EasyChair Preprint

№ 11027

The Usage of Machine Learning Techniques to Identify Frauds on Energy Distribution

Thiago de Almeida Ribeiro and Luciana Martinez

October 6, 2023

Utilização de Técnicas de Aprendizado de Máquina Para Identificação de Fraudes na Distribuição de Energia Elétrica

Thiago de Almeida Ribeiro * **Luciana Martinez** **

* *Departamento de Engenharia Elétrica e de Computação, Universidade Federal da Bahia, BA, (e-mail: ribeiro.thiago@ufba.br).*
** *Departamento de Engenharia Elétrica e de Computação, Universidade Federal da Bahia, BA (e-mail: lucianam@ufba.br)*

Abstract: This article presents a proposal for a tool to combat non-technical losses in electricity distribution systems, based on Machine Learning techniques and consumer data from a Brazilian energy distributor. An exploratory analysis of the data was carried out to identify the variables to be used in training the models, with the purpose of selecting facilities to be inspected on site. The model with the best theoretical result was used on site tests.

Resumo: Este artigo apresenta uma proposta de ferramenta para auxiliar o combate às perdas não técnicas em sistemas de distribuição de energia elétrica, baseada em técnicas de Aprendizado de Máquina e dados de consumidores de uma distribuidora de energia brasileira. Uma análise exploratória dos dados foi realizada para se identificar as variáveis a serem utilizadas no treinamento nos modelos, com o objetivo de selecionar instalações a serem inspecionadas em campo. O modelo com melhor resultado teórico foi utilizado para testes em campo.

Keywords: non-technical losses; energy distribution; data analysis; machine learning.
Palavras-chaves: perdas não técnicas; distribuição de energia; análise de dados; aprendizado de máquina.

1. INTRODUÇÃO

Em um sistema de distribuição de energia elétrica, as perdas podem ser definidas como a diferença entre a energia elétrica adquirida pelas distribuidoras e a faturada aos seus consumidores. Essas perdas podem ser técnicas ou não técnicas. As Perdas Não Técnicas (PNTs) decorrem em geral de fraudes ou erros de medição e faturamento, e tem seus valores regulatórios determinados pela Agência Nacional de Energia Elétrica (ANEEL), com base em critérios de eficiência e características socioeconômicas das áreas de concessão.

No Brasil, no ano de 2020 as PNTs reais representaram um custo de aproximadamente R$ 8,6 bilhões, enquanto o custo das PNTs regulatórias foi de aproximadamente R$ 5,6 bilhões, o que representou aos consumidores cerca de 2,9% do valor da tarifa de energia elétrica, variando por distribuidora (ANEEL, 2021). Segundo a ANEEL, a diferença de custos entre o valor regulatório e o real é de responsabilidade da concessionária de energia. Portanto, as PNTs impactam financeiramente tanto a distribuidora de energia quanto o consumidor final. Além de impactos financeiros, as PNTs causam ainda impactos sociais, como insegurança, concorrência desleal, desperdício de energia e não arrecadação de impostos (Dantas, 2006). Assim, o combate às PNTs em sistemas de distribuição pode trazer benefícios tanto para as distribuidoras de energia como para a sociedade em geral.

Para que seja identificada a origem das possíveis perdas em um sistema de distribuição, normalmente é necessária a visita de uma equipe técnica à instalação, o que gera custos e demanda tempo. Para as distribuidoras é interessante que a rotina das inspeções seja otimizada, com o envio de equipes apenas para instalações onde exista algum indício prévio de perda. Outro ponto importante a ser considerado é a automatização do processo de seleção da instalação a ser vistoriada, em especial diante de um número elevado de consumidores.

O uso de tecnologias que facilitem a identificação de PNTs, assim como auxiliem a construção de modelos analíticos, tem sido cada vez mais recorrente. Em Viegas et al. (2017) é apresentada uma revisão bibliográfica de trabalhos propostos para a detecção de PNTs. Os autores dividem os estudos analisados, 103 no total, em três categorias: estudos teóricos (6), soluções de *hardware* (25) e soluções sem *hardware* (72), destacando vantagens e limitações de cada uma das soluções. Para os estudos teóricos e com uso de *hardware*, os autores citam as limitações de uma baixa precisão de detecção de PNTs ou investimento de capital significativo por parte da distribuidora. Entretanto, essas soluções podem se tornar viáveis para identificar e combater perda em locais considerados críticos. Já as soluções sem *hardware* são consideradas mais acessíveis, por se basearem em informações dos consumidores e medidores na identificação de possíveis perdas, como, por exemplo, o uso de dados para a detecção da probabilidade de um comportamento de consumo ilegal.

As soluções sem *hardware* compreendem a maior parte dos estudos presentes na literatura. Esse tipo de solução, normalmente baseada em um *software* ou algoritmo, abrange técnicas como classificação, estimação, teoria dos jogos, entre outras. A primeira fornece previsões quanto à presença ou não de PNTs significativas em uma zona ou instalação específica, enquanto a segunda realiza um estimativa absoluta ou relativa de tais perdas. Já a teoria dos jogos diz respeito a modelos matemáticos que levam em consideração as relações entre consumidores, fraudadores e a concessionária, no estudo de detecção de perdas.

O termo Aprendizado de Máquina (AM) refere-se a um campo da inteligência artificial voltado ao desenvolvimento de algoritmos para análise automática de dados. Uma das principais aplicações do AM está no problema de classificação, com as soluções baseadas em técnicas como Máquinas de Vetores de Suporte (SVM), Redes Neurais Artificiais (RNA), Árvore de Decisão (*Random Forest* e *Gradient Boosting*) e Florestas de Caminhos Ótimos (OPF).

É possível encontrar na literatura diferentes trabalhos científicos que utilizam técnicas de classificação para a detecção de PNTs em distribuidoras nacionais, como é o caso dos trabalhos de Paulo (2020) e Barros (2021) (Árvore de Decisão, *Random Forest*, *Gradient Boosting*, RNA e SVM) e Ramos (2010) (OPF, RNA e SVM), adotados aqui como os principais referenciais metodológicos. As abordagens diferem consideravelmente a depender dos tipos de dados disponíveis para a construção dos modelos. Vale citar ainda Saeed et al. (2020), que pode ser considerado uma atualização dos trabalhos de Viegas et al. (2017), trazendo algumas informações adicionais, como uma lista das métricas mais utilizadas para avaliar os modelos: acurácia, precisão, sensibilidade, valor preditivo negativo, F1-*score*, tempos de classificação e treinamento.

Este artigo tem como foco soluções sem *hardware* na detecção de PNTs, embora informações de *hardwares* já existentes na rede de distribuição, como alguns sensores inteligentes, discutidos em Oliveira (2021), sejam também utilizadas. Foram construídos 5 modelos utilizando técnicas de AM, a fim de verificar qual se adaptaria melhor ao problema em questão: SVM, RNA (*Perceptron* Multicamadas), *Random Forest*, *Gradient Boosting* e OPF. O estudo considerou especificamente consumidores do Grupo B de determinada distribuidora brasileira. O modelo que forneceu melhores resultados teóricos foi então utilizado para selecionar instalações para serem inspecionadas em campo.

O artigo está organizado como segue. Na seção 2 é apresentada a metodologia adotada no trabalho, com definição das variáveis, análise e tratamento dos dados. Na seção 3 é apresentada a concepção e testes dos modelos propostos. Na seção 4 é realizada uma discussão sobre os resultados obtidos. A seção 5 apresenta a conclusão do estudo realizado.

2. METODOLOGIA

Neste artigo, a proposta de identificação de fraudes na distribuição de energia elétrica teve como base a metodologia adotada em Paulo (2020), originalmente proposta por Pyle (1999). Seu fluxograma pode ser visto na Figura 1.

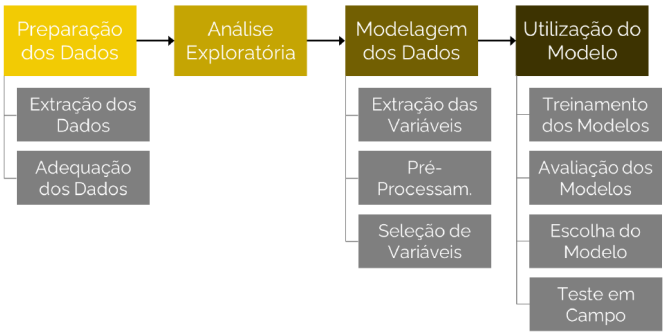

Figura 1. Fluxograma da metodologia (Paulo, 2020)

2.1 Preparação dos Dados e Criação das Variáveis

O estudo aqui proposto teve como base dados de uma empresa distribuidora de energia elétrica brasileira, que por questões éticas não terá o nome divulgado. Os sistemas comerciais da distribuidora em questão possuem informações sobre os clientes da empresa, relacionadas tanto ao serviço prestado, quanto à unidade consumidora. As informações utilizadas neste trabalho foram classificadas em cinco categorias: inspeções, cadastro, consumo, leitura e medição, serviços e pagamentos. A escolha de tais informações foi baseada na literatura e na experiência de colaboradores da própria distribuidora.

Ressalta-se que a utilização de técnicas de AM requer a definição de uma variável *target*, que no caso trata-se da existência ou não de fraude. No entanto, como a fraude na instalação só pode ser constatada a partir de uma inspeção, e cada unidade consumidora do banco de dados foi inspecionada em uma data diferente, foi necessário se fazer uma adequação da base considerando uma data de inspeção como a data referência (no caso foi escolhida a inspeção mais recente). Assim, apenas as informações dessa data, ou de datas anteriores a essa, foram consideradas.

Para que a mesma quantidade de informações fosse considerada para todos as amostras do banco de dados, assumiu-se uma janela de 36 meses anteriores à data de referência. Essa janela foi aplicada às bases de dados de consumo, irregularidades e serviços. Além disso, foram atualizados os dados de cadastro com as informações válidas naquela data específica. Como o banco de dados da empresa fornece informações de consumo dos últimos cinco anos, apenas unidades inspecionadas nos últimos dois anos foram consideradas. As variáveis inicialmente consideradas, construídas a partir das informações contidas no banco de dados, podem ser vistas nas Tabelas 1, 2, 3, 4 e 5.

2.2 Análise Exploratória

No desenvolvimento de algoritmos para análise de dados é interessante que o melhor subconjunto dos atributos originais disponível seja selecionado, preservando-se toda ou a maior parte da informação dos dados, eliminando-se aqueles que são irrelevantes ou redundantes. Mesmo que,

Tabela 1. Variáveis de Inspeção

Atributo	Descrição
QTD_FRAUDE	Quantidade de fraudes já encontradas na instalação.
FAMILIAINSPANT	Família da última inspeção (indica se houve fraude ou não).
INSTPARCEIROFRAUDE	Flag que indica se já houve fraude em alguma instalação do cliente.

Tabela 2. Variáveis de Leitura e Medição

Atributo	Descrição
OCLE_ATUAL	Ocorrência de Leitura no mês da fraude.
TENSAOMED	Tensão medida.
PERDA_NAO_TECNICA_PRCT	PNT percentual por unidade consumidora de uma área (obtida a partir de Sensores Inteligentes) ou do alimentador.

Tabela 3. Variáveis de Cadastro

Atributo	Descrição
LATITUDE	Latitude da instalação.
LONGITUDE	Longitude da instalação.
MOTIVO	Razão pela qual a instalação foi incluída na base de dados.
TIP_INSTALACAO	Tipo de instalação.
FAB_MEDIDOR	Fabricante do medidor.
MODELO_MEDIDOR	Modelo do medidor.
IDADE_ATIVA_MEDIDOR	Idade ativa do medidor.
IDADE_INSTALACAO	Idade da instalação.
CLASSE_CONTACONTRATO	Classe da conta contrato.
SUBCLASSE_CONTACONTRATO	Subclasse da conta contrato.
TENSAOFORNEC	Tensão de fornecimento.
MODLIGA	Modo de ligação (direta ou indireta).
TIPO_LOCAL	Tipo de local (urbano ou rural).
BAIXARENDA	Flag de clientes de baixa renda.

em geral, uma redução na dimensão do vetor de entrada signifique uma redução de informação, em aplicações reais, como a quantidade de dados é limitada, a maldição da dimensionalidade leva a dados esparsos e pode reduzir a performance de sistemas de classificação (Bishop, 1995).

Para se analisar os dados utilizados neste trabalho, foi determinado o cálculo da correlação entre as variáveis das Tabelas 1 a 5, utilizando-se o coeficiente de Pearson, dado por (1), onde $\overline{x}$ e $\overline{y}$ são as médias aritméticas e x_1, x_2, ..., x_n e y_1, y_2, ..., y_n são os valores que compõem os dois conjuntos.

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2} \cdot \sqrt{\sum (y_i - \overline{y})^2}} \qquad (1)$$

O coeficiente r varia entre -1 e $+1$. A correlação é considerada forte para coeficientes maiores que $0,8$ ou menores que $-0,8$ (Devore, 2010). Com base em Paulo (2020), foram eliminadas aqui as variáveis numéricas que possuíam correlação forte com alguma outra, de forma a reduzir a redundância de dados. A matriz de correlação determinada pode ser vista na Figura 2, e as variáveis eliminadas foram: **MAXIMO_VENCIMENTO, QTD_QUEDA_24M, QTD_QUEDA_36M, QTD_QUEDA_48M, COMP_CV24M_CVVZ24M, CORTEEXECUTADO, RELIGACAORECORTE, RELIGACAOEFETIVO e RELIGACAOCORTE**.

Tabela 4. Variáveis de Consumo

Atributo	Descrição
CONS_MES_FRAUDE	Consumo do mês anterior à fraude.
QUEDA	Quantidade de meses desde a última queda.
TX_QUEDA_POS	Taxa de queda pós-queda.
TX_QUEDA_POS	Taxa de queda recente.
QTD_QUEDA_12M	Quantidade de quedas últimos 12 meses.
QTD_QUEDA_24M	Quantidade de quedas últimos 24 meses.
QTD_QUEDA_36M	Quantidade de quedas últimos 36 meses.
QTD_QUEDA_48M	Quantidade de quedas últimos 48 meses.
QTD_QUEDA_60M	Quantidade de quedas últimos 60 meses.
COMP_MED12M_MEDVZ12M	Comparação da média de consumo 12 meses com a média de consumo 12 meses do vizinho.
COMP_MED24M_MEDVZ24M	Comparação da média de consumo 24 meses com a média de consumo 24 meses do vizinho.
COMP_MED36M_MEDVZ36M	Comparação da média de consumo 36 meses com a média de consumo 36 meses do vizinho.
COMP_CV12M_CVVZ12M	Comparação do coeficiente de variação do consumo 12 meses com o coeficiente de variação do consumo 12 meses do vizinho.
COMP_CV24M_CVVZ24M	Comparação do coeficiente de variação do consumo 24 meses com o coeficiente de variação do consumo 24 meses do vizinho.
COMP_CV36M_CVVZ36M	Comparação do coeficiente de variação do consumo 36 meses com o coeficiente de variação do consumo 36 meses do vizinho.
COMP_MEDIA12_MED13_24M	Comparação da média de consumo do último ano com a média de consumo do 13º ao 24º mês.
COMP_MEDIA12_MED25_36M	Comparação da média de consumo do último ano com a média de consumo do 25º ao 36º mês.

Tabela 5. Variáveis de Serviços e Pagamentos

Atributo	Descrição
FATPAGVENC_12M	Quantidade de faturas pagas até o vencimento nos últimos 12 meses.
MEDIA_ATRASO_12M	Diferença média de dias de atraso entre o vencimento e o pagamento das 12 últimas faturas.
MAXIMO_VENCIMENTO	Máximo de dias para compensação de uma fatura em atraso nos últimos 12 meses.
MINIMA_VENCIMENTO	Mínimo de dias para a compensação de uma fatura em atraso nos últimos 12 meses.
CORTEEFETIVO	Quantidade de cortes efetivos.
CORTEEXECUTADO	Quantidade de cortes executados.
RECORTEEFETIVO	Quantidade de recortes efetivos.
RECORTEEXECUTADO	Quantidade de recortes executados.
RELIGACAORECORTE	Quantidade de religações recorte.
RELIGACAOEFETIVO	Quantidade de religações efetivas.
RELIGACAOCORTE	Quantidade de religações corte.

2.3 Tratamento dos Dados

Eliminadas do conjunto de dados as variáveis com forte correlação, um subconjunto final de 45 variáveis passou a ser considerado. Entretanto, muitas dessas variáveis ainda precisaram passar por algum tratamento antes do treinamento dos modelos.

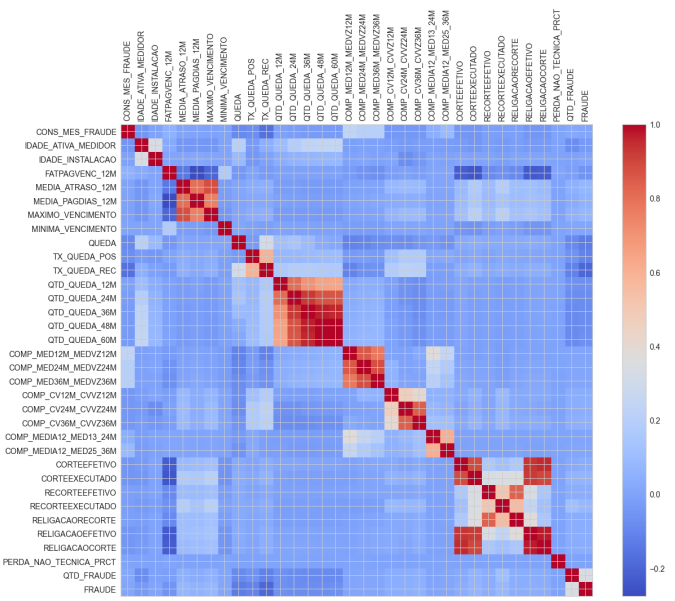

Figura 2. Matriz de Correlação de Pearson

Inicialmente foram tratadas as variáveis com ocorrências de valores nulos ou mesmo impossíveis. Para a variável PERDA_NAO_TECNICA_PRCT, os valores nulos e menores do que zero foram substituídos pela média aritmética dos valores positivos. Ressalta-se que, na prática, é impossível que haja uma perda percentual negativa. Isso normalmente ocorre devido a erros de cadastro. Para as variáveis LATITUDE e LONGITUDE, os valores nulos foram substituídos pela latitude e longitude do município de localização da instalação e, quando nem isso foi possível, pela média aritmética de todos os valores.

Em seguida, foi feita a codificação das variáveis categóricas em variáveis numéricas, ou seja, cada ocorrência diferente foi representada por um número específico, sendo o processo automatizado com o auxílio da função *LabelEncoder( )* da biblioteca *Scikit-Learn* (linguagem Python). Por fim, as linhas do arquivo de dados que ainda possuíam algum valor nulo foram descartadas, resultando num conjunto final de 46.977 instalações, de um total de 60.432 da base de dados extraída e 290.391 do nicho de clientes considerado.

3. CONCEPÇÃO E TESTE DOS MODELOS

3.1 Criação dos modelos

Como proposta para identificação das PNTs, 5 diferentes modelos foram construídos, a fim de verificar qual se adaptaria melhor ao problema em questão. As diferentes técnicas de classificação consideradas foram: SVM, RNA (*Perceptron* Multicamadas), *Random Forest*, *Gradient Boosting* e OPF. Detalhes destas técnicas podem ser vistos em Barros (2021), Ramos (2010) e Ribeiro (2022).

Para a implementação computacional dos modelos, foi escolhida a linguagem Python, amplamente adotada no contexto de ciência de dados. As seguintes bibliotecas foram utilizadas:

- **Scikit-Learn**: ferramental para trabalhar com grande parte das técnicas de AM, trazendo, além dos classificadores, funções de pré-processamento dos dados e métricas de avaliação dos modelos.
- **OPFython**: classificador OPF.
- **Yellowbrick**: visualização dos dados (como em matrizes de confusão, por exemplo).
- **Pickle**: exportar modelos na forma de arquivo .sav.
- **Pandas**: manipulação e análise dos dados.
- **NumPy**: funções matemáticas.
- **Datetime**: classes para manipulação de datas e horas.

Em todos os modelos implementados, o valor inicial dos parâmetros foi escolhido com base na literatura e na experiência dos autores. Em seguida, esses parâmetros foram ajustados empiricamente em busca de um melhor resultado. Os valores finais considerados são apresentados na Tabela 6. Os parâmetros que não foram citados assumiram os valores padrão da biblioteca utilizada (*OPFython* para OPF e *Scikit-Learn* para as demais técnicas).

Tabela 6. Parâmetros dos Modelos

Técnica	Parâmetros Utilizados
OPF	distance = "log_squared_euclidean" pre_computed_distance = None n_iterations = 5
Gradient Boosting	n_estimators = 10 learning_rate = 1
SVM	kernel = 'linear' C = 1.0
Random Forest	criterion = 'gini' n_estimators = 60
RNA (Perceptron Multicamadas)	early_stopping = True max_iter = 1000 tol = 0.0000010 hidden_layer_sizes = (32,) n_iter_no_change = 100

Ressalta-se que, para o treinamento dos modelos, a base de dados disponível foi dividida de forma aleatória entre as bases de treinamento (75%) e teste (25%) (*holdout*).

3.2 Testes Teóricos

Para os modelos desenvolvidos, foram calculadas métricas de interesse, com o auxílio das funções *score( )* (*Scikit-Learn*) e *metrics( )* (*OPFython*). A comparação de tais métricas, para os 5 modelos, pode ser vista na Tabela 7.

Tabela 7. Comparação dos Modelos

	Precisão	Sensibilidade	F1-Score
OPF	0,20	1,00	0,34
Gradient Boosting	0,82	0,23	0,37
SVM	-	-	-
Random Forest	0,96	0,24	0,38
RNA	0,68	0,20	0,31

Nota-se que os resultados para o modelo com a técnica SVM não foram apresentados na Tabela 7. A construção do modelo com uso dessa técnica mostrou-se extremamente custosa computacionalmente, demorando mais de 24hrs para a realização do treinamento e seleção, sendo sua utilização considerada inviável.

As métricas sensibilidade e precisão são consideradas as mais importantes para um classificador binário (Saeed et al., 2020). Tais métricas guardam uma relação de compromisso, de forma que é simples obter uma alta precisão

com baixa sensibilidade ou uma alta sensibilidade com uma baixa precisão. O desafio para um bom classificador é obter a combinação de alta sensibilidade com alta precisão, o que pode ser mensurado a partir do F1-*score*, que consiste na média harmônica entre a sensibilidade e a precisão, conforme expresso em (2).

$$\text{F1-}score = \frac{2}{\frac{1}{sensibilidade} \times \frac{1}{precisao}} \quad (2)$$

Tomando-se o F1-*score* como base de comparação, pode-se observar que, de acordo com a Tabela 7, o melhor resultado foi o do modelo que utiliza a técnica de *Random Forest*, seguido de perto pelo modelo com *Gradient Boosting*.

3.3 Validação dos Resultados e Testes em Campo

Depois de realizado os testes teóricos, 50 instalações foram selecionadas para serem inspecionadas pela distribuidora, de forma a testar o desempenho dos modelos em campo. Embora tal amostra não seja tão significativa quando considerado o espaço amostral em questão, pode ser considerada razoável para um teste preliminar, sem que o faturamento da distribuidora seja afetado de forma considerável.

Usando o F1-*score* como base, o modelo com a técnica de *Random Forest* foi, a princípio, escolhido como ferramenta para selecionar as instalações, a partir de uma base de dados com 200.600 elegíveis para inspeção fornecida pela distribuidora. Neste caso, 2.986 instalações foram selecionadas. Foi testado ainda o modelo com a técnica de *Gradient Boosting*, que apresentou um F1-*score* muito próximo ao da *Random Forest* nos estudos teóricos. Nesse caso, 5.761 instalações foram selecionadas. Considerando que a meta de quantidade de inspeções da distribuidora é relativamente elevada, esse segundo foi considerado mais adequado, já que fornece mais opções de instalações a serem inspecionadas.

Considerando-se o tamanho da base de dados e o uso do *holdout* com amostragem aleatória, resultados de treinamento razoavelmente precisos eram esperados. Não obstante, antes do envio a campo, optou-se por validar o modelo escolhido, tendo sido adotado para isso o método de validação cruzada K-*fold* com K=5 e amostragem estratificada. A validação resultou em médias de precisão, sensibilidade e F1-*Score* de 0,70, 0,26 e 0,37, respectivamente. Vale observar que os valores obtidos foram similares para as duas técnicas consideradas, sendo o F1-*Score* igual.

Em testes teóricos, as instalações a serem inspecionadas são escolhidas aleatoriamente. Para a distribuidora, contudo, a quantidade de energia recuperada com a inspeção é o que mais importa na detecção de perdas. Como a base de dados da distribuidora possui informações de valores de previsão de recuperação, foram escolhidas as 50 instalações com maior previsão de recuperação dentre as selecionadas pelo modelo para inspeção. Como resultado prático, do total de instalações inspecionadas, em 4 foram detectadas perdas, representando num acerto (precisão) de 8% e uma energia recuperada de mais de 88 MWh.

4. DISCUSSÃO DOS RESULTADOS

O modelo proposto apresentou resultado em campo razoável em comparação a outras metodologias usualmente adotadas pela distribuidora em questão para o combate às PNTs. No entanto, de acordo com os resultados obtidos, pode-se verificar uma divergência considerável entre os resultados dos testes teóricos e de campo, para as instalações com maior previsão de recuperação selecionadas para inspeção. Este comportamento da solução pode ser justificado devido a uma série de fatores.

Em primeiro lugar, tem-se a qualidade das bases de dados fornecida pela distribuidora. É possível que hajam variáveis com incorreções, enviesando assim resultados do modelo (*leakage*). Pode-se citar, por exemplo, a variável INSTPARCEIROFRAUDE que, através do atributo *feature_importances* da biblioteca *Scikit-Learn*, demonstrou uma importância elevada no modelo. Para contornar esse problema seria necessário o acesso aos códigos de criação da base de dados e aos dados mãe, não disponibilizados pela empresa.

Outra questão que pode ter afetado os resultados é a desconsideração da evolução temporal dos dados. Como citado na subseção 2.1, uma adequação na base de dados foi realizada considerando-se a data da última inspeção na instalação como data de referência para a obtenção dos dados. Desta forma, foi considerado como se todas as inspeções tivessem ocorrido na mesma data, o que não é verdade na prática. Isso pode ser um problema, visto que determinado padrão pode ser uma característica de fraude em determinada época, mas não em outra, afetando o treinamento do modelo. Ressalta-se ainda a influência da sazonalidade nas inspeções. Existem fraudes que são mais comuns no verão do que no inverno, por exemplo, o que contribui para um pior desempenho do modelo, que não leva esse fato em consideração.

A influência da evolução temporal dos dados pode ser demonstrada na teoria, dividindo-se as bases de teste e treinamento de acordo com a data de inspeção. Para isso, foi considerado aqui o ano de 2021 como base, de forma a se ter também a influência da sazonalidade, e mantida a proporção de 75% dos dados para treinamento e 25% para teste. Além disso, foi também removida do modelo a variável INSTPARCEIROFRAUDE. Com essa abordagem, a performance dos resultados foi reduzida consideravelmente, tendo o F1-*score* baixado de 0,37 para 0,12, a sensibilidade de 0,23 para 0,09 e a precisão, que representaria o acerto, de 0,82 para 0,18, comportamento mais próximo do resultado dos testes em campo.

Outro fator a ser considerado é a qualidade do serviço de campo realizado pela distribuidora. Se as inspeções não forem feitas de forma minuciosa, é possível que as fraudes existentes não sejam detectadas. Assim, a qualidade da inspeção influencia diretamente o resultado dos testes em campo, já que, se as fraudes não forem encontradas, o acerto será reduzido. Tal problema, contudo, pode influenciar ainda o treinamento dos modelos, afinal, se os dados de uma instalação fraudadora forem fornecidos com indicativo que não há fraude na variável *target*, o modelo tomará tal padrão como característico de uma instalação não fraudadora.

Por fim, como citado anteriormente, o número de instalações inspecionadas é muito pequeno, sendo inviável a realização de um processo de inferência dos resultados. A distribuidora normalmente inspeciona cerca de 2.500 instalações desse nicho por mês. Além disso, a própria sazonalidade pode influenciar os testes em campo: o acerto da distribuidora varia ao longo do ano, havendo épocas mais comuns de se encontrar fraudes. Assim, idealmente, os testes deveriam ter sido realizados ao longo do período de um ano. Por questões de cronograma, contudo, isso não foi possível para este trabalho.

5. CONCLUSÃO

Neste artigo foi abordado o problema das perdas não técnicas na rede de distribuição de energia elétrica e apresentada uma possível forma de detectá-las fazendo uso de técnicas de aprendizado de máquina. Embora trate de um tema bastante abordado na literatura, é sempre importante o estudo de novas metodologias que possam colaborar com a diminuição do desperdício de energia e também trazer benefícios financeiro tanto para distribuidoras como consumidores em geral. Além disso, foi possível apresentar o resultado de testes em campo, evidenciando uma discrepância entre a prática e a teoria e revelando algumas possibilidades de melhoria na metodologia em questão.

Os modelos construídos foram desenvolvidos com base em metodologias propostas na literatura. Os algorítimos utilizados foram: Máquinas de Vetores de Suporte, Redes Neurais Artificiais, Árvore de Decisão (*Random Forest* e *Gradient Boosting*) e Florestas de Caminhos Ótimos. Dados de uma distribuidora de energia elétrica brasileira foram utilizados como base, considerando-se especificamente consumidores do Grupo B. Para a implementação computacional dos modelos propostos, foi adotada a linguagem Python.

Os modelos que ofereceram melhores resultados teóricos em termos de F1-*score* e custo computacional foram os baseados em Árvores de Decisão: *Random Forest* e *Gradient Boosting*. O modelo com a técnica *Gradient Boosting* foi então escolhido para a realização de testes em campo, por selecionar um número maior de instalações a serem inspecionadas. O acerto final das 50 inspeções realizadas em campo foi de 8%, sendo recuperados mais de 88 MWh, representando um resultado razoável comparado a outras metodologias adotadas pela distribuidora, embora com resultado consideravelmente divergente da teoria. Tal comportamento pode ser devido a uma série de fatores, como qualidade das bases de dados e serviços de campo, universo amostral reduzido e desconsideração de evolução temporal e sazonalidade.

A qualidade das bases de dados e serviços é de responsabilidade da distribuidora, sendo assim um fator externo aos modelos construídos, difícil de ser alterado. Também não seria viável tomar uma amostra de tamanho considerável para teste em campo devido ao risco de se afetar o faturamento da empresa. Dessa forma, acredita-se que seria interessante para trabalhos futuros a investigação uma forma de se considerar a evolução temporal no treinamento do modelo, o que poderia ser feito, por exemplo, considerando-se apenas dados dos meses imediatamente anteriores ou dados do mesmo período do ano anterior (de forma a levar em conta a sazonalidade). A eficiência de tais métodos, entretanto, pode variar de consumidor para consumidor, além da possibilidade de ser afetada pela redução da quantidade de dados de treinamento, sendo necessário um estudo mais aprofundado sobre o assunto para possíveis conclusões.

O estudo foi realizado após o período da pandemia de Covid-19, durante a qual houve alteração drástica dos padrões de consumo de energia elétrica, devido ao isolamento social, resultando em dados com comportamento diferente daqueles encontrados na literatura usada como base para este trabalho.

REFERÊNCIAS

ANEEL (2021). Perdas de energia elétrica na distribuição. Disponível em: `https://www.aneel.gov.br/documents/654800/18766993/Relat\%C3\%B3rio+Perdas+de+Energia_+Edi\%C3\%A7\%C3\%A3o+1-2021.pdf/143904c4-3e1d-a4d6-c6f0-94af77bac02a`.

Barros, R.M.R. (2021). *Advanced Analytics Aplicado à Gestão da Perda Não Técnica em Energia em Sistemas Elétricos de Distribuição*. Tese de doutorado. Universidade Federal de Campina Grande.

Bishop, C. (1995). *Neural networks for pattern recognition*. Oxford University Press, USA.

Dantas, P.R.P. (2006). *Avaliação de perdas de energia elétrica não-técnicas metodologia aplicada no Município de Salvador*. Dissertação de mestrado. Universidade Salvador.

Devore, J. (2010). *Probabilidade e estatística para engenharia e ciências*. Cengage Learning Edições Ltda.

Oliveira, D.R.M. (2021). *Utilização de Sensores Inteligentes na detecção e compare às perdas não técnicas na distribuição de energia*. Trabalho de Conclusão de Curso. Universidade Federal da Bahia.

Paulo, F.R. (2020). *Detecção de fraude em unidades consumidoras não telemedidas com uso de técnicas de aprendizado de máquina*. Dissertação de mestrado. Universidade Federal da Paraíba.

Pyle, D. (1999). *Data Preparation for Data Mining*. ITPro collection. Elsevier Science.

Ramos, C.C.O. (2010). *Desenvolvimento de Ferramentas Computacionais Inteligentes para Identificação de Perdas Comerciais em sistemas de Energia*. Dissertação de mestrado. Universidade Estadual Paulista.

Ribeiro, T.d.A. (2022). *Utilização de técnicas de aprendizado de máquina para identificação de fraudes na distribuição de energia elétrica*. Trabalho de Conclusão de Curso. Universidade Federal da Bahia.

Saeed, M.S., Mustafa, M.W., Hamadneh, N.N., Alshammari, N.A., Sheikh, U.U., Jumani, T.A., Khalid, S.B.A., and Khan, I. (2020). Detection of non-technical losses in power utilities—a comprehensive systematic review. *Energies*, 13(18). doi:10.3390/en13184727.

Viegas, J.L., Esteves, P.R., Melício, R., Mendes, V., and Vieira, S.M. (2017). Solutions for detection of non-technical losses in the electricity grid: A review. *Renewable and Sustainable Energy Reviews*, 80, 1256–1268.