Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Vinícius Alboneti Aguiar

Algoritmo de Aprendizado de Máquina na Predição de Perdas não Técnicas na Rede Elétrica de Distribuição

Curitiba
2022

Vinícius Alboneti Aguiar

Algoritmo de Aprendizado de Máquina na Predição de Perdas não Técnicas na Rede Elétrica de Distribuição

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Setor de Ciências Exatas, Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Dr. Luiz Eduardo S. de Oliveira

Curitiba

2022

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Algoritmo de Aprendizado de Máquina na Predição de Perdas não Técnicas na Rede Elétrica de Distribuição

Vinícius Alboneti Aguiar[1]
Prof. Dr. Luiz Eduardo S. de Oliveira[2]

Resumo

As perdas não técnicas ou comerciais decorrem principalmente de furto (ligação clandestina, desvio direto da rede) ou fraude de energia (adulterações no medidor), popularmente conhecidos como "gatos", erros de medição e de faturamento. No entanto, os métodos existentes para a detecção deste comportamento de fraude são complexos e manuais.

Este projeto aborda uma nova proposta para detecção de furtos de eletricidade composta por duas etapas: 1) Foram experimentadas várias características e combinadas em conjuntos caracterizados em quatro critérios: temporalidade, localidade, similaridade e infraestrutura. 2) Em seguida, foi utilizado um conjunto de características para treinar três algoritmos de aprendizado de máquina. A hipótese é que o conjunto de características derivadas apenas de dados independentes, são adequadas para uma detecção precisa de fraude.

Os experimentos foram realizados utilizando dados reais de consumo de eletricidade, e os resultados mostram que o método proposto supera os métodos tradicionais em termos de detecção de fraude.
Palavras-chave: Ciência de dados, Inteligência artificial, Perdas não técnicas, Aprendizado de Máquina

Abstract

Non-technical or commercial losses are mainly due to theft (clandestine connection, direct network detour) or energy utilities fraud (meter tampering), popularly known as "gatos", metering and billing errors. However, existing methods for the detection of this fraud behavior are complicated and manual.

This project addresses a new proposal for electricity theft detection consisting of two steps: 1) Several features were sampled and combined into sets characterized by four criteria: seasonality, location, similarity, and infrastructure. 2) Next, a set of features were used to train three machine learning algorithms. The hypothesis is that the feature sets derived only from independent data, are suitable for accurate fraud detection.

The experiments were conducted utilizing real electricity consumption data, and the results show that the proposed method surpasses the traditional methods in terms of fraud detection.
Keywords: *Data Science, Artificial Intelligence, Non-technical losses, Machine Learning*

1 Introdução

A eletricidade é um fator chave para reduzir a pobreza e melhora a qualidade de vida em todo o mundo. Cerca de 86% da população mundial tem acesso à eletricidade e este número tende a crescer (World Bank, 2020). Uma vez que a eletricidade é gerada, ela é distribuída através de redes elétricas. Durante a fase de distribuição, as perdas ocorrem com bastante frequência, e podem ser definidas como a diferença entre a energia elétrica adquirida pelas distribuidoras e a faturada aos seus consumidores e são classificadas em dois grupos: As perdas técnicas (PT) ou não técnicas (PNT).

As **perdas técnicas** são inerentes à atividade de distribuição de energia elétrica e relacionadas aos componentes dos sistemas elétricos e suas propriedades físicas, como a transformação de energia elétrica em energia térmica nos condutores (efeito Joule) [1], [2]. Os montantes de perdas técnicas são divididos pela energia injetada, que é a energia elétrica inserida na rede de distribuição para atender aos consumidores, incluindo perdas.

Já as **perdas não técnicas** incluem manutenção deficiente do equipamento na rede de distribuição, medidores quebrados, fornecimento não medido e roubo de eletricidade. O roubo de eletricidade é a principal causa das PNTs. Este roubo pode ser causado por contornar ou invadir o medidor de eletricidade, adulterar a leitura do medidor entre outras manipulações técnicas no medidor [3].

Os impactos financeiros das perdas na tarifa de energia também podem ser segregados pelas peras técnicas e não técnicas.

Em 2020, no Brasil, o custo com perdas técnicas, é da ordem de R$ 8,5 bilhões e com perdas não técnicas, representaram um custo de aproximadamente R$ 8,6 bilhões [4]. No entanto, as perdas não técnicas regulatórias, que são calculadas conforme metodologia da ANEEL, considerou um custo de aproximadamente R$ 5,6 bilhões,

[1]Aluno do programa de Especialização em Data Science & Big Data, vinicius.aguiar15@gmail.com.
[2]Professor do Departamento de Informática - DINF/UFPR.

o que representou 2,6% do valor da tarifa de energia elétrica, variando por distribuidora [4].

Em montante de energia, as perdas técnicas na distribuição corresponderam a cerca de 38,8 TWh e as perdas não técnicas 37,9 TWh em 2020 [4]. A figura 1 ilustra a participação dessas perdas.

Figura 1: Perdas sobre a energia injetada (2020). Fonte: Aneel

No cenário do setor elétrico, ter como padrão a utilização de técnicas de modelagem estatística e aprendizagem de máquina são a premissa para uma vantagem competitiva quando consideramos perdas não técnicas, derivadas do furto de energia e de falhas na medição e faturamento. A construção de um processo automatizado e adaptativo capaz de identificar a probabilidade de existência de perda e priorizar vistorias *in loco*, é de extrema importância para o processo.

2 Conjunto de dados

O conjunto de dados utilizado neste trabalho corresponde a dados reais do setor elétrico, com uma amostra total de 38.355 vistorias realizadas pela concessionária de energia. A maioria dos dados é composta por consumidores não fraudulentos e consumidores fraudadores, ao qual são uma parte muito pequena, o que acarreta em um desbalanceamento do conjunto de dados. Os dados são semestrais entre 2015 até o primeiro semestre de 2021. A estrutura da base de dados é apresentada na Tabela 1.

Resultado das Vistorias	Total
Falso	36.319
Verdadeiro	2.036

Tabela 1: Resultado total agrupado por verdadeiro, onde indica-se que encontrou fraude e falso, onde indica-se que não foi encontrado fraude.

2.1 Dados para treinamento

A solução proposta para criação de um conjunto de dados para treinamento do modelo foi obtida a partir de entrevistas com inspetores da companhia de energia elétrica e com anos de experiência na detecção de perdas não técnicas. Foram implementadas regras para validação de informações contratuais, histórico de consumo, histórico de inspeções e principalmente questões legais de Lei Geral de Proteção de Dados Pessoais (LGPD). Os dados mais relevantes, obtidos a partir das entrevistas, para criar um conjunto significativo de características para dar suporte ao treinamento de um modelo de aprendizagem de máquina robusto é apresentada na Figura 1.

Categoria	Descrição
Perfil de consumo	Variáveis relacionadas às series de consumo. Ex.: média, desvio padrão.
Vizinhança	Variáveis relacionadas à possíveis fraudes na vizinhança. Ex.: contador de fraudes, vizinhança geográfica.
Cadastro Comercial	Variáveis relacionadas às informações comerciais. Ex.: tipo de atividade, tipo de disjuntor.
Cadastro Técnico	Variáveis relacionadas às informações técnicas. Ex.: tipo de medidor, tipo de ramal.
Ordem de serviço	Variáveis relacionadas às execuções de ordens de serviços juntos aos consumidores. Ex.: cortes, religações, ligações.

Figura 2: Resumo das variáveis obtidas após as entrevistas.

2.2 Pré processamento dos dados

Inicialmente foi realizada a etapa de pré-processamento, que é um conjunto de atividades englobando a preparação, organização e estruturação dos dados. É uma etapa fundamental que precede a realização de análises e predições a serem aplicadas aos dados. Um método para detectar PNT é derivar características de dados históricos de consumo do cliente, como em [5]: média de consumo, consumo máximo, consumo mínimo, desvio padrão, número de inspeções e consumo médio do bairro residencial, e utilizando uma série temporal de 36 meses desde a última inspeção, como regra.

A fim de obter uma maior representação do padrão de consumo de um cliente, foi criado variáveis de sazonalidade, pensando em período de férias, verão, inverno e etc. No total, de forma resumida, foram utilizadas características: mais de 50 variáveis continuas, pelo menos 4 variáveis discretas, 10 variáveis categóricas e foram criadas 6 variáveis de sazonalidade, a fim de classificar um possível comportamento de consumo e 6 variáveis comportamentais regionalizadas, para classificação de um comportamento coletivo.

Conforme apresentado na Tabela 1, as variáveis foram obtidas a partir das informações disponíveis nos bancos de dados da concessionária.

A maioria das variáveis são relacionadas ao consumo de energia, ou seja, foram obtidas a partir de dados históricos de consumo de energia elétrica. O resultado deve-se ao fato de que o processo de criação de variáveis é focado na descoberta de informações relacionadas à ocorrência de PNT, é um dos principais efeitos da PNT em um consumidor é a redução de consumo de energia elétrica [6].

O próximo passo é combinar os efeitos dessas características por meio da criação de modelos.

3 Metodologia

A literatura sobre detecção de PNT não contribui adequadamente para identificar os tipos de algoritmo de aprendizagem de máquina que são mais adequados para a detecção de PNT. Nesta seção, foram ajustados 3 tipos de modelos de classificação na tentativa de classificar fraudes no sistema elétrico. O conjunto de dados é dividido, de forma aleatória, em conjuntos de treinamento e validação com uma proporção de 80%, e 20%, respectivamente. Devido ao processo de pré processamento de dados, a base de dados foi redimensionada, conforme tabela 2.

Base de dados	Tamanho	Resultado	Total
Treinamento	27.179	Falso	25.705
		Verdadeiro	1474
Validação	6.671	Falso	6308
		Verdadeiro	363

Tabela 2: Apresentação da base de dados após divisão aleatória.

Para cada um dos três modelos, o classificador treinado que realizou o melhor no conjunto de validação é selecionado e testado no conjunto de teste para gerar a métrica de precisão e sensibilidade.

O modelo 1 é o Gradient Boosting o qual é um algoritmo baseado na montagem sucessiva de árvores de decisão simples. A ideia é construir árvores de decisão sucessivas, de tal modo que exemplos classificados incorretamente por árvores anteriores sejam melhores classificados nas árvores seguintes. Ou seja, a árvore atual depende das anteriores tendo maior foco no erro das últimas, exemplos classificados incorretamente anteriormente são ponderados com maior peso nas iterações seguintes [6].

O modelo 2 é o XGBoost (Extreme Gradient Boosting) o qual é um algoritmo de aprendizado de máquina baseado em árvore de decisão que usa uma estrutura de aumento de gradiente. Quando se trata de dados estruturados/tabulares, algoritmos baseados em árvore de decisão são considerados os melhores da sua classe no momento [6][7][8]. Algumas diferenças entre XGBoost e o GBM(Gradient Boost) [6] são:

i. O algoritmo suporta computação distribuída;
ii. Possui um parâmetro de randomização para o treinamento, o que o torna mais resistente ao overfitting;
iii. Explora a existência de matrizes esparsas para reduzir o custo computacional.

O modelo 3 é um Random Forest o qual é um estimador de conjunto que compreende um número de árvores de decisão. Cada árvore é treinada em uma subamostra dos dados e recursos definidos a fim de controlar a sobreposição. Na fase de previsão, é feita uma votação majoritária sobre as previsões das árvores individuais [7].

3.1 Métricas de avaliação

Os resultados foram avaliados utilizando precisão, sensibilidade e F_1-Score, ao qual é uma medida útil do sucesso da previsão quando as classes estão muito desequilibradas. Na recuperação de informações, a precisão é uma medida da relevância do resultado, enquanto a recuperação é uma medida de quantos resultados realmente relevantes são retornados.

A curva de precisão e sensibilidade mostra a compensação entre precisão e sensibilidade para diferentes limites. Uma área alta sob a curva representa alta sensibilidade e alta precisão, onde alta precisão está relacionada a uma baixa taxa de falsos positivos e alta sensibilidade está relacionada a uma baixa taxa de falsos negativos. Pontuações altas para ambos mostram que o classificador está retornando resultados precisos (alta precisão), bem como retornando a maioria de todos os resultados positivos (sensibilidade alta).

1. Precisão (P) é definido como o número de verdadeiros positivos (T_p) sobre o número de verdadeiros positivos mais o número de falsos positivos (F_p), e é definida por:

$$P = \frac{T_p}{T_p + F_p}$$

2. Sensibilidade (R) é definido como o número de verdadeiros positivos (T_p) sobre o número de verdadeiros positivos mais o número de falsos negativos (F_n), e é definida por:

$$R = \frac{T_p}{T_p + F_n}$$

3. F_1-Score (F_1) é definido como a média harmônica de precisão e sensibilidade. Ou seja, quando tem-se um F_1-Score baixo, é um indicativo de que ou a precisão ou a sensibilidade, e é definida por:

$$F_1 = 2\frac{PxR}{P + R}$$

4 Resultados

Os modelos foram gerados durante o período de desenvolvimento e com utilização da base de dados de treinamento. O resultado foi plotado utilizando o Modelo estimado *versus* a base de validação e tem sua acurácia medida através de precisão, sensibilidade e F_1-Score. Os hiperparâmetros são partes fundamentais de um modelo e são variáveis do algoritmo definidas antes do treinamento. Os hiper parâmetros de cada modelo, foram

ajustados para encontrar as configurações otimizadas a serem usadas através, da técnica de otimização chamada Grid Search, a qual é um algoritmo de busca que recebe um conjunto de valores de um ou mais hiper parâmetros e testa todas as combinações dentro dessa vizinhança.

4.1 Modelo 1 - Gradient Boosting

O modelo GBM, teve os hiper parâmetros estimados conforme visto na Tabela 3.

Parâmetro	Valor
Números de Árvores	41
Profundidade	8
Amostra Mínima	10
Taxa de Aprendizado	0.1
Distribuição	Bernoulli

Tabela 3: Hiper parâmetros do modelo GBM.

De maneira adicional, o desempenho do modelo também foi avaliado, onde foi possível obter uma Precisão/Sensibilidade de 0.66, conforme Figura 3. O F_1-Score obtido pelo modelo foi de 0.61.

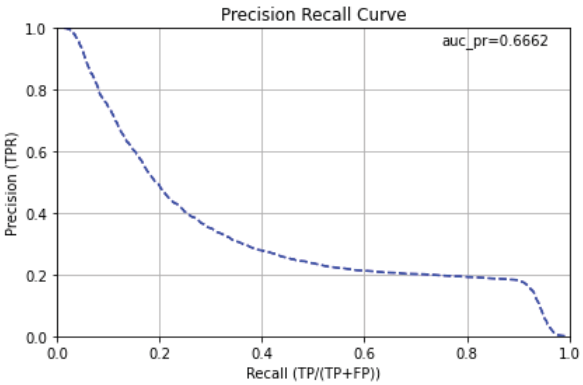

Figura 3: Área sob a curva de precisão e sensibilidade do modelo GBM.

4.2 Modelo 2 - Extreme Gradient Boosting

O modelo XGBoost, teve os hiper parâmetros estimados conforme visto na Tabela 4.

Parâmetro	Valor
Números de Árvores	49
Profundidade	5
Amostra Mínima	10
Taxa de Aprendizado	0.3
Distribuição	Bernoulli

Tabela 4: Hiper parâmetros do modelo XGBoost.

O classificador XGBoost, foi o classificador com o pior desempenho dentre os modelos avaliados, com uma Precisão/Sensibilidade de 0.48 e F_1-Score de 0.44, conforme Figura 4. O que indica que o modelo erra muito mais do que está acerta ao classificar clientes fraudadores e não fraudadores.

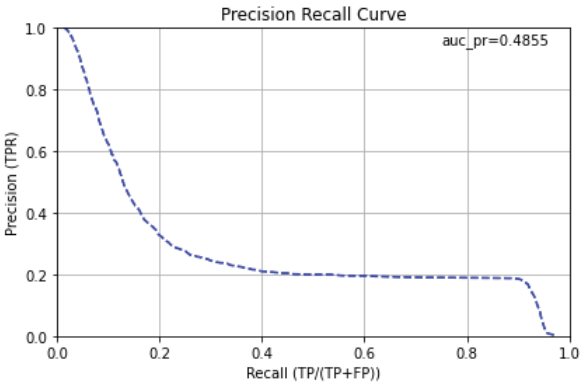

Figura 4: Área sob a curva de precisão e sensibilidade do modelo XGBoost.

4.3 Modelo 3 - Random Forest

O modelo Random Forest, teve os hiper parâmetros estimados conforme visto na Tabela 5.

Parâmetro	Valor
Números de Árvores	48
Profundidade	10
Amostra Mínima	1
Distribuição	Multinomial

Tabela 5: Hiperparâmetros do modelo Random Forest.

O desempenho do modelo três, Random Forest, foi melhor entre os modelos testados obtendo uma Precisão/Sensibilidade de 0.98, conforme Figura 5. O F_1-Score foi de 0.94.

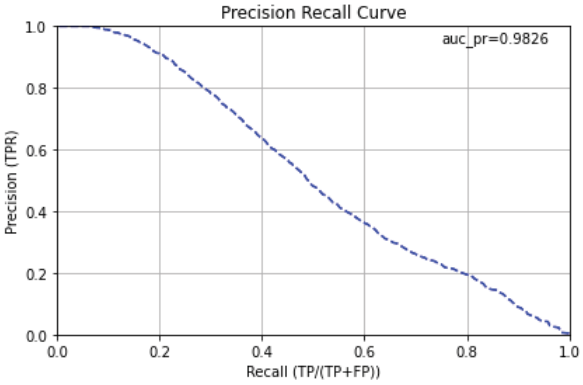

Figura 5: Área sob a curva de precisão e sensibilidade do modelo Random Forest.

5 Discussão

Neste documento, foi utilizado uma abordagem de engenharia de características para a detecção de perdas não técnicas. Foram utilizados três classificadores de aprendizagem de máquinas em conjuntos de características reais computadas usando cinco critérios: temporalidade, localidade, similaridade, sazonalidade e consumo. Os resultados experimentais foram muito satisfatórios e, com base nesta nossa abordagem, a concessionária realizou inspeções *in loco*.

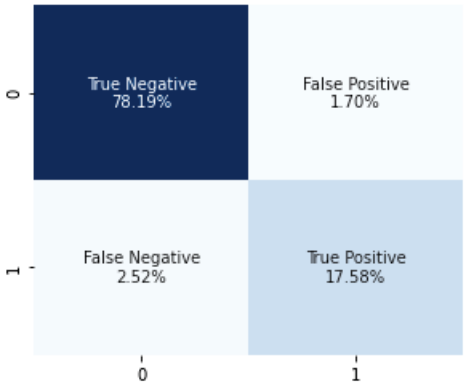

Figura 6: Matriz de confusão do modelo Random Forest.

A matriz de confusão da Figura 6, é gerada conforme resultados do modelo Random Forest, onde é possível identificar uma acurácia 95.77% em classificação de fraudes e não fraudes, sendo que os dados são desbalanceados. No cenário de perdas não técnicas e considerando o mercado ao qual a concessionária está envolvida, este resultado obtido pode ser considerado satisfatório e com um retorno financeiro alto.

A figura 7, demonstra que o modelo Random Forest obteve o melhor desempenho dentre os modelos testados.

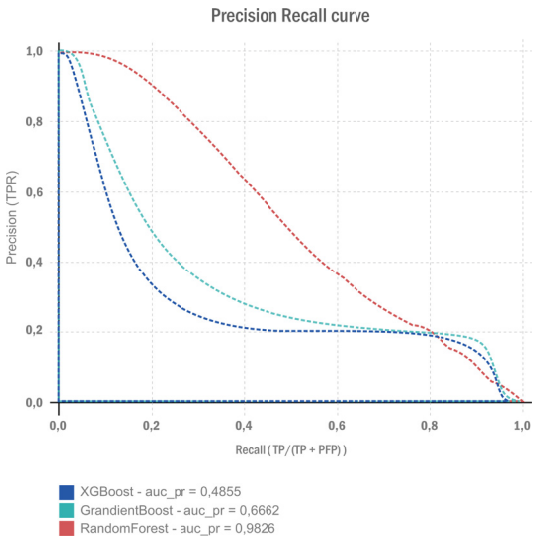

Figura 7: Comparação área sob a curva de precisão e sensibilidade de todos os modelos.

Tendo em vista que os métodos de análise de perdas técnicas e não técnicas estão sendo utilizados por uma grande variedade de pessoas que possuem pouco tempo para a preparação e análise dos dados, uma melhor estruturação dos resultados de saída gerados pelo sistema pode auxiliar em muito o processo de detecção de fraudes. Em pesquisas futuras, técnicas como oversampling, downsampling ou smote serão utilizadas, afim de equilibrar o conjunto de dados. Também será estudado com mais detalhes a correlação de características comportamentais, a fim de entender melhor o padrão comportamental dos clientes. Algoritmos de séries temporais, também serão estudados, afim de detectar anomalias no consumo por um periodo pré determinado.

6 Agradecimentos

Em primeiro lugar, agradeço à minha esposa Jessica por todo amor, carinho e compreensão que massa ao longo de nossa jornada. Seu apoio incodicional durante todo o desenvolvimento deste trabalho, mesmo quando houve necessidade de algumas renuncias para que fosse possível a realização deste trabalho.

Agradeço aos meus pais e minha irmã por me apoiarem e sempre incentivarem a minha dedicação aos estudos.

Ao Professor Luiz Eduardo , pelas ideias, orientações e discussões durante todo o desenvolvimento deste trabalho. Agradeço ainda mais pelos conhecimentos repassados e ensinamentos.

Referências

[1] R. Bhat,R. Trevizan, R. Sengputa, X. Li, and A. Bretas, *Identifying Nontechnical Power Loss via Spatial and Temporal Deep Learning*, (15a Conferência Internacional IEEE sobre Aprendizagem e Aplicações de Máquinas, ICMLA, 2016), pag. 272-279.

[2] E. Sankari e R. Rajesh, *Detection of Non-Technical Loss in Power Utilities using Data Mining Techniques*, (International Journal for Innovative Research in Science Technology), vol. 1, no. 9, pag. 97-101, 2015.

[3] S. McLaughlin, B. Holbert, A. Fawaz, R. Berthier, e S. Zonouz, *A Multi-Sensor Energy Theft Detection Framework for Advanced Metering Infrastructures*, (IEEE Journal on Selected Areas in Communications), vol. 31, no. 7, pag. 1319-1330, 2013.

[4] ANEEL, Relatório de Perdas. Perdas de Energia Elétrica na Distribuição. Brasília, DF. Disponível em: <https://www.aneel.gov.br/documents/654800/18766993/Relat%C3%B3rio+Perdas+de+Energia+Edicao+1-2021
.pdf/>. Acesso em: 12 Dez. 2021.

[5] E. W. S. Angelos, O. R. Saavedra, O. A. C. Cortes, e A. N. d. Souza, *Detection and identification of abnor-*

malities in customer consumptions in power distribution systems, (IEEE Transactions on Power Delivery, 2011.)

[6] R. B. Rocha, , *Advanced Analytics Aplicado à Gestão da Perda Não Técnica de Energia em Sistemas Elétricos de Distribuição.*, <http://dspace.sti.ufcg.edu.br:8080/jspui/bitstream/ri ufcg/20761/1/RAFAEL%20MENDON%C3%87A%20RO CHA%20BARROS%20- %20TESE%20%28PPGEE%29%2 02021.pdf/>. Acesso em: 12 Dez. 2021.

[7] M. Anwar, and N. Javaid, and A. Khalid, and M. Imran, and M. Shoaib, *Electricity theft detection using pipeline in machine learning.* (2020 International Wireless Communications and Mobile Computing, 2020), pag 2138-2142.

[8] C.D. Daykin, *Practical Risk Theory for Actuaries*, (Chapman & Hall, London, 1994), pag. 32.