Análise do Dataset SIGEC - Multa de Fiscalização

João Pedro da Silva de Andrade - 2458810
Vitor Hugo Melo Ribeiro - 2399121

1 de dezembro de 2025

Resumo

O presente trabalho tem como objetivo analisar o comportamento do pagamento das taxas de fiscalização do serviço de energia elétrica no Brasil, utilizando dados disponibilizados pela ANEEL (Agência Nacional de Energia Elétrica). O estudo busca compreender como variáveis como valor previsto e atraso no pagamento influenciam o valor efetivamente pago e o risco de inadimplência. A partir da exploração do dataset `sigec-taxa-de-fiscalizacao.csv`, pretende-se identificar padrões e relações que auxiliem na formulação de estratégias mais eficazes de arrecadação e monitoramento da adimplência no setor elétrico.

1 Introdução

A gestão eficiente da arrecadação é crucial para a sustentabilidade das operações regulatórias da ANEEL. A inadimplência ou atraso no pagamento da taxa de fiscalização geram incertezas orçamentárias. O problema abordado nesse estudo analisa o dataset [1] que contém registros de taxas de fiscalização aplicadas pela ANEEL. A pergunta estudada neste dataset foi: a análise da influência do valor previsto e atraso influenciam o valor efetivamente pago e o risco de inadimplência. As hipóteses testadas foram:

1.1 Hipoteses

- *H1*: O número de dias em atraso tem relação negativa com a proporção de pagamento, ou seja, quanto maior o atraso, menor é o percentual efetivamente pago.
- *H2*: Créditos com valores previstos mais altos apresentam maior probabilidade de inadimplência total.
- *H3*: O número de dias em atraso é influenciado pelo pelo valor total previsto para arrecadação?

2 Métodos

O conjunto de dados bruto, extraído de [1], apresentou diversas inconsistências, ocorrendo nos registros de pagamentos, datas e categorização de situações de crédito. Para garantir a integridade da análise estatística e da modelagem subsequente, foi realizado um tratamento nesses dados.

2.1 Padronização e Conversão de Tipos

A etapa inicial consistiu na conversão explícita dos tipos de dados para formatos adequados à manipulação algébrica e temporal:

- Dados Temporais: As colunas referentes a datas `DatGeracaoConjuntoDados`, `DatVencimentoTitulo`, `DatIncidenciaMultaMora` e `DatPagamentoTitulo` foram convertidas para o formato `datetime`.

- Dados Numéricos e Identificadores: As colunas `Codcvnarr` e `QtdDiasEmAtraso` foram convertidas para inteiros `int64`. A coluna `NumCPFCNPJ` foi tratada como *string*, com a remoção de quaisquer caracteres não numéricos.

- Dados Financeiros: As variáveis monetárias `VlrPcpPrvArr`, `VlrTotPvrArr`, `VlrTotPagArr`, `VlrTotDifPvrPagArr` e `VlrSelic`, originalmente formatadas no padrão brasileiro, foram tratadas e convertidas para *float*.

- Dados Categóricos: As colunas textuais `AnmArrecadacao`, `SigNomAgente`, `NumAutoInfracao`, `DscSituacaoArrecadacao` e `DscSituacaoCredito`foram normalizadas para *string*, e a coluna `NomEmpreendimento`, devido à alta cardinalidade e baixa relevância para o modelo de inadimplência, foi removida.

2.2 Tratamento de Inconsistências Temporais e Atrasos

A variável `QtdDiasEmAtraso` apresentou inconsistências, contendo valores negativos ou nulos onde logicamente deveria haver atraso. O tratamento aplicado foi:

1. Recálculo de Dias: Foi calculada a diferença entre a data de pagamento e a data de vencimento.

2. Correção Lógica: Nos casos onde o cálculo diferia do valor registrado, a coluna foi atualizada. Valores negativos (pagamentos antecipados) foram truncados para zero.

3. Tratamento de *Outliers*: Para mitigar o efeito de *outliers* extremos na distribuição temporal, aplicou-se um limite superior de 1000 dias na variável de atraso.

2.3 Normalização e Imputação de Valores Financeiros

Detectou-se uma discrepância significativa entre os valores previstos `VlrTotPvrArr` e os valores efetivamente pagos `VlrTotPagArr`, incluindo registros nulos e valores negativos inconsistentes. As seguintes correções foram aplicadas:

- Recuperação de Valores Nulos: Valores ausentes no total pago foram imputados calculando a diferença entre a coluna de diferença de valor `VlrTotDifPvrPagArr` e o valor previsto.

- Correção de Outliers: Foi calculado o percentual de excesso entre o valor pago e o previsto. Esse percentual foi limitado ao intervalo $[-1.0, 1.0]$ para conter distorções extremas. O valor pago foi então recalculado com base nesse percentual ajustado, garantindo consistência aritmética entre as colunas.

- Valores Negativos: Registros de pagamento que permaneceram negativos após o cálculo foram convertidos para seus valores absolutos ou zerados, conforme o contexto da transação.

2.4 Padronização Categórica e Regras de Negócio

A coluna `DscSituacaoArrecadacao` possuía múltiplas categorias redundantes (e.g., "Quitada pela conciliação", "Quitação manual"). Estas foram mapeadas para categorias macro: 'quitada', 'pendente', 'cancelada' e 'restituída'.

Para a variável alvo `DscSituacaoCredito`, que apresentava dados ausentes (missing values), utilizou-se inferência baseada em regras de negócio:

- Se não houve pagamento e a situação estava nula, o registro foi imputado como `Inscrição em Dívida Ativa`.

- Se houve atraso, mas o valor pago cobriu o previsto, ou se o atraso foi igual a zero com pagamento efetuado, a situação foi imputada como `Quitada`.

- O texto final desta coluna foi normalizado para letras minúsculas e categorias simplificadas.

Ao final do processo, os registros restantes contendo valores nulos em colunas críticas, cuja imputação não foi possível via lógica de negócio, foram removidos do *dataset*, resultando em uma perda mínima de dados.

2.5 Engenharia de Atributos

Para enriquecer a análise temporal, foram derivadas novas variáveis:

- Decomposição de Data: A competência de arrecadação foi desmembrada em `AnoArrec` e `MesArrec`.
- Trimestralidade: Criou-se a coluna `TrimestreVencimento` baseada na data de vencimento do título.
- Flags de Status: Foram geradas variáveis binárias indicadoras: `fatura_paga`, `fatura_atrasado` e `fatura_nao_paga`.

A análise exploratória e os testes de hipóteses foram conduzidos utilizando as bibliotecas NumPy e SciPy para o tratamento estatístico, com suporte de visualização e manipulação de dados. Na etapa de modelagem preditiva, empregaram-se algoritmos implementados nas bibliotecas Scikit-Learn e XGBoost. Adicionalmente, utilizaram-se as ferramentas de pré-processamento, construção de pipelines e validação cruzada nativas do ecossistema Scikit-Learn.

2.6 Consultas

Para aplicação da Análise exploratória foi criado uma coluna que indica a proporção efetivamente paga.

Após isso os dados foram convertidos para o formato tidy para serem utilizados para consultas com duckdb.

As consultas avaliada foram:

- Tendência temporal e médias móveis: Analisa como o comportamento de atrasos e pagamentos evolui mensalmente, utilizando médias móveis.
- Correlação e variabilidade: Verifica estatisticamente se existe uma correlação entre o tempo de atraso e a proporção efetivamente paga de um título, além de medir a variância desses dados
- Análise por quintis de valor previsto: Investiga se o valor do título `VlrTotPvrArr` influencia o risco de inadimplência total
- Ranking e suavização de atraso médio: Testa qual faixa de valor (quintil) tende a demorar mais para pagar, gerando um ranking e suavizando a média.
- Análise sazonal e variação trimestral: Busca identificar mudanças bruscas ou padrões sazonais entre trimestres consecutivos. os padrões analisados são: atraso médio, percentual medio de pago, atraso previsto, pagamento previsto suavizações e média.

2.7 EDA

Nessa etapa, descrevem-se os procedimentos adotados para a realização da análise exploratória dos dados (EDA).

As análises conduzidas para o estudo do conjunto de dados foram:

- Análise univariada: Avaliação da frequência da quantidade de dias de atraso.

- Análise bivariada: Aplicação do teste de correlação de Spearman entre as variáveis quantidade de dias em atraso e percentual pago.
- Análise multivariada: Aplicação do teste de correlação de Spearman entre todas as colunas presentes no *dataset*.

Os testes de hipótese foram realizados com base nas hipóteses previamente definidas, estabelecendo um caso base e verificando sua aceitação ou rejeição. Os testes aplicados foram os seguintes:

H1: Existe uma associação negativa entre o tempo de atraso e o valor pago.
Método de análise: Correlação de Spearman.
H_0: Não existe correlação ($p = 0$).
H_1: Existe correlação negativa.

H2: Créditos com valores previstos mais elevados apresentam maior probabilidade de resultar em inadimplência total, isto é, de não receberem nenhum pagamento.
Método de análise: Teste U de Mann-Whitney.
H_0: Não há diferença entre as distribuições.
H_1: O valor previsto é menor para as faturas pagas do que para as não pagas.

H3: O valor total previsto para arrecadação exerce influência sobre o número de dias em atraso, indicando que montantes maiores podem estar associados a atrasos mais longos.
Método de análise: Correlação de Spearman.
H_0: Não existe associação.
H_1: Existe uma associação significativa entre o valor total previsto e o número de dias em atraso.

2.8 Modelagem

A etapa de classificação foi conduzida utilizando, como modelo base, o algoritmo de *Logistic Regression*. Para fins de comparação de desempenho, também foram empregados os modelos *XGBoost* e *Random Forest*.

Para a execução dessa tarefa, foi necessário remover as seguintes colunas do conjunto de dados: `fatura_atrasado`, `fatura_paga`, `fatura_nao_paga`, `QtdDiasEmAtraso`, `VlrTotPvrArr`, `VlrTotPagArr`, `VlrTotDifPvrPagArr`, `DscSituacaoArrecadacao`, `DscSituacaoCredito`, `Codcvnarr`, `NumCPFCNPJ`, `DatIncidenciaMultaMora`, `DatVencimentoTitulo`, `DatGeracaoConjuntoDados`.

A variável *target* considerada nesta etapa foi `fatura_atrasado`.

2.8.1 Prever a diferença entre o valor previsto e o valor efetivamente pago

Para esta tarefa de regressão, foi utilizado como baseline o modelo de Linear Regression, e comparado com o XGBoost Regressor. Nessa etapa foi necessário trabalhar um uma amostra de 20% dos dados pois o modelo não estava conseguindo executar com a porção total de dados. O modelo RandomForest, testado inicialmente, foi descartado devido ao tempo excessivo de execução.

Para evitar data leak, foram removidas as seguintes colunas: `VlrTotDifPvrPagArr`, `VlrTotPvrArr`, `VlrTotPagArr`, `fatura_paga`, `fatura_atrasado`, `fatura_nao_paga`, `NumCPFCNPJ`, `DatGeracaoConjuntoDado` `DatIncidenciaMultaMora`, `AnoArrec`, `MesArrec`, `TrimestreVencimento` e `prop_pago`.

A variável *target* considerada nesta etapa foi `VlrTotDifPvrPagArr`

2.8.2 Prever a quantidade de dias em atraso

A terceira modelagem teve como objetivo prever a variável alvo `QtdDiasEmAtraso`. Os modelos avaliados foram a Linear Regression baseline e o XGBoost Regressor.

As colunas removidas para garantir a integridade do teste foram: `VlrTotPagArr`, `VlrTotDifPvrPagArr`, `prop_pago`, `fatura_paga`, `fatura_atrasado`, `fatura_nao_paga`, `DscSituacaoCredito`, `DscSituacaoArreca` `DatIncidenciaMultaMora`, `DatGeracaoConjuntoDados`, `AnoArrec`, `MesArrec`, `NumCPFCNPJ` e `QtdDiasEmAtr`

A variável *target* considerada nesta etapa foi `QtdDiasEmAtraso`

3 Resultados

Nesta seção, apresentamos os resultados obtidos a partir da análise exploratória, testes de hipóteses e modelagem preditiva.

3.1 Análise Estatística e Teste de Hipóteses

A investigação inicial buscou compreender as relações entre atrasos, valores monetários e a efetividade do pagamento.

3.1.1 Relação entre Tempo de Atraso e Pagamento H1

A primeira hipótese avaliou se o tempo de atraso influencia a proporção do valor pago. Utilizamos o teste de correlação de Spearman, obtendo um coeficiente $\rho = -0.2956$ com valor-p estatisticamente significativo ($p < 0.05$).

- Conclusão: Existe uma correlação negativa moderada. Quanto maior o número de dias em atraso, menor tende a ser o percentual da dívida liquidado pelo contribuinte.

3.1.2 Influência do Valor Previsto na Inadimplência Total H2

A segunda hipótese testou se valores previstos mais altos `VlrTotPvrArr` estão associados a uma maior probabilidade de inadimplência total. O teste de Mann-Whitney U resultou em um valor-p de 1.0.

- Conclusão: Não rejeitamos a hipótese nula (H_0). Não há evidências estatísticas de que a distribuição dos valores previstos difira significativamente entre faturas pagas e não pagas.

3.1.3 Variação do Atraso por Faixa de Valor (H3)

Ao analisar a relação entre o valor previsto e os dias de atraso, identificamos uma correlação fraca, mas estatisticamente significativa. A segmentação por quartis revelou que faturas de menor valor 1º e 2º quartis tendem a apresentar médias de dias em atraso superiores às faturas de alto valor 4º quartil. Isso acontece devido a distribuição dos dados onde a uma quantidade significativamente maior de dados no primeiro e segundo quartil.

3.2 Modelagem Preditiva

Dois tipos de modelos foram avaliados nesse etapa, modelos de classificação e regressão. A seguir são apresentados os resultados dos experimentos.

3.2.1 Classificação de Atraso na Fatura

O objetivo foi prever a variável binária de atraso. Três algoritmos foram testados: Regressão Logística, Random Forest e XGBoost.

Tabela 1: Comparativo de Desempenho - Classificação de Atraso

Modelo	**Precisão**	**Recall**	**F1-Score**
Logistic Regression	0.91	0.76	0.82
Random Forest	0.98	0.92	0.95
XGBoost	0.98	0.85	0.90
XGBoost (Tuned)	0.99	0.85	0.91

3.2.2 Prever a diferença entre o valor previstro e o valor efetivamente pago

Nessa etapa o objetivo foi prever a variável `VlrTotDifPvrPagArr` que é calculada `VlrTotPagArr - VlrTotPvrArr`. Dessa forma temos valores negativos para contas que não foram totalmente pagas e valores maiores que 0 para taxas em que o valor final foi diferente ou houve juros.

A Tabela 2 apresenta o desempenho comparativo dos modelos. O *LinearRegression* foi estabelecido como *baseline*. Nota-se uma melhora progressiva no coeficiente de determinação (R^2) e uma redução no erro com o uso do *XGBRegressor* e sua posterior otimização.

Tabela 2: Comparativo de desempenho dos modelos preditivos

Modelo	**MSE**	**R^2**
LinearRegression (Baseline)	2.75×10^{30}	0.0037
XGBRegressor (Default)	2.60×10^{30}	0.0565
XGBRegressor (Tuning)	$\mathbf{2.56 \times 10^{30}}$	**0.0709**

Para validar a generalização do melhor modelo, aplicou-se a validação cruzada no *XGBRegressor* otimizado. Os resultados demonstraram consistência, conforme observado na Tabela 3.

Tabela 3: Resultados da Validação Cruzada (XGBRegressor Tunado)

Métrica	**Valor**
MSE Médio (CV)	2.49×10^{30}
Desvio Padrão (CV)	2.72×10^{29}

3.2.3 Previsão de Dias de Atraso (Regressão)

Para estimar a quantidade exata de dias de atraso `QtdDiasEmAtraso`, testamos Regressão Linear e XGBoost Regressor.

A Tabela 4 detalha a evolução do desempenho. Diferente do experimento anterior, nota-se aqui um ajuste muito superior do modelo aos dados. O *XGBRegressor* otimizado atingiu um R^2 de aproximadamente 0.80, reduzindo o Erro Quadrático Médio (MSE) para menos da metade em comparação ao *baseline*.

Tabela 4: Comparativo de desempenho na previsão de dias de atraso

Modelo	**MSE**	**R^2**
LinearRegression (Baseline)	7666.39	0.5625
XGBRegressor (Default)	7038.85	0.5983
XGBRegressor (Tuning)	**3353.35**	**0.8086**

A validação cruzada confirmou a robustez do modelo tunado. Abaixo na Tabela 5 é apresentado os resultados do desempenho alcançado aplicando validação cruzada:

Tabela 5: Resultados da Validação Cruzada (XGBRegressor Tunado)

Métrica	**Valor**
MSE Médio (CV)	3752
Desvio Padrão (CV)	31.58

3.3 Fatores Determinantes

A análise de importância das variáveis (*feature importance*) permitiu identificar quais atributos exerceram maior influência nas decisões dos modelos em cada um dos cenários propostos:

1. Classificação de Atraso: Para determinar se haverá atraso, as variáveis mais importantes foram o identificador do agente e o ano. A variável `SigNomAgente` apresentou importância de 0.017, seguida por `AnoArrec` com 0.0003.

2. Previsão da Diferença de Valores: Na tentativa de prever a discrepância entre o valor previsto e o pago, notou-se que nenhuma variável apresentou preponderância significativa. A variável mais influente foi o ano de arrecadação `AnmArrecadacao`, porém com um coeficiente muito baixo 0.0057. Isso corrobora os resultados obtidos anteriormente, onde o modelo apresentou dificuldade em explicar a variância dos dados.

3. Estimativa de Dias de Atraso: Diferente do cenário anterior, para estimar a quantidade de dias em atraso `QtdDiasEmAtraso`, três variáveis mostraram alto poder explicativo: `AnmArrecadacao` 4.0262, `Codcvnarr` 2.1947 e `SigNomAgente` 0.9820. As demais variáveis tiveram impacto praticamente nulo no modelo.

3.4 Discussão Consolidada

As análises realizadas foram influenciadas pelas inconsistências presentes no dataset, especialmente a elevada quantidade de taxas em atraso e a presença de taxas com valores destoantes da média. Ainda assim, foi possível identificar padrões relevantes. Os testes de hipótese mostraram que o tempo de atraso possui relação negativa com o percentual efetivamente pago, confirmando H1, enquanto não foram encontradas evidências de que valores previstos mais altos aumentem a inadimplência total, resultando na não rejeição de H2. Já para H3, observou-se uma associação estatisticamente significativa, porém fraca, entre o valor previsto e os dias de atraso, influenciada sobretudo pela concentração de valores nos menores quartis.

Na etapa de modelagem, verificou-se um ganho consistente ao utilizar modelos mais complexos em comparação aos baselines simples, especialmente nas tarefas de classificação. Entretanto, algumas previsões como a diferença entre o valor previsto e pago apresentaram baixo poder explicativo, sugerindo que fatores externos ao dataset podem exercer papel relevante.

De modo geral, apesar das limitações identificadas, os resultados obtidos indicam que técnicas de análise estatística e de modelagem preditiva podem oferecer suporte valioso ao monitoramento do comportamento de pagamento das taxas de fiscalização, contribuindo para estratégias mais eficazes de gestão e arrecadação.

Referências

[1] Agência Nacional de Energia Elétrica. SIGEC - Sistema de Gestão de Créditos, 2023. Acessado em: 2025-10-14.