma variável e uma instância. A posição no eixo y é determinada pela importância da variável no modelo e no eixo x pelo valor de shap. A cor representa o valor da variável, do baixo (azul) para o alto (vermelho). Os pontos sobrepostos são distorcidos na direção do eixo y, então temos uma noção da distribuição dos valores de shap por variável. Tal como SHAP *feature importance*, as variáveis são ordenadas de acordo com sua importância.

No exemplo da figura 5, um alto nível de teor de álcool (*alcohol*) tem um impacto alto e positivo na classificação. O alto nível de teor de álcool (*alcohol*) é representada pela cor vermelha e o impacto positivo é mostrado no eixo x. Por outro lado, um alto valor de acidez volátil (*volatile acidity*) tem um impacto relativamente alto, mas negativo na classificação, representado pela cor azul.

Para observar como o modelo se comportou em previsões individuais, pode-se visualizar os valores shap como ”forças”, através do *force plot* (figura 6). No *force plot*, cada valor individual das variáveis é uma força que aumenta ou diminui a previsão. A previsão começa a partir de uma linha de base (*baseline*). A *baseline* para os valores de shap é a média de todas as previsões. Ou seja, a previsão final (*output*

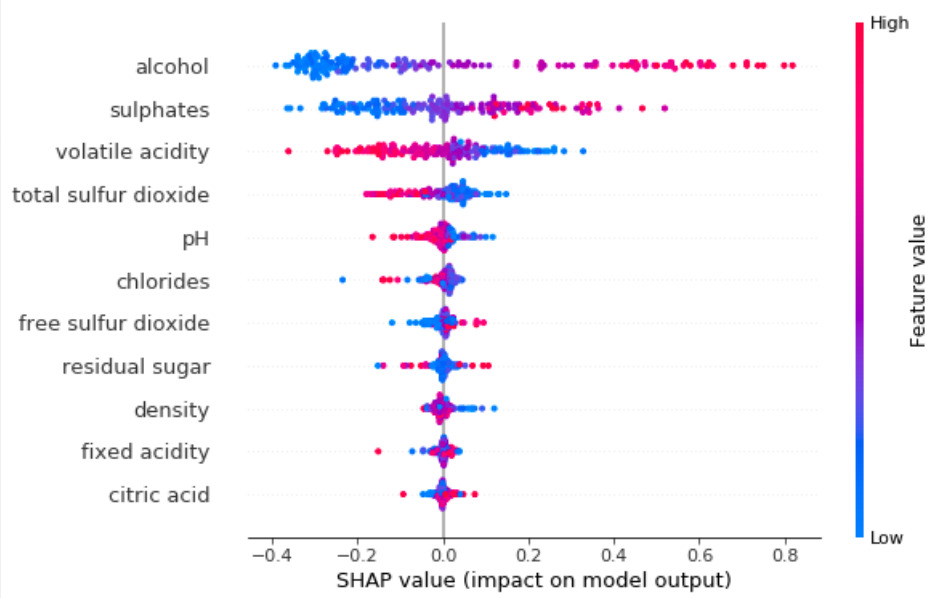

Figura 5 – **Exemplo de um SHAP *summary plot* para observação global.**
Fonte: Elaborado pelos autores.

value, em *log odds ratio*) é igual à previsão média (linha de base) mais os valores de shap de todas as variáveis. No gráfico, cada variável possui um valor de shap que empurra para aumentar (valor positivo, cor vermelha) ou diminuir (valor negativo, cor azul) a previsão. Setas maiores indicam uma força com maior impacto, essas forças se equilibram na previsão real da instância de dados. Se uma variável estiver positivamente correlacionada com o variável resposta, um valor maior que sua própria média contribuirá positivamente para a previsão. Se uma variável estiver correlacionada negativamente com a variável resposta, um valor maior que sua própria média contribuirá negativamente para a previsão.

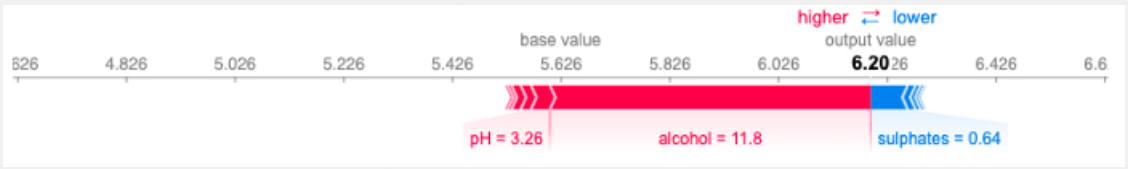

Figura 6 – **Exemplo de um *force plot* para uma observação individual.**
Fonte: Elaborado pelos autores.

Voltando ao exemplo da figura 6, observa-se que o valor de álcool (*alcohol*) para essa predição individual possui um impacto positivo na predição (cor vermelha), além de ter o maior impacto entre as variáveis que impactam positivamente a previsão (por exemplo, pH). Sulfatos (*Sulphates*), ao contrário, possui impacto negativo (cor azul) tendo o maior impacto comparado as demais que impactam negativamente. Para observar exemplos de predições individuais nos resultados, escolhemos quatro predições (duas com *output* de CNPJ e duas sem CNPJ) para as variações do modelo (sem UFs e com UFs).

3 RESULTADOS

3.1 ANÁLISE EXPLORATÓRIA

No caso das variáveis categóricas, realizou-se primeiramente o teste de hipótese Qui-quadrado de independência com a variável resposta ”CNPJ”e todas as variáveis independentes. O resultado do p-valor menor que 5% para todos, indica a rejeição da hipótese nula que diz não haver associação entre as variáveis. Ou seja, o teste mostrou evidências em existir associação entre a variável CNPJ e as variáveis de negócio categóricas. A tabela 1 mostra os resultados das medidas de associação adotadas.

Tabela 1 – **Medidas de correlação.**

Variável	Coeficiente de Contingência	V de Crámer
Rendimento	0,41	0,45
Local	0,42	0,39
Pessoas	0,29	0,28
Setor	0,26	0,27
Sócios	0,26	0,27
UF	0,25	0,24

Fonte: Resultados da pesquisa.

Dentre as variáveis numéricas, aplicou-se a correlação de Spearman e a matriz com os resultados está representada na figura 7

A partir dos resultados das análises de correlação, as variáveis com maior correlação com a variável resposta ”CNPJ”são: ”rendimento”, ”local”, ”anos de estudo”, ”pessoas”, ”cliente”, ”horas_trabalhadas”e ”socios”.

3.2 MODELO FINAL

Conforme as métricas adotadas para classificar o desempenho dos modelos (i.e. AUC e Balanced Accuracy), Catboost obteve melhor desempenho dentre os modelos-base. Após a escolha do modelo final, houve a otimização dos hiperparâmetros através do Optuna (25 trials). Catboost possui vários parâmetros (para mais informações consulte o site: https://catboost.ai/en/docs/), segundo os resultados obtidos pelo Optuna, o conjunto de parâmetros que trouxeram maior desempenho para o modelo foram:

> **Parâmetros**: *learning rate*: 0.06, *colsample by level* 0.04, *n estimators: 780*, *silent*: True, *max depth*: 7, *early stopping rounds*: 100, *boosting type*: Plain, *bootstrap type*: MVS, *random state*: 42.

Desta forma, os parâmetros foram usados dentro da *pipeline* final, composta de uma fase de pré-processamento (imputação de variáveis numéricas e categóricas, e uso de *one-hot encoding* como *encoder* das variáveis categóricas) e outra relacionada ao modelo, ambas descritas nos métodos. Podemos representar a *pipeline* conforme o seguinte diagrama (Figura 8):

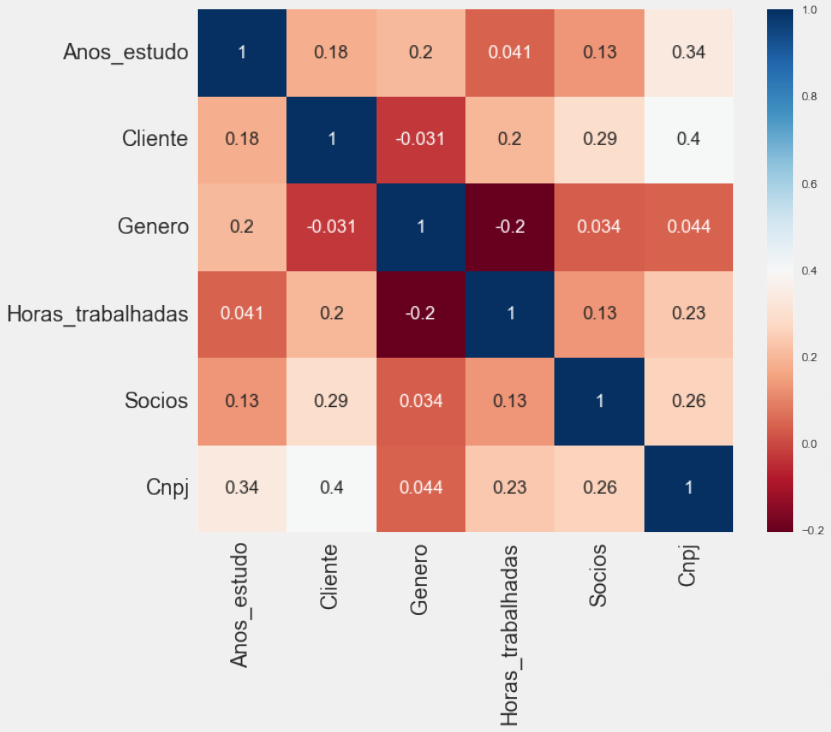

Figura 7 – **Representação gráfica da matriz de Correlação de Spearman.**
Fonte: Elaborado pelos autores.

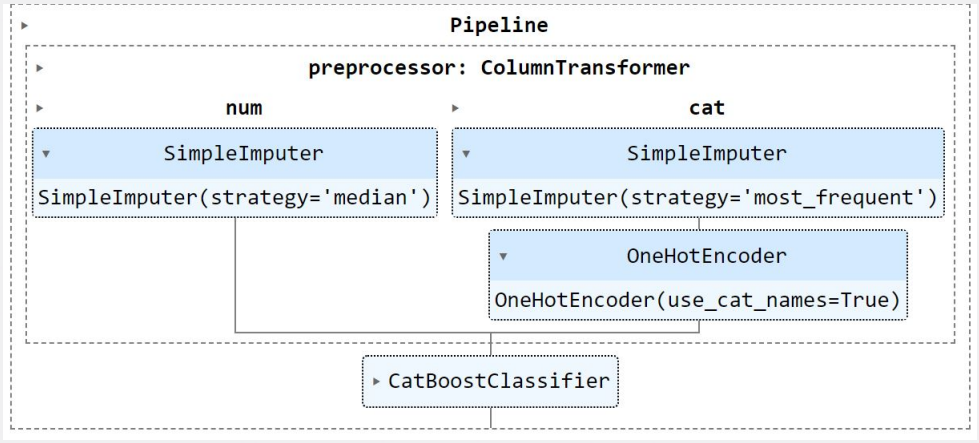

Figura 8 – ***Pipeline* do modelo final.**
Fonte: Elaborado pelos autores.

Os resultados do modelo estão descritos na Tabela 2). Como observado, os resultados ficaram bem similares, com um pequeno aumento nas métricas para quando é considerada a variável de UFs. Conforme a análise de *feature importance*, para a variação do modelo sem a inclusão das UFs (Figura 9) e com inclusão de UFs (Figura 10), as variáveis mais importantes para a predição do modelo foram similares, entre as principais estão: rendimento, anos de estudo, cliente, local, horas trabalhadas e

sócios.

Tabela 2 – **Resultados do modelo final.**

Métricas	Validação Treino/Teste	Validação cruzada (5 folds)
AUC	0.88	0.87
Balanced accuracy	0.73	0.75
Precision	0.75	0.73
Recall	0.52	0.58

2.1 Variação sem UFs

Métricas	Validação Treino/Teste	Validação cruzada (5 folds)
AUC	0.89	0.88
Balanced accuracy	0.77	0.76
Precision	0.74	0.74
Recall	0.54	0.59

2.2 Variação com UFs

Fonte: Resultados da pesquisa.

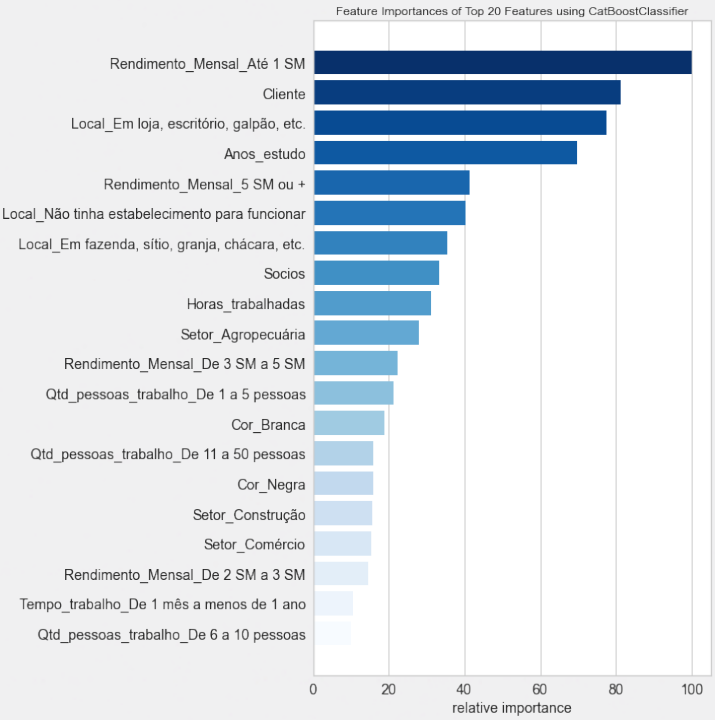

Figura 9 – ***Feature importance* do modelo final - variação sem UFs.**
Fonte: Elaborado pelos autores.

Os valores de shap demonstraram importância similar das variáveis comparada a análise de *feature importance*, porém com uma série de padrões relacionados aos valores das 20 variáveis mais importantes. A figura (11) mostra estes padrões para a variação do modelo sem a inclusão das UFs e a figura (12) com a inclusão das UFs. Em geral, os valores das variáveis influenciaram tanto positivamente

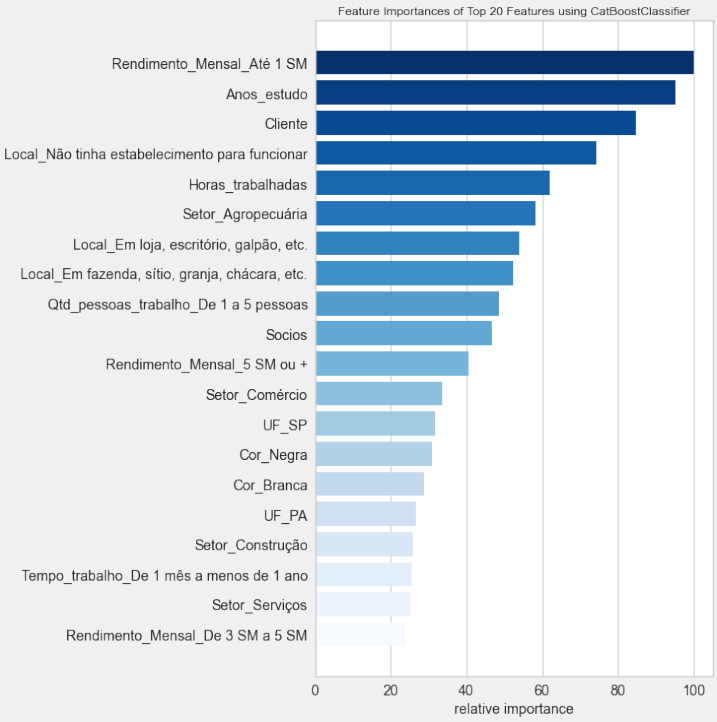

Figura 10 – ***Feature importance* do modelo final - variação com UFs.**
Fonte: Elaborado pelos autores.

quanto negativamente a probabilidade de um dono de negócio ter CNPJ ou não. Entre as variáveis mais importantes, foi observado a maior influência do rendimento para ambas as variações do modelo (Valor médio de SHAP entre: 0.45), onde indivíduos com um rendimento mensal de até 1 SM, tendem a não ter CNPJ. A segunda variável mais importante foi anos de estudo (Valor médio de SHAP entre: 0.35 e 0.42), sendo que quanto maior a quantidade de anos, maior a probabilidade do dono de negócio ser CNPJ. A terceira variável mais importante na variação sem UFs foi o local do cliente, onde se o local era em loja, escritório, galpão, e etc (Valor médio de SHAP: 0.26), a probabilidade do negócio ser CNPJ é maior. Já na variação com UFs, se não existir local para o negócio funcionar (Valor médio de SHAP: 0.29), sua probabilidade de ser CNPJ é menor.

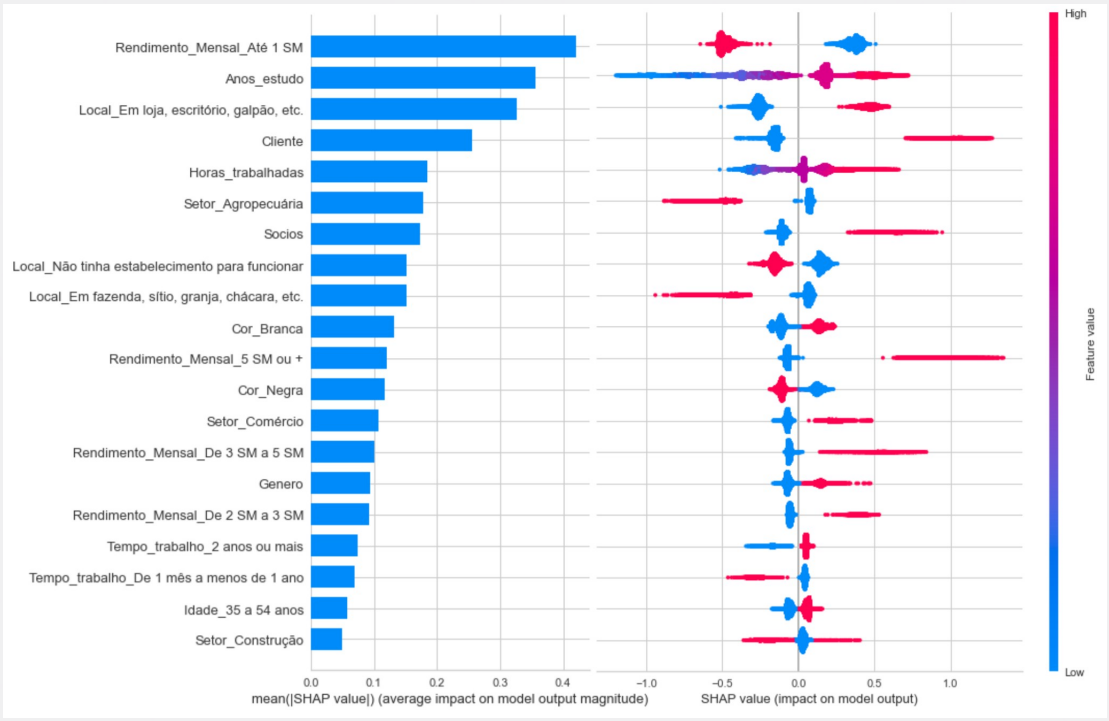

Figura 11 – **SHAP *feature importance* e *summary plot* - variação sem UFs.**
Fonte: Elaborado pelos autores.

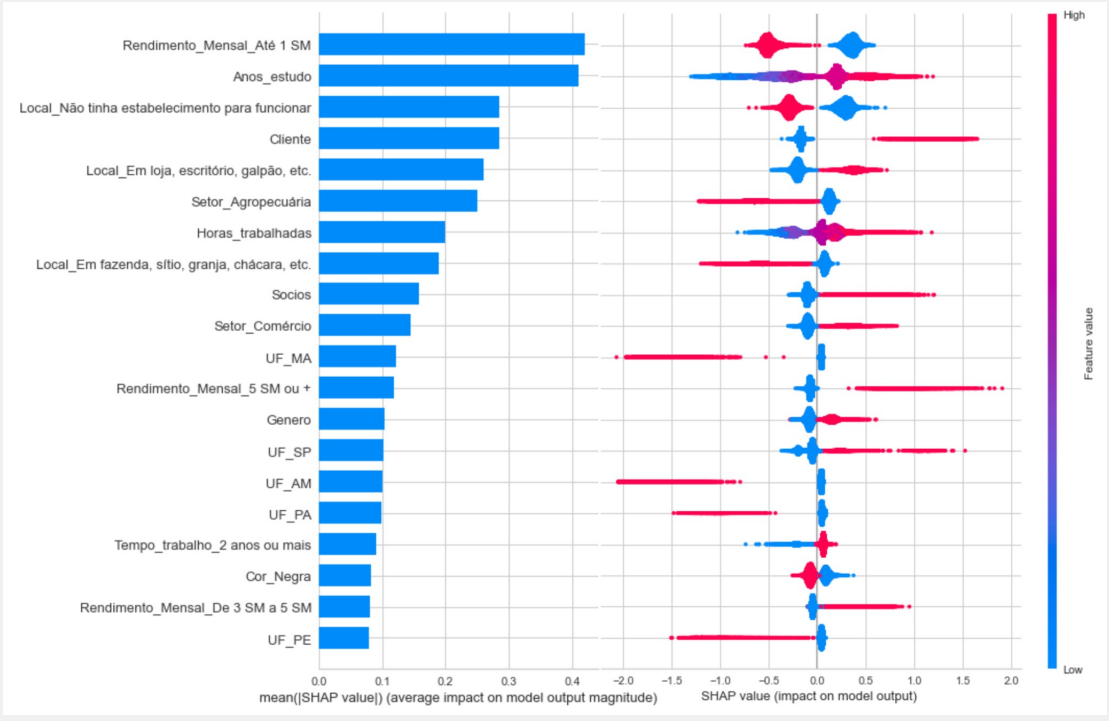

Figura 12 – **SHAP *feature importance* e *summary plot* - variação com UFs.**
Fonte: Elaborado pelos autores.

Considerando somente a influência das UFs no território nacional (Figura 13), as maiores foram MA (Valor médio de SHAP: 0.12), SP (Valor médio de SHAP: 0.10), AM (Valor médio de SHAP: 0.09) e PA (Valor médio de SHAP: 0.08) e PE (Valor médio de SHAP: 0.0). Quando a UF do dono de negócio foi de São Paulo, maior foi a probabilidade do negócio ter CNPJ, diferentemente, quando a UF foi do Maranhão, Amazonas Pará ou Pernambuco, menor foi a probabilidade do empreendimento ter CNPJ.

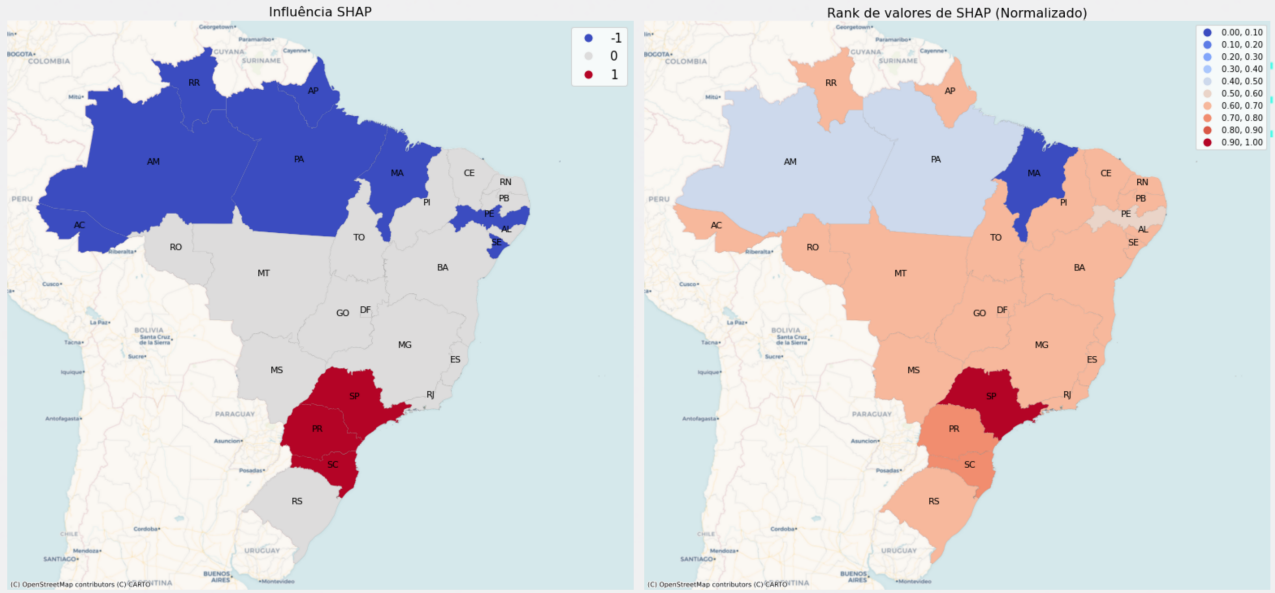

Figura 13 – **Influência do valor de SHAP e rank de valores normalizados.**
Fonte: Elaborado pelos autores.

Podemos também fazer uma análise das predições individuais com os *force plots*. Na figura (14) com exemplos de predições sem a inclusão das UFs, para a primeira predição individual, observa-se que a ausência de anos de estudos (0), e o tempo de trabalho curto (Menos de 1 mês) teve um grande impacto para o modelo indicar uma probabilidade baixa do negócio não ter CNPJ. Já para a segunda predição, os maiores responsáveis foram o tempo de trabalho curto (Menos de 1 mês) e o rendimento mensal ser de até 1 SM. A terceira predição, já caracterizando negócios com CNPJ, teve um rendimento alto (5 SM ou mais) e o cliente ser empregador (Cliente = 1) como grandes responsáveis para uma probabilidade alta do negócio ser CNPJ. Na quarta predição, tanto o cliente ser empregador, quanto a presença de sócios contribuiu significativamente para uma maior probabilidade do negócio ser CNPJ.

Na figura (15) com exemplos de predições com a inclusão das UFs, padrões similares a primeira variação são encontrados, porém, é observado a influência das UFs presente, alguns estados tiverem um impacto considerável na diminuição da probabilidade do negócio ser CNPJ, como nos casos do estado de Pernambuco na primeira predição, Roraima na segunda e Amazonas na quarta.

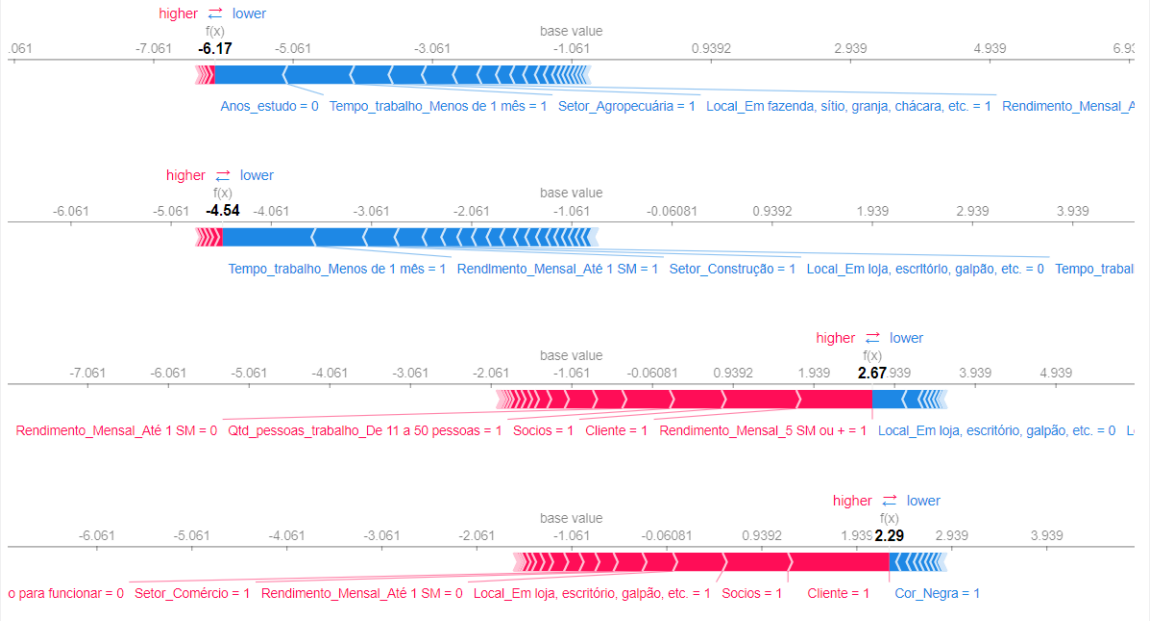

Figura 14 – ***Force plots* de observações individuais - variação sem UFs.** As duas primeiras predições, de cima para baixo, são de negócios sem CNPJ, as duas últimas, de empreendimentos com CNPJ.
Fonte: Elaborado pelos autores.

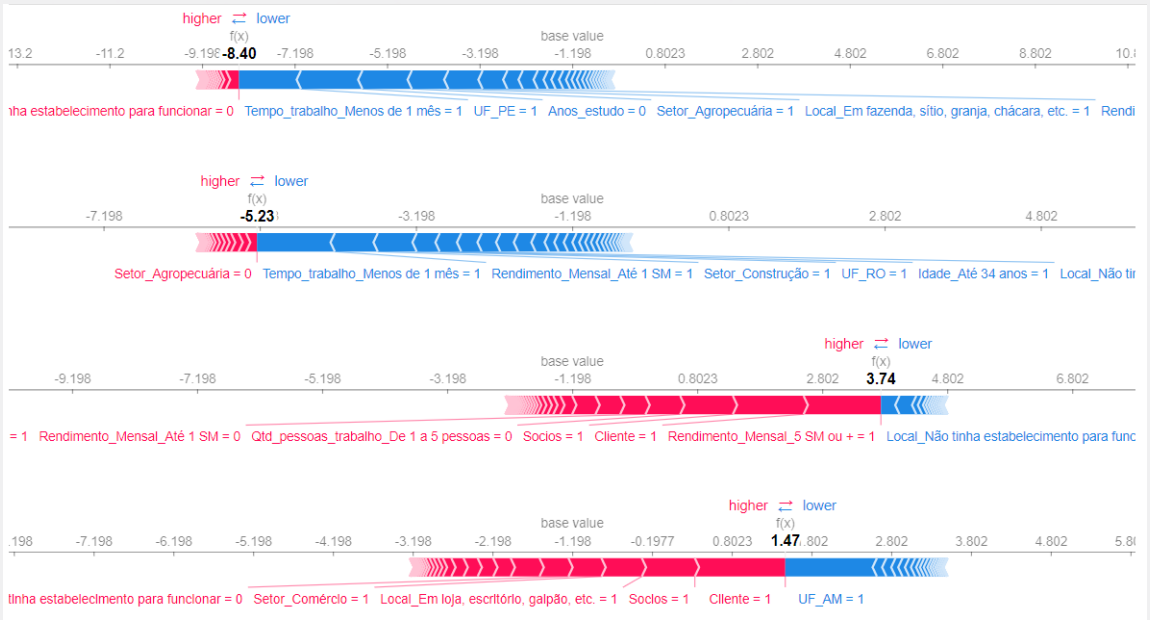

Figura 15 – ***Force plots* de observações individuais - variação com UFs.** As duas primeiras predições, de cima para baixo, são de negócios sem CNPJ, as duas últimas, de empreendimentos com CNPJ.
Fonte: Elaborado pelos autores.

4 CONCLUSÃO

Dentre os Donos de Negócios, apenas 32,08% dos negócios/empresas possui registro CNPJ, ou seja, existem cerca de 20 milhões de potenciais empresários que não estão formalizados. Em geral, dentre os fatores mais importantes para a formalização do negócio, o modelo apontou que quanto maior o rendimento do indivíduo, maior a probabilidade do negócio/empresa ter CNPJ, o que condiz com os limites de faturamento existentes para o enquadramento jurídico da organização. Um número maior de anos de estudo do indivíduo também caracterizou uma probabilidade maior do negócio ser CNPJ, tal resultado condiz com estudo socioeconômicos que indicam uma associação de renda maior com alta escolaridade. A categoria do local onde o empreendimento funciona também é uma variável significativa. Dado que se tal atividade é desempenhada em um escritório ou galpão, a probabilidade é maior desse negócio possuir CNPJ. Em contrapartida, a falta de um estabelecimento para funcionar diminui essa probabilidade. Outras variáveis importantes, mas com um peso menor, foram se o cliente fosse empregador, caracterizando negócios com CNPJ. Segmento de atuação do negócio/empresa, em que se o empreendimento estiver classificado na categoria de agropecuária (por exemplo), causa menor probabilidade dele ter CNPJ. A presença de sócios e o número de horas trabalhadas na semana, aumentam a probabilidade do negócio estar formalizado. A visão do modelo por Unidade Federativa apontou em destaque o estado de São Paulo como uma variável que aumenta a probabilidade de o negócio/empresa possuir CNPJ. De outra forma, se o empreendimento do indivíduo tiver residência no Maranhão, Amazonas, Pará e Pernambuco, essa probabilidade é menor, apontando uma maior necessidade de suporte para esses estados no sentido do empreendedorismo. O bom desempenho do modelo conforme as métricas consideradas, apenas reforçou que os resultados aqui apresentados conseguem ser robustos o suficiente para descrever a importância que determinadas variáveis possuem para caracterizar os fatores que levam um empreendimento estar formalizado.

REFERÊNCIAS

AKIBA, Takuya et al. Optuna: A Next-generation Hyperparameter Optimization Framework. arXiv, 2019. Disponível em: <https://arxiv.org/abs/1907.10902>.

BENGFORT, Benjamin et al. **Yellowbrick**. 14 nov. 2018. DOI: 10.5281/zenodo.1206264. Disponível em: <http://www.scikit-yb.org/en/latest/>.

BUITINCK, Lars et al. API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning. 2013. P. 108–122.

DOROGUSH, Anna Veronika; ERSHOV, Vasily; GULIN, Andrey. CatBoost: gradient boosting with categorical features support. arXiv, 2018. Disponível em: <https://arxiv.org/abs/1810.11363>.

GÉRON, Aurélien. Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems. O'Reilly, 2019.

IBGE. **PNAD Contínua - Pesquisa Nacional por Amostra de Domicílios Contínua**. 2021. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/habitacao/17270-pnad-continua.html?=&t=resultados>. Acesso em: 4 mai. 2022.

LUMLEY, Thomas. **survey: analysis of complex survey samples**. 2020. R package version 4.0.

LUNDBERG, Scott M; LEE, Su-In. **A Unified Approach to Interpreting Model Predictions**. Edição: I. Guyon. Curran Associates, Inc., 2017. P. 4765–4774. Disponível em: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>.

MCGINNIS, William D. et al. Category Encoders: a scikit-learn-contrib package of transformers for encoding categorical data. **Journal of Open Source Software**, The Open Journal, v. 3, n. 21, p. 501, 2018. DOI: 10.21105/joss.00501. Disponível em: <https://doi.org/10.21105/joss.00501>.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2021. Disponível em: <https://www.R-project.org/>.

SIEGEL, S.; CASTELLAN, N.J. **Estatı́stica não-Paramétrica Para Ciências do Comportamento**. Artmed Editora. ISBN 9788536313580. Disponível em: <https://books.google.com.br/books?id=eHejDgAAQBAJ>.

VAN ROSSUM, Guido; DRAKE JR, Fred L. **Python tutorial**. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

WICKHAM, Hadley. **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. Disponível em: <https://ggplot2.tidyverse.org>.


SEBRAE
50+50