

Score de Pagamento Espontâneo para Direcionamento de Ações de Cobrança na CPFL Energia^{*}

Dennis Eduardo, Douglas Akassaka, Lídia Gusmão e
Hugo Helito

*Centro Analítico, CPFL Energia, Campinas-SP,
(e-mail: dennis.eduardo@gmail.com, {douglasakassaka, lguimaraes,
hugoh}@cpfl.com.br})*

Abstract: Payment delinquency stands for a major financial issue for distribution utilities, since debit collection actions are costly and their profit margins, reduced when compared to recovered amounts. Hence, it is essential to guarantee efficient collection processes, in order to reduce financial losses and increase revenue recovery, leading to operational gains. For that regard, this paper proposes the implementation of a spontaneous payment score for those customers with overdue bills, which indicates their probability to pay their debts spontaneously, in other words, without recharging actions to be applied over them. This indicator was created through a prediction model implemented directly from databases, with no need for specific softwares, and validated by an A/B testing, making use of charging, billing and payment data from customers for its consolidation. In feature selection stage, the most relevant variables were selected through other models application, and the best results were obtained through Random Forest. The predictive model designed for score calculation has obtained a R^2 value of 83% and it has proposed an efficient and appropriate approach to its application dynamics, which require a simpler solution concerning the demanded processing.

Resumo: A inadimplência prejudica financeiramente as distribuidoras de energia de forma significativa, uma vez que os processos de cobrança são custosos e suas margens de lucro, pequenas, frente aos volumes arrecadados. Sendo assim, garantir a eficiência nos processos de cobrança é essencial para reduzir as perdas financeiras e ampliar a recuperação de receita, garantindo ganhos operacionais. Esse trabalho propõe a implementação de um *score* de pagamento espontâneo para os consumidores com faturas em atraso, que indica sua propensão de quitar as dívidas espontaneamente, ou seja, sem receber qualquer ação de cobrança. Através desse índice, é possível otimizar a seleção dos mecanismos de cobrança, direcionando-os para grupos de clientes com menor probabilidade de pagarem suas faturas de forma espontânea. Esse índice foi criado através de um modelo preditivo capaz de ser implementado diretamente a partir do banco de dados, sem necessidade de software específico, e validado por um Teste A/B, fazendo uso de dados de cobrança, faturamento e pagamento dos consumidores para sua consolidação. A modelagem preditiva obteve uma aderência aos dados de 83% para cálculo de espontaneidade e propõe uma abordagem eficaz e adequada à dinâmica de aplicação, que requer uma solução mais simplificada no que diz respeito ao processamento exigido.

Keywords: Payment Delinquency, Debt Collection, Spontaneous payment, Logistic Regression, Linear Regression, Score, Data Mining.

Palavras-chaves: Inadimplência, Recuperação de Receita, Pagamento espontâneo, Regressão Logística, Regressão Linear, *Score*, Mineração de Dados, Ação de cobrança.

1. INTRODUÇÃO

Na conjuntura em que se estrutura o mercado de energia, processos de faturamento são permeados por diversas diretrizes regulatórias, que garantem, entre outras deliberações, que os direitos dos consumidores sejam desempenhados – a exemplo no que diz respeito a definição de prazos para vencimento e garantias a serviços e atendimentos – e

que tributações e tarifas sejam homologadas, as quais estabelecem a operação de sistemas de cobrança ao redor do país (Resolução ANEEL nº 414/2010). Embora normatizadas, as distribuidoras enfrentam diferentes demandas relativas a processos de faturamento e cobrança, dado o caráter multiforme de sua estruturação e regulamentações. Tendo em vista que as contas recebíveis compõem o principal pilar do sistema de arrecadação de uma distribuidora, é evidente que inadimplência e riscos de crédito representam enfoques fundamentais para garantir a saúde financeira da companhia.

^{*} Este estudo foi financiado e realizado no âmbito do projeto de P&D ANEEL PD-00063-3037/2018.

Dessa forma, frente aos representativos impactos econômicos que permeiam esse cenário, o desenvolvimento de mecanismos de arrecadação e cobrança apropriados é fundamental para a efetividade não apenas do sistema financeiro, como também operacional da distribuidora – através da redução das taxas de inadimplência e otimização de ações de recuperação de receita. É importante destacar que ações de cobrança mais precisas permitem a mitigação de tempo, esforços e custos perdidos, de tal forma que diversos procedimentos internos possam ser beneficiados por uma operacionalidade eficiente – o que implica na melhor qualidade dos serviços prestados pela companhia. Um mecanismo relevante para direcionar as ações de cobrança e garantir sua maior efetividade está na correta identificação dos consumidores que devem ser preferencialmente cobrados em detrimento daqueles mais propensos a pagarem suas faturas de forma espontânea, ou seja, sem sofrerem ações de cobrança.

Nessa perspectiva, o estudo apresentado nesse artigo consiste em uma proposta para direcionamento de ações de cobrança – testada e validada na CPFL Energia – baseada na criação de um *score* de pagamento espontâneo para os consumidores, responsável por indicar a probabilidade de que paguem suas faturas sem que ações de cobrança sejam a eles dirigidas. Esse conceito propõe que consumidores com *scores* mais altos tenham menor prioridade de serem cobrados, o que garante que se evite custos dispensáveis para essa finalidade – dado que as ações de cobrança contam com recursos definidos para sua execução e devem ser selecionadas de forma a garantir precisão em sua aplicação. Esse mecanismo foi implementado através de algoritmos de *Machine Learning*, que consolidam o valor do *score* baseando-se em dados granulares, isto é, relativos a cada unidade consumidora. Essas informações, entre elas registros de faturamento, cobrança e pagamento, foram estruturadas a partir de técnicas de mineração de dados.

Em um contexto em que o mercado de energia passa por decisivas mudanças estruturais com a incorporação de novas tecnologias, os paradigmas comerciais são significativamente afetados. A disseminação de recursos energéticos distribuídos, Jiayi et al. (2008), medidores inteligentes, Depuru et al. (2011) e a expansão na abertura do mercado de energia, Magalhães (2009), Streimikiene and Siksnelyte (2014), preveem a adequação na forma de se comercializar energia. Integrar soluções *Data-driven* – movidas a dados – em operações diversas no setor de energia, com ênfase aqui naquelas relacionadas a processos de cobrança, corresponde a uma reestruturação primordial na garantia da proteção de receita e da consolidação da companhia no mercado, Brynjolfsson et al. (2011). A aplicação apresentada nesse trabalho faz uso desses mecanismos para aprimorar a efetividade dos sistemas de cobrança, com expressiva relevância dentro do método tradicionalmente aplicado na CPFL Energia, distribuidora responsável pelo fornecimento de energia a cerca de 9,8 milhões de consumidores em mais de 5 mil municípios brasileiros.

No âmbito de inadimplência no setor elétrico, o estudo desenvolvido em Araújo et al. (2014) propôs uma investigação estatística para seleção de variáveis comerciais relacionadas aos débitos de consumidores, a fim de direcionar a tomada de decisões nos sistemas de cobrança. Foram iden-

tificadas, a partir da análise de Coeficiente de Correlação de *Pearson* maiores correlações entre inadimplência e as variáveis número de faturas vencidas, montante e tempo da dívida. Já o trabalho descrito em Silva (2012) aborda a identificação de perfis de pagamento de consumidores de energia através de técnicas de Regressão Logística e análise de agrupamento, de forma a investigar a propensão de clientes à adesão ao financiamento de Crédito Direto ao Consumidor. Uma análise direcionada para seleção de ações de cobrança a consumidores inadimplentes, fazendo uso de métodos de Regressão Logística, foi proposta em Araújo et al. (2016). A abordagem apresentada pela análise de pagamento espontâneo, embora não objetive selecionar as ações de cobrança, se utiliza do conceito de *score* de probabilidade para classificar os consumidores de acordo com sua propensão ao pagamento, e para isso faz uso de variáveis comerciais diversas e técnicas de Regressão Logística para classificar. Após validar essa classificação se buscou desenvolver um *score* simples que pudesse ser calculado diretamente em SQL, avaliando a aderência dos resultados por meio de um modelo de Regressão Linear. Um dos objetivos principais foi simplificar a abordagem tradicionalmente empregada para esse fim, de forma a entregar uma solução ágil e que exigisse menor processamento para viabilizar sua aplicação diária. A validação dos resultados obtidos foi executada através de um teste sobre o cenário prático de seleção de cobranças.

O trabalho aqui contemplado foi desenvolvido pelo Centro Analítico da CPFL Energia – departamento da companhia designado à criação de valor a partir da integração e análise de dados, propondo soluções baseadas em técnicas de Ciência de Dados e *Machine Learning*. A estruturação do histórico de dados para cada consumidor, a consolidação de seus *scores* de pagamento espontâneo, assim como os resultados obtidos, são descritos nas próximas seções.

2. PAGAMENTO ESPONTÂNEO

A definição do conceito daquilo que se entende por pagamento espontâneo foi inicialmente estruturada nesse trabalho, para que se pudesse trabalhar sobre uma fundamentação consolidada. Mensalmente, as distribuidoras emitem para o consumidor uma fatura para a quitação da prestação do serviço de distribuição de energia elétrica. Quando ele deixa de pagar essa fatura em sua data de vencimento, ele é qualificado como “em atraso”, situação na qual taxas legalmente previstas são aplicadas a título de juros e mora. Essas faturas que não foram pagas na data prevista são classificadas de acordo com circunstâncias diversas, entre elas valor do débito e idade do atraso, as quais determinam as diferentes ações de cobranças às quais podem ser submetidos.

Perante o levantamento de diferentes hipóteses, como a data ou período em que o pagamento foi efetivado, ou recorrência daquilo que se denota espontâneo, foi possível sintetizar uma definição para o termo “espontaneidade de pagamento”. Restituições dessa natureza são aquelas que não têm como principal motivador uma ação de cobrança, mas outros fatores particulares à realidade do consumidor, como seu fluxo de recebimento, rotina, disponibilidade de recurso, dentre outros. A distinção entre uma fatura paga de forma espontânea e uma que não tem essa característica

é importante, já que a fatura espontânea não responde à ação de cobrança (que implicaria em um dispêndio desnecessário), enquanto a fatura que não tem essa característica só será paga mediante cobrança. Como a possibilidade de cobrança é custosa e insuficiente para todos, se faz necessário discriminar esses dois grupos para otimizar os recursos e ser mais assertivo nas ações com os clientes inadimplentes.

3. AÇÕES DE COBRANÇA

Previamente a esse estudo, a estruturação do direcionamento das ações de cobrança se dava com base no sistema de réguas de cobrança. As réguas de cobrança consistem em um sistema de agrupamento dos consumidores inadimplentes a partir do caráter da dívida, com base em dois critérios principais: valor devido e idade da dívida. É importante destacar que uma ação de cobrança é classificada como efetiva quando ela é seguida (dentro de períodos de tempo pré-definidos) pelo pagamento da dívida, sugerindo que ele foi efetuado como consequência da cobrança.

Tendo em vista a distribuição das unidades consumidoras dentro das réguas, ações de cobrança específicas são direcionadas a cada grupo, com periodicidade diária, conforme critérios pré-estabelecidos para as réguas. O êxito das ações de cobrança é contabilizado de forma individual, de acordo com o tempo de pagamento da fatura contado a partir execução da cobrança para cada tipo, em outras palavras, cada tipo de ação de cobrança tem um prazo de pagamento após ser feita pra que seja considerada efetiva. Os diferentes tipos de ações de cobrança executadas pela empresa estão listados a seguir.

- (1) Negativação;
- (2) Ação da cobradora;
- (3) Cobrança via carta;
- (4) Aviso de pré-negativação via URA (unidade de resposta audível);
- (5) Corte de energia;
- (6) Aviso de corte via URA (unidade de resposta audível);
- (7) Recorte de energia;
- (8) Desligamento.

Com base em evidências a respeito da redução na efetividade dessas ações no contexto das réguas de cobrança, a nova abordagem nesse trabalho apresentada foi proposta como forma de aprimorar a seleção dos consumidores inadimplentes aos quais as ações de cobrança são destinadas.

4. METODOLOGIA

4.1 Mineração de Dados

A preparação dos dados a serem trabalhados na análise descritiva e nos modelos preditivos que compuseram a estruturação do *score* demandou a junção de informações disponibilizadas em diferentes bancos de dados internos. Registros comerciais foram extraídos de tabelas diversas, as quais foram conectadas de forma a constituir um cenário histórico com dados dos consumidores: identificação da UC (unidade consumidora); dados do cliente tais como faturamento, datas de pagamento, ações de cobrança, dias

em atraso e existência de valores em PDD (Provisão para Devedores Duvidosos).

Essa estrutura foi filtrada de forma a retornar apenas unidades que possuíssem alguma fatura referente a pagamentos em atraso no período de 01/01/2018 a 30/11/2018. Foram considerados consumidores do grupo B de tensão, das classes residencial, industrial, comercial e rural e que fizeram uso de qualquer canal de pagamento dentre aqueles disponíveis. O tratamento dos dados foi feito de forma granular e sobre um massivo volume de dados. Isso exigiu a aplicação de metodologias de extração, carregamento e transformação, Vassiliadis (2009) para a manipulação e lapidação das informações que foram estruturadas para análises descritivas e modelos preditivos, cujos resultados são apresentados na seção 5. A base final para o período considerado continha um total de 4.832.227 registros, cada qual correspondente a uma UC com histórico de pagamento em atraso. Cada um desses registros foi rotulado como espontâneo (1) ou não (0) considerando se aquele pagamento em atraso ocorreu devido uma ação de cobrança ou se foi realizado independente dela, de acordo com os critérios descritos para cada metodologia, como apresentado em 4.2.

4.2 Modelo Preditivo

Para predição do *score* de pagamento espontâneo dos consumidores inadimplentes, foram testadas duas diferentes abordagens de *Machine Learning* empregadas em problemas de classificação, implementadas, nesse estudo, em *Python* com aplicação de funções do pacote *scikit-learn*. A primeira delas, empregou técnicas robustas de Regressão Logística, Hosmer Jr et al. (2013) para predição se o cliente paga ou não espontaneamente. Em seguida, objetivando simplificar não apenas a implementação, como também os esforços computacionais do modelo, foi desenvolvido um segundo algoritmo empregando técnicas de Regressão Linear, Seber and Lee (2012).

Feature Selection: As bases de dados que serviram como entrada na etapa de *feature selection* eram compostas por 27 variáveis explicativas. O objetivo, nessa etapa, foi aplicar a técnica de Floresta Randômica com a finalidade de retornar o índice de explicabilidade de cada variável, o qual expressa o quanto ela contribui para a predição assertiva do *score* de pagamento espontâneo, (Saeys et al. (2008) explana sobre o uso desse modelo para esse propósito). Os resultados dessa seleção permite então que se realize um filtro sobre essas características, de forma a serem utilizadas apenas aquelas que favorecem o sucesso da predição.

Inicialmente, os dados selecionados para o teste do modelo compuseram 50% da base analítica. Na aplicação de Floresta Randômica, foram definidos os valores para os hiper-parâmetros quantidade de estimadores, quantidade máxima de características (*features*) e número mínimo de folhas de cada árvore de decisão da floresta. Sendo assim, os pesos que estruturam os ramos das diferentes “árvores de decisão aleatórias” que compõem a floresta, foram definidos e adequados até que se obtivesse o melhor ajuste, retornando as variáveis mais explicativas e seu nível de importância para a predição.

Regressão Logística: Após realizada a etapa de *feature selection*, em que foram selecionadas somente as variáveis mais relevantes, o algoritmo preditivo foi modelado a partir da técnica de Regressão Logística. Nesse algoritmo, as variáveis independentes (X_k) se referem aos dados explicativos do consumidor e a dependente pelo seu *score* S_{esp} (representada pela variável $Y_{score_{(log)}}$, resultado da transformação logarítmica da metodologia), conforme a equação 1. Os parâmetros da função no *Python* foram ajustados de forma a adequar da melhor forma os coeficientes b_k para aprimorar as taxas de sucesso do modelo considerando critérios de negócio pré-definidos.

$$Y_{score_{(log)}} = b_0 + b_1 \cdot X_1 + \dots + b_N \cdot X_N \quad (1)$$

A variável resposta S_{esp} foi consolidada na base de dados a partir da informação de espontaneidade de pagamento definida: Consumidores com registros de faturas pagas sem ação de cobrança recebiam a *tag* 1 - espontâneo e, no caso contrário, a *tag* 0 - não espontâneo. 50% da base foi utilizada no treino do modelo e 50% no teste. O objetivo dessa modelagem supervisionada foi utilizar os dados históricos previamente classificados no treino e no teste para, quando fosse aplicado em dados reais ainda não classificados, estimasse um valor entre 0 e 1 para seu *score* de espontaneidade. O valor usado para corte (*threshold*) no momento da classificação foi de 0,5, o que significa que consumidores com $S_{esp} < 0,5$ (probabilidade menor que 50%) seriam classificados como não espontâneos e, conseqüentemente, selecionado para as ações de cobrança. De forma complementar, aqueles com $S_{esp} > 0,5$ (probabilidade maior que 50%) estariam na classe de pagadores espontâneos.

É importante levar em consideração que, durante a etapa de *tuning* dos modelos as regras de negócio que permeiam esse processo. No cenário abordado nesse trabalho, os falsos positivo correspondem aos casos em que o modelo prevê como pagamento espontâneo um consumidor que, na realidade, não pagaria sua fatura espontaneamente, indicando que ações de cobrança não devam ser a ele dirigidas (quando na verdade deviam). Já os falsos negativos se referem àqueles clientes classificados como não espontâneos, sugerindo a atuação de cobrança sobre eles quando, na realidade, eles têm maiores probabilidades de pagarem espontaneamente (o que sugere a aplicação de uma ação de cobrança que não seria necessária). Nesse contexto, é relevante entender as prioridades e necessidades de negócio para definição da “qualidade” do modelo. A minimização de ambas classes (falsos positivos ou negativos) tem significativa importância no estudo, no entanto, foi definida a ênfase na redução dos falsos negativos, de forma a priorizar a recuperação de receita através dos processos de cobrança.

Embora atendessem os objetivos, a robustez do modelo exigia um alto processamento computacional e uma arquitetura de sistemas que se mostraram inviáveis para sua implementação e execução diária no departamento de cobranças. Sendo assim, uma nova abordagem simplificada, através de Regressão Linear, foi implementada para solução desse problema, a qual tinha como principais norteadores a possibilidade de ser executada diretamente no banco de dados (SQL), consumir pouco recurso computacional e ter uma interpretação analítica simples. A metodologia

de Regressão Logística descrita foi, então, utilizada para validar esse novo modelo proposto.

Modelo Simplificado: A abordagem através da técnica de Regressão Linear trabalha, assim como a técnica de Regressão Logística, com o ajuste dos coeficientes b_k , conforme a formulação dessa metodologia (equação 2).

$$S_{esp} = b_0 + b_1 \cdot X_1 + \dots + b_N \cdot X_N \quad (2)$$

Nesse caso, a variável dependente (S_{esp}), que representa o *score*, foi calculada – diferentemente do que foi apresentado para Regressão Logística – a partir da fórmula mostrada na equação 3.

$$S_{esp} = \frac{\text{faturas pagas sem ação de cobrança}}{\text{total de faturas pagas}}, 0 < S_{esp} < 1 \quad (3)$$

O valor do *score* (S_{esp}) representa uma pseudo-probabilidade de pagamento espontâneo e varia de 0 a 1, sendo os valores mais próximos de 0 representados por menores probabilidades de espontaneidade, enquanto que valores próximos de 1 se referem a maiores propensões de pagamentos espontâneos. A fórmula foi proposta com base na premissa de que o histórico do percentual de pagamento espontâneo pode ser um bom balizador do comportamento futuro de cada cliente. Dessa forma, a regressão permite avaliar se o comportamento histórico do S_{esp} é um preditor do comportamento futuro, e ainda se seus valores guardam relação. Para comparar o resultado dessa metodologia com o modelo de Regressão Logística, utilizou-se um corte no S_{esp} , com a finalidade de torná-lo uma variável binária. Dessa forma, foram classificados como 1 valores de *score* acima de 0,5 e como 0, aqueles abaixo de 0,5. De modo a corroborar os resultados obtidos da aplicação do modelo de Regressão Logística, quando utilizou-se a técnica da Regressão Linear modelada a partir das variáveis independentes mais relevantes (X_k) selecionadas na etapa da *feature selection*, pôde-se verificar a aderência na resposta dos dois modelos, de forma que o resultado do modelo simplificado poderia replicar o resultado do modelo de Regressão Logística na maior parte das vezes.

5. RESULTADOS

5.1 Análise Descritiva

A partir das bases de dados consolidadas, análises foram executadas com a finalidade de mapear o cenário de inadimplência. Inicialmente, foi consolidada a distribuição de faturas pagas em atraso dentro dos dias dos meses, de acordo com a Fig. 1.

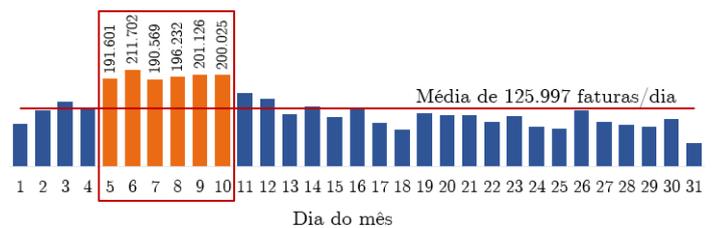


Figura 1. Distribuição do volume de faturas pagas em atraso por dia do mês.

Foi identificado que 35% dos pagamentos realizados em atraso ocorrem de forma recorrente entre os dias 05 e 11 de cada mês, sendo sua média diária (193 mil faturas/dia) 53% superior àquela de todos os pagamentos realizados em atraso (média de 125 mil faturas/dia). Em média, são pagas 3,9 milhões de faturas em atraso por mês, sendo um terço delas (1,6 milhões) pagas entre esses sete dias. Apesar das datas de pagamento em atraso das faturas se concentram próximas ao 5º dia útil de cada mês, sendo 14% superior à média mensal, não se observa diferença de comportamento entre as faturas pagas em atraso de forma espontânea que (que não sofreram ações de cobrança) e as que foram cobradas (que sofreram ações de cobrança) em relação ao dia do mês de seu pagamento.

A Fig. 2 apresenta a distribuição do volume de faturas em atraso pagas por dias de atraso (idade da dívida em dias). Elas foram divididas de acordo com três critérios: sem ação de cobrança (laranja), com ação de cobrança efetiva (cinza) e com ação de cobrança inefetiva (azul). É importante ressaltar que pagamentos em atraso classificados com ação de cobrança inefetiva se referem àqueles que ocorreram fora dos prazos de efetividade definidos para cada ação, o que sugere que a quitação da dívida não foi consequência da cobrança.

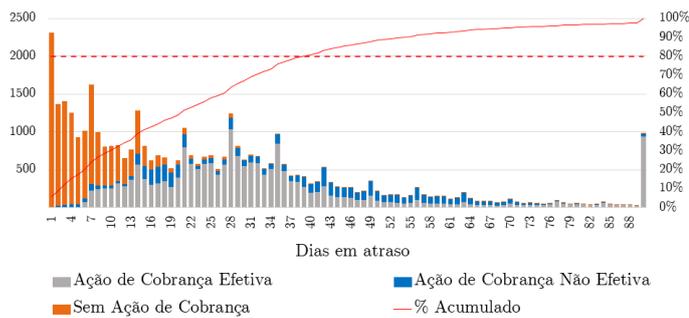


Figura 2. Distribuição do volume de faturas pagas pela idade da dívida em dias.

Analisando a espontaneidade e a efetividade das ações de cobrança, observou-se que 80% das faturas em atraso são pagas em até 39 dias após o seu vencimento, sendo que 99% dos pagamentos sem ação de cobrança ocorrem até o 35º dia de atraso.

Além disso, é possível identificar picos recorrentes de pagamentos de faturas a cada intervalo de 7 dias, os quais acontecem às segundas-feiras – observa-se predominância equivalente a 30% dos pagamentos registrados neste dia da semana –, como mostra a Fig. 3. Nos demais dias, a distribuição dos pagamentos é aproximadamente equilibrada, seguindo uma leve tendência de queda no fim de semana.

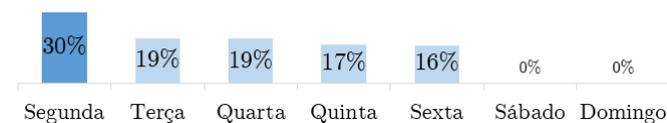


Figura 3. Distribuição do percentual de faturas em atraso pagas por dia da semana.

A participação das faturas pagas em atraso por dias de atraso é apresentada em barras empilhadas na Fig. 4,

para análise mais precisa da concentração das diferentes categorias de pagamento.

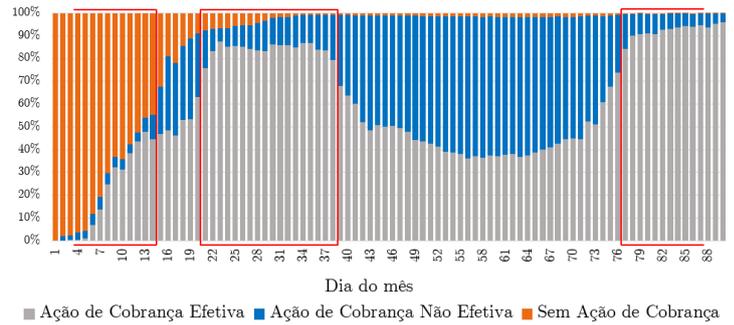


Figura 4. Participação de cada tipo de pagamento por dia de atraso.

A efetividade das ações de cobrança em comparação com a quantidade de dias em atraso, sinaliza uma maior incidência de pagamentos espontâneos (sem ação de cobrança) nos primeiros dias de atraso, indo até o 14º dia de atraso. A partir deste tempo de atraso, há uma queda nítida da espontaneidade nos pagamentos. Já as ações de cobrança possuem dois picos de efetividade, o primeiro se iniciando após 20º dia e o segundo após 77º dia. Verifica-se uma baixa na sua eficiência no intervalo que vai do 39º ao 76º dia de atraso.

No que diz respeito aos tipos de ações de cobrança, descritos na seção 3, a distribuição da efetividade de cada uma delas sobre o total de faturas pagas é mostrada na Fig. 5, assim como o percentual de faturas pagas espontaneamente.

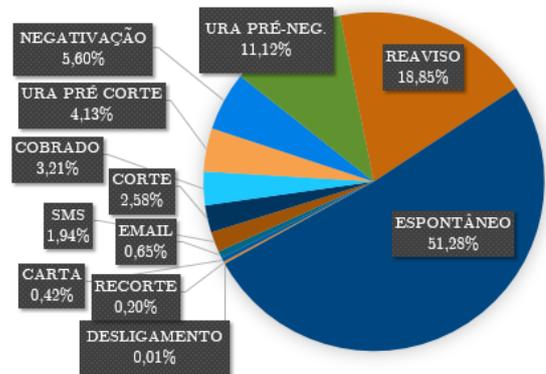


Figura 5. Distribuição da efetividade das ações.

Foi identificada que a maior efetividade está no reaviso, que não configura uma ação em si, mas um atendimento regulatório de aviso de débito apresentado na fatura. As demais ações de maior representatividade, foram a URA pré-negativação, negativação, a URA pré-corte e as cobradoras. As outras ações de cobrança tiveram representatividade inferior a 3%.

5.2 Modelo Preditivo

Feature Selection: Os resultados obtidos para a seleção das características com os maiores índices de explicabili-

dade, extraídos da aplicação de Floresta Randômica, são aqueles apresentados na Fig. 6.

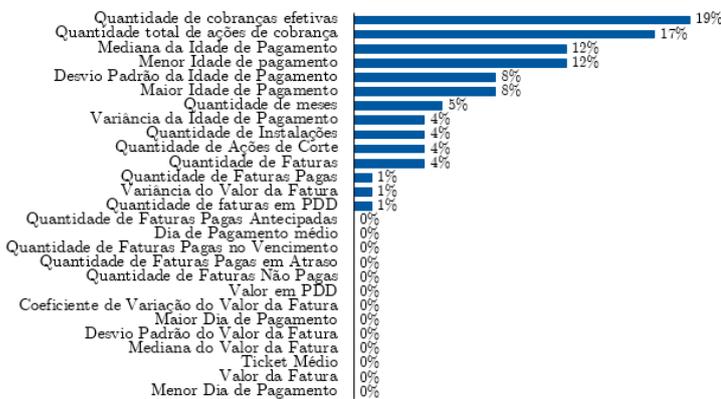


Figura 6. Relevância das variáveis explicativas empregadas.

Dentre as 27 variáveis explicativas inicialmente empregadas nos modelos, apenas 14 apresentaram relevância considerável na predição e foram consideradas para implementação dos modelos, sendo elas:

- Quantidade de faturas analisadas para aquele consumidor;
- Quantidade de faturas pagas por aquele consumidor;
- Quantidade de meses considerados na análise daquele consumidor;
- Quantidade total de ações de cobrança efetuadas para aquele consumidor;
- Quantidade de ações de cobrança efetivas computadas para aquele consumidor;
- Quantidade de instalações registradas no nome daquele consumidor;
- Quantidade de ações de corte executadas em instalações daquele consumidor;
- Mediana da idade de pagamento (cada “idade” corresponde ao tempo que levou para que dada UC pagasse cada fatura após seu período de faturamento);
- Menor idade de pagamento;
- Desvio padrão da idade de pagamento;
- Variância da idade de pagamento;
- Maior idade de pagamento;
- Variância calculada a partir dos valores das faturas daquele consumidor;
- Quantidade de faturas em PDD (Provisão de Devedores Duvidosos) para aquele consumidor;

Predição do Score de Pagamento Espontâneo: As variáveis selecionadas na etapa anterior foram utilizadas no modelo de predição de pagamento espontâneo. O modelo inicialmente empregado de Regressão Logística apresentou uma acurácia e de 76% para previsão de pagamento espontâneo, se mostrando com uma taxa de sucesso para identificar quem necessita de uma ação de cobrança de 81% (verdadeiro negativo), conforme mostrado na matriz de confusão da Fig. 7.

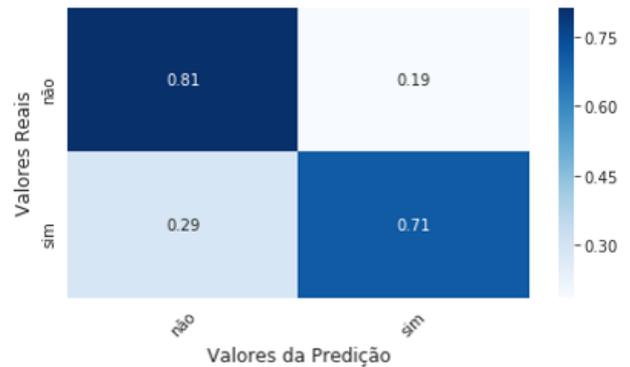


Figura 7. Matriz de confusão do modelo de Regressão Logística.

Já no modelo simplificado, a relação entre o índice de pagamento espontâneo histórico e futuro se comprovou verdadeiro, tendo como resultado P-valor $< 0,000$ e R^2 de 0,83, como mostra a Fig. 8.

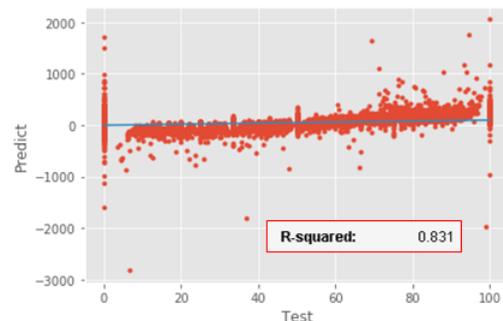


Figura 8. Resultado da Regressão Linear.

Ainda no modelo simplificado, o tempo de processamento levou em torno de quatro horas, que é considerado um tempo razoavelmente bom dado o volumetria de dados processados.

Dada a necessidade de utilizar um modelo de predição com performance mais ágil e adequada à implementação e uso recorrente na área de negócio responsável por faturamento e cobrança, foi empregado para predição do *score* o algoritmo simplificado de Regressão Linear. O valor final atribuído ao corte (*threshold*) foi mais conservador do que aquele utilizado nos testes iniciais: 0,8, em oposição ao de 0,5 anteriormente aplicado, indicando que consumidores com $S_{esp} < 0,8$ (probabilidade de espontaneidade menor do que 80%) seriam selecionados para ações de cobrança, enquanto os demais ($S_{esp} < 0,8$) seriam classificados como pagadores espontâneos. Com isso, os clientes da carteira da distribuidora poderiam ter *scores* calculados diariamente, sem impacto no processamento do servidor, contribuindo assim para tomada de decisão sobre as ações de cobrança e em quais clientes deverão ser realizadas.

Teste do modelo: Com o intuito de validar o *score* de pagamento espontâneo, foi realizado o Teste A/B. Para a realização desse teste, primeiramente foram selecionados 10 mil clientes considerados como espontâneos, ou seja, aqueles que pagariam sem precisar de ações de cobrança,

em seguida eles foram divididos em dois grupos. O primeiro correspondeu aos clientes em que as ações de cobrança não foram realizadas, que foi denominado como grupo de teste. Já o segundo, as ações de cobrança ocorreram normalmente, esse grupo foi denominado como grupo de controle. Os dados coletados dos testes foram do período de abril a julho de 2019, no qual depois foram divididos em dois subgrupos, de quem pagou e não pagou. A fim de validar os dados coletados foi utilizada a razão de chances ou razão de possibilidades, definida como a razão entre a chance ou probabilidade de ocorrência em um grupo e a chance ou probabilidade de ocorrência em outro grupo. A razão de possibilidades teve média obtida de 0,87, como pode ser verificado na Fig. 9, isso indica que a chance de inadimplência no grupo de teste foi menor que no grupo de controle. O esperado era encontrar valores próximos de 1, demonstrando que a chance de não pagamento no período do estudo seria igual nos dois grupos.

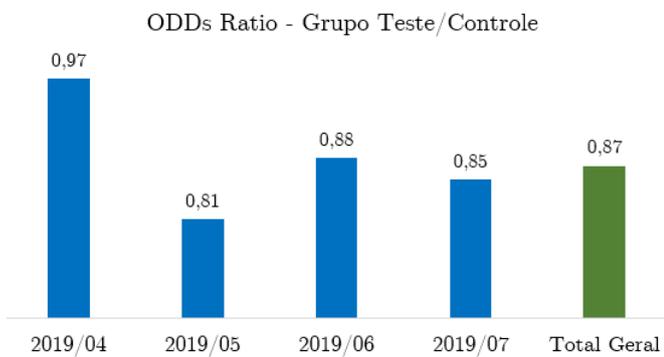


Figura 9. Razão de possibilidades do teste do *score*.

6. CONCLUSÃO

A incorporação de análises de dados em processos de tomadas de decisão no meio corporativo tem se tornado um mecanismo crucial na garantia de melhoras na eficiência operacional, serviços prestados e recuperação de receitas. Com esse propósito, o Centro Analítico – departamento da CPFL Energia financiado por um projeto P&D da ANEEL e destinado a criar valor a partir da integração e análises de dados – propôs a implementação de um modelo preditivo para otimização dos sistemas de faturamento e cobrança, através da identificação dos consumidores mais propensos a pagarem suas faturas de forma espontânea.

A partir de metodologias de *Machine Learning*, foi desenvolvido um *score* de pagamento espontâneo para os consumidores, que corresponde a um índice que indica a propensão de consumidores inadimplentes pagarem seus débitos espontaneamente, ou seja, sem a necessidade de ações de cobrança. Dado que as companhias dispõem de números limitados de cobranças a serem efetivadas, garantir que elas sejam direcionadas de maneira apropriada permite um aumento na sua efetividade e, conseqüentemente, na receita recuperada e eficiência dos processos internos.

Propriamente tratados e selecionados, dados diversos de faturamento e cobrança foram estruturados de forma a servir de entrada para o modelo de predição do *score*, implementado a princípio através da técnica de Regressão Logística, o qual obteve uma acurácia de 76%. Os resulta-

dos alcançados pelos ajustes evidenciaram a possibilidade de simplificação dessa abordagem através de uma modelagem de Regressão Linear. A solução simplificada obtida apresentou aderência aos dados de (R^2) de 83%, que foi validada no cenário prático por meio de um Teste A/B. A partir do desenvolvimento simplificado, foi possível facilitar o processo de seleção das ações de cobrança em relação aos algoritmos tradicionalmente empregados para esse fim. Isso permitiu ganhos de performance e processamento, de acordo com a demanda requerida para aplicações práticas, garantindo resultados eficientes e de aplicação otimizada.

REFERÊNCIAS

- Araújo, R.V., Martinez, L., and Moreira, F.A. (2014). Statistical Analysis of the Relationship Between Payment Behavior Variables and Delinquency in Electricity Consumption. In *2014 IEEE PES Transmission & Distribution Conference and Exposition-Latin America (PES T&D-LA)*, 1–6. IEEE.
- Araújo, R.V., Martinez, L., and Moreira, F.A. (2016). Seleção de Clientes para Ações de Cobrança através de Regressão Logística na Inadimplência do Consumo de Energia Elétrica.
- Brynjolfsson, E., Hitt, L.M., and Kim, H.H. (2011). Strength in numbers: How does data-driven decision-making affect firm performance? *Available at SSRN 1819486*.
- Depuru, S.S.S.R., Wang, L., Devabhaktuni, V., and Gudi, N. (2011). Smart meters for power grid—challenges, issues, advantages and status. In *2011 IEEE/PES Power Systems Conference and Exposition*, 1–7. IEEE.
- Hosmer Jr, D.W., Lemeshow, S., and Sturdivant, R.X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.
- Jiayi, H., Chuanwen, J., and Rong, X. (2008). A review on distributed energy resources and microgrid. *Renewable and Sustainable Energy Reviews*, 12(9), 2472–2483.
- Magalhães, G.S.C. (2009). *Comercialização de energia elétrica no ambiente de contratação livre: uma análise regulatório-institucional a partir dos contratos de compra e venda de energia elétrica*. Ph.D. thesis, Universidade de São Paulo.
- Saeyns, Y., Abeel, T., and Van de Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 313–325. Springer.
- Seber, G.A. and Lee, A.J. (2012). *Linear regression analysis*, volume 329. John Wiley & Sons.
- Silva, A.M.M. (2012). *Técnicas de Data Mining na aquisição de clientes para financiamento de Crédito Direto ao Consumidor-CDC*. Ph.D. thesis, Universidade de São Paulo.
- Streimikiene, D. and Siksnelyte, I. (2014). Electricity market opening impact on investments in electricity sector. *Renewable and Sustainable Energy Reviews*, 29, 891–904.
- Vassiliadis, P. (2009). A survey of extract–transform–load technology. *International Journal of Data Warehousing and Mining (IJDWM)*, 5(3), 1–27.