XXVIII SIMPEP
Simpósio de Engenharia de Produção
"ENSINO EM ENGENHARIA DE PRODUÇÃO: COMO PREPARAR AS NOVAS GERAÇÕES PARA OS DESAFIOS DO SÉCULO XXI"
10, 11 E 12 DE NOVEMBRO DE 2021

BIG DATA E REGRESSÃO LOGÍSTICA APLICADOS À ANÁLISE DE PEDIDOS DE EMPRÉSTIMO

ANDRÉ FONTANEZ BRAVO – andrebravo@usp.br
UNIVERSIDADE DE SÃO PAULO - USP

RENATO DE OLIVEIRA MORAES – remo@usp.br
UNIVERSIDADE DE SÃO PAULO - USP

HUGO MARTINELLI WATANUKI – hugo.watanuki@lexisnexisrisk.com
LEXISNEXIS RISK SOLUTIONS - LNRS

Área: *6 – ENGENHARIA ORGANIZACIONAL*
Sub-Área: *6.4 – GESTÃO DA INFORMAÇÃO*

Resumo: *BIG DATA E SUAS APLICAÇÕES VÊM SE TORNANDO CADA VEZ MAIS PRESENTES E IMPORTANTES EM DIFERENTES ÁREAS DO CONHECIMENTO, DESDE A MEDICINA ATÉ O MUNDO DOS NEGÓCIOS. NESSE CONTEXTO, TÉCNICAS E FERRAMENTAS QUE SEJAM CAPAZES DE PROCESSAR O IMENSO FLUXO DE INFORMAÇÕES DISPONÍVEIS PARA AGREGAR VALOR SÃO INSTRUMENTOS PODEROSOS. O OBJETIVO DESTE TRABALHO É CONSTRUIR UM MODELO DE REGRESSÃO LOGÍSTICA QUE FILTRE OS PEDIDOS DE EMPRÉSTIMO FEITOS NA PLATAFORMA DA EMPRESA LENDINGCLUB, SELECIONANDO AQUELES COM MAIOR PROBABILIDADE DE SEREM PAGOS EM DIA, OU SEJA, AQUELES QUE SERÃO OS MELHORES INVESTIMENTOS. PARA TANTO, FOI UTILIZADA A PLATAFORMA HPCC SYSTEMS, CAPAZ DE PROCESSAR GRANDES VOLUMES DE DADOS. NELA, O ARQUIVO DE DADOS UTILIZADO FOI LIMPO E ORGANIZADO PARA QUE, ENTÃO, FOSSE CONSTRUÍDO O PRIMEIRO MODELO. APÓS ANÁLISE INICIAL, TAL MODELO FOI APERFEIÇOADO POR DOIS MÉTODOS DISTINTOS. AO FINAL, FORAM OBTIDOS DOIS MODELOS ÚTEIS PARA A FILTRAGEM DE PEDIDOS DE EMPRÉSTIMO. O PRIMEIRO DELES É MELHOR PARA OBTER UMA FILTRAGEM COM A MAIOR PROPORÇÃO POSSÍVEL DE BONS PAGADORES E O SEGUNDO É MELHOR PARA QUE A FILTRAGEM DESCARTE O MÍNIMO POSSÍVEL DE BONS PAGADORES.*

Palavras-chaves: *BIG DATA; HPCC SYSTEMS; REGRESSÃO LOGÍSTICA; ANÁLISE DE PEDIDOS DE EMPRÉSTIMO.*

BIG DATA AND LOGISTIC REGRESSION APPLIED TO ANALYSIS OF LOAN REQUESTS

Abstract: *BIG DATA AND ITS APPLICATIONS ARE BECOMING MORE AND MORE IMPORTANT ACROSS DIFFERENT FIELDS, FROM MEDICINE TO THE BUSINESS WORLD. IN THAT SENSE, TECHNIQUES AND TOOLS THAT ARE ABLE TO PROCESS THE IMMENSE FLOW OF INFORMATION TO CREATE VALUE ARE POWERFUL INSTRUMENTS. THE GOAL OF THIS PROJECT IS TO BUILD A LOGISTIC REGRESSION MODEL TO FILTER LOAN REQUESTS AVAILABLE AT LENDINGCLUB'S WEBSITE. THE MODEL SHOULD ONLY SELECT REQUESTS WITH HIGH PROBABILITY OF BEING PAID ON TIME, WHICH WOULD MEAN THEY ARE A GOOD INVESTMENT. THE DATASET WAS CLEANED AND THE FIRST MODEL WAS BUILT USING HPCC SYSTEMS, A PLATFORM SPECIALIZED IN DEALING WITH LARGE VOLUMES OF DATA. THIS MODEL WAS THEN ANALYSED AND IMPROVED THROUGH TWO DIFFERENT METHODS, LEADING TO TWO FINAL MODELS. THE FIRST ONE IS BETTER TO FILTER LOAN REQUESTS TO OBTAIN A HIGHER PROPORTION OF GOOD LOANS, WHILE THE SECOND ONE IS BETTER TO FILTER LOAN REQUESTS DISCARDING AS FEW GOOD LOANS AS POSSIBLE.*

Keywords: *BIG DATA; HPCC SYSTEMS; LOGISTIC REGRESSION; ANALYSIS OF LOAN REQUESTS.*

1. INTRODUÇÃO

A análise de quantidades massivas de dados ou Big Data e o desenvolvimento de modelos de predição possuem importantes aplicações em diversas áreas do conhecimento. Segundo Oracle (2021), o Big Data se tornou quase imprescindível no mundo atual para processar constantemente o gigantesco fluxo de informações disponíveis para as empresas. Os avanços tecnológicos recentes possibilitam lidar com um volume de dados maior, mais barato e mais acessível e, a partir disso, é possível tomar decisões de negócios mais precisas.

Por meio de técnicas estatísticas e computacionais, é possível treinar modelos, a partir de condições e resultados passados, para prever resultados futuros, o que é chamado de aprendizado supervisionado (GRUS, 2016). No campo da medicina aplicada ao esporte, por exemplo, Ruiz-Peréz et al. (2021) desenvolveram um modelo para prever o risco de lesão em atletas de futsal a partir de fatores como a qualidade de seu sono, seu controle postural e o impacto psicológico dos resultados de sua equipe.

O objetivo deste projeto foi desenvolver um processo para identificar pedidos de empréstimo que serão bons investimentos por meio de um modelo de regressão logística. Para tanto, foi utilizado o banco de dados da empresa norte-americana LendingClub (YASH, 2020), que oferece empréstimos entre pessoas físicas.

Na plataforma da empresa, um cliente aplica para um empréstimo e outras pessoas financiam partes desse empréstimo. Os financiadores podem, então, receber sua parte de cada parcela do pagamento ou comercializar os títulos de dívida que possuem com outras pessoas.

Conquanto a LendingClub faça uma filtragem inicial dos pedidos de empréstimo, disponibilizando em sua plataforma apenas aqueles cujo risco não é considerado alto demais, cerca de 12,5% dos empréstimos acabam inadimplentes. Evidentemente, investir em um empréstimo que acaba não sendo pago de volta acarreta uma perda financeira. Logo, o modelo desenvolvido neste projeto deve filtrar ainda melhor tais empréstimos, servindo de auxílio na decisão de investimento.

Para treinamento dos modelos, foi usada a base pública de dados da LendingClub (YASH, 2020). Em seu arquivo de dados, estão contidos mais de 2,2 milhões de registros de pedidos de empréstimo aprovados (que ficaram disponíveis na plataforma), com informações sobre renda anual do indivíduo, prazo e taxa de juros do empréstimo, entre outras (totalizando 151 colunas). Tais informações, bem como o resultado final de cada empréstimo (adimplente ou inadimplente), foram utilizadas para desenvolvimento de modelos de regressão logística.

No segundo capítulo deste artigo, é apresentada a revisão da literatura, abordando os principais conceitos que guiaram o desenvolvimento deste trabalho. O terceiro capítulo traz a metodologia utilizada, isto é, como foi organizado o projeto, desde a filtragem e organização dos dados até a construção dos modelos. No quarto capítulo são discutidos os resultados fornecidos pelos modelos construídos, bem como medidas tomadas para melhorá-los. O quinto capítulo, por sua vez, trata das conclusões finais acerca do trabalho desenvolvido e sua importância para análise de empréstimos. Ao final, são listadas as referências bibliográficas consultadas ao longo do trabalho.

2. REVISÃO DA LITERATURA

Neste capítulo, são apresentados conceitos e técnicas essenciais para o desenvolvimento do trabalho. Primeiramente, explica-se o que é Big Data, ideia que engloba o trabalho como um todo. Então, é exposta a técnica de regressão logística, que foi utilizada para a construção dos modelos. Por fim, são abordadas as métricas usadas para avaliação dos modelos.

2.1 Big Data

Segundo Gartner (2021), Big Data são ativos de informação de grande volume, grande velocidade e/ou grande variedade. Tais ativos requerem formas inovadoras de processamento de informação para possibilitarem melhor entendimento do assunto analisado e melhores tomadas de decisão.

Para Oracle (2021), Big Data envolve a integração de dados de diversas fontes diferentes, o gerenciamento ou armazenamento dos dados, e a análise, que permite a obtenção de informações úteis para a atividade a ser desenvolvida. Dentre seus casos de uso, podem-se citar o desenvolvimento de produtos, a manutenção preditiva e o aprendizado de máquina, escopo deste trabalho.

2.2 Regressão logística

Dória (2019) define a regressão logística como um modelo que prevê a variável dependente binária Y (0 ou 1) a partir de n variáveis independentes (X_1, X_2, ..., X_n). A probabilidade de Y = 1 é dada pela equação 1, sendo $f(x) = B_0 + B_1X_1 + ... + B_nX_n$.

$$P(Y = 1) = \frac{1}{1+e^{-f(x)}} \quad (1)$$

Determinam-se os coeficientes B_0, B_1, ..., B_n por meio do método de máxima verossimilhança, o qual maximiza a probabilidade de a amostra ter sido observada. Para problemas de classificação, define-se Y = 1 quando P(Y=1) > 0,5. Caso contrário, define-se Y = 0.

2.3 Indicadores de desempenho

Um método bastante direto para a avaliação do desempenho de um modelo preditivo é a matriz de confusão. Nela, são expostos os quatro possíveis resultados de uma classificação binária:

a) Positivo verdadeiro: o dado era positivo (1) e foi corretamente classificado pelo modelo.
b) Falso positivo: o dado era negativo (0) e foi incorretamente classificado pelo modelo.
c) Negativo verdadeiro: o dado era negativo (0) e foi corretamente classificado pelo modelo.
d) Falso negativo: o dado era positivo (1) e foi incorretamente classificado pelo modelo.

A partir da quantidade de cada um desses resultados, constrói-se uma matriz de confusão, como mostrada no Quadro 1:

QUADRO 1 – Estrutura de uma matriz de confusão.

Estrutura de uma matriz de confusão		
Previsão \ Real	**Positivo (1)**	**Negativo (0)**
Positivo (1)	Positivos verdadeiros	Falsos positivos
Negativo (0)	Falsos negativos	Negativos verdadeiros

Fonte: Elaborado pelos autores (2021).

A partir da matriz de confusão, é possível calcular outros indicadores bastante úteis para a análise de desempenho de classificações binárias. Segundo Balaji, Annavarapu e Bablani (2021), têm-se:

a) Acurácia: eficácia total do modelo (número total de acertos dividido pelo total de previsões).
b) Precisão: proporção das classificações positivas que estavam corretas (positivos verdadeiros divididos pelo total de classificações positivas).

c) Sensibilidade: proporção dos dados positivos classificados corretamente (positivos verdadeiros divididos pelo total de positivos).

Por último, tem-se o Akaike Information Criterion (AIC). Miller e Forte (2017) descrevem o AIC como um indicador que utiliza uma abordagem baseada na teoria da informação para avaliar a qualidade de um modelo, levando em consideração sua acurácia e sua complexidade. Com isso, o índice tenta lidar tanto com o risco de sobreajuste quanto com o risco de sub-ajuste do modelo.

3. METODOLOGIA

Neste capítulo, são detalhados os métodos que foram utilizados ao longo do trabalho. Primeiramente, descreve-se a plataforma usada para a construção dos modelos e sua importância considerando a natureza do conjunto de dados. Então, explicam-se a limpeza e a formatação dos dados para prepará-los para a construção dos modelos. Por fim, é abordado o processo de amostragem para treinamentos e testes dos modelos.

3.1 Plataforma HPCC Systems

Devido à quantidade massiva de registros e colunas (característica comum em aplicações de Big Data), a manipulação dos dados utilizados em um computador comum levaria tempo demais ou seria até impossível, pela insuficiência de memória. Portanto, foi utilizada a plataforma HPCC Systems.

A HPCC Systems é uma plataforma código-aberto desenvolvida pela empresa norte-americana LexisNexis (MIDDLETON e CHALA, 2015) e que permite o processamento massivo de dados. Sua linguagem é otimizada para aplicações de Big Data e os programas nela desenvolvidos foram rodados em um ambiente em nuvem.

3.2 Limpeza dos dados

Antes de se iniciar a análise dos dados em si, foi importante limpar e organizar o arquivo de dados de forma a prepará-lo para uso dentro do projeto.

Primeiramente, foi analisada a coluna-alvo *loan_status*, que indica a condição em que se encontra o pagamento do empréstimo. Tal coluna apresentava diversos possíveis valores, como *Fully Paid*, *Charged Off*, *30_days_late*, *90_days_late*, entre outros. Como ainda não é possível determinar se um cliente com pagamentos atrasados sanará suas dívidas no futuro ou

não, os dados foram filtrados de modo a manter somente os registros cujo *loan_status* fosse *Charged Off*, *Current* ou *Fully Paid*. Essa variável também foi transformada em binária, sendo atribuído o valor 1 para os registros adimplentes e 0 para os inadimplentes. Com isso, o banco de dados passou a ter somente dados de empréstimos definitivamente adimplentes (*Current* (em dia) e *Fully Paid* (totalmente pago)) ou inadimplentes (*Charged Off*).

Também foram removidos os registros que representassem empréstimos conjuntos, feitos por mais de uma pessoa. Tais registros possuem colunas a mais que os registros de empréstimos individuais (como informações da segunda pessoa) que dificultam a comparação entre os dois tipos.

Após essa filtragem inicial, as mais de 150 colunas do arquivo de dados foram analisadas com base em seu significado e agrupadas conforme o tema comum que abordam (dados de renda, de inadimplência, de histórico bancário, informações pessoais do cliente, entre outros) para melhor compreensão dos dados como um todo. As colunas com dados posteriores ao financiamento do empréstimo, como *last_fico_score* (pontuação de crédito mais recente do indivíduo, a qual avalia seu histórico financeiro), foram removidas por não se encaixarem no objetivo do projeto. Só faz sentido tentar prever se um cliente pagará ou não seu empréstimo usando informações obtidas no momento em que o empréstimo foi solicitado.

Em seguida, foram comparadas as médias de cada coluna para empréstimos inadimplentes e adimplentes. Colunas cujas médias eram estatisticamente diferentes para cada grupo, indicativo de que podem ser usadas para diferenciar clientes adimplentes e inadimplentes, foram selecionadas.

Como muitas colunas apresentavam informações semelhantes, foi feita nova filtragem para que restassem somente colunas independentes entre si. A lista final de variáveis independentes é mostrada a seguir no Quadro 2:

QUADRO 2 – Colunas selecionadas e suas definições.

Nome da coluna	Descrição
All_util	Razão entre balanço e limite de crédito
Annual_inc	Renda anual
Dti	Razão entre gasto mensal com dívidas (exceto hipoteca e o empréstimo da LendingClub) e renda mensal
Fico_range_low	FICO Score (pontuação de crédito) do indivíduo
Home_ownership	Indica se o indivíduo tem casa própria, aluga, está pagando hipoteca, etc.

Initial_list_status	Indica se o empréstimo foi primeiro oferecido para que investidores o financiassem por inteiro ('w') ou se foi oferecido no mercado fracionário da LendingClub desde o início ('f')
Int_rate	Taxa de juros do empréstimo
Loan_amnt	Valor total do empréstimo
Mort_acc	Número de hipotecas que o indivíduo possui
Purpose	Propósito do empréstimo
Tax_liens	Número de propriedades penhoradas para pagamento de impostos
Term	Prazo do empréstimo (em meses)
Disbursement_method	Método pelo qual o indivíduo receberá o dinheiro do empréstimo
Debt_settlement_flag	Indica se o indivíduo, que está endividado, está trabalhando com uma empresa de liquidação de dívidas

Fonte: Adaptado de Yash (2020).

A limpeza realizada nos dados não representou perda significativa de dados. Ao final, o arquivo apresentava 2.106.701 registros (93,2% da quantidade original) e 16 colunas. Destes registros, 87,55% eram adimplentes.

3.3 Desenvolvimento dos modelos

Para o desenvolvimento de modelos de regressão logística, as variáveis independentes devem ser numéricas. Por isso, as variáveis categóricas foram transformadas em variáveis *dummy* por meio do processo de *one-hot encoding*. Tal processo consiste em transformar todos os valores possíveis de uma variável categórica em novas colunas, cujo valor é verdadeiro (1) ou falso (0). Com isso, o arquivo de dados passou de 14 para 32 variáveis independentes.

3.4 Amostragem

Mesmo com o uso de uma plataforma especializada em processamento massivo de dados, em muitos casos, foi mais eficiente utilizar amostras reduzidas dos dados para diminuir o tempo gasto com análises e, com isso, permitir melhor andamento do projeto. O tamanho das amostras foi decidido caso a caso com base na quantidade de modelos a serem rodados em cada teste. Porém, para um mesmo teste, as amostras eram de mesmo tamanho e continham os mesmos registros aleatoriamente selecionados do arquivo de dados total.

Ademais, convencionou-se o uso de 75% da amostra para treinamento e 25% para validação dos modelos.

4. ANÁLISE DOS DADOS E RESULTADOS

Neste quarto capítulo, são analisados os modelos construídos ao longo do trabalho. Primeiro, analisa-se um modelo inicial e, então, são detalhados diferentes processos feitos com o intuito de aperfeiçoá-lo.

4.1 Modelo inicial

Em um primeiro momento, foi testado um modelo inicial com amostra de 100 mil registros. Com base nos 25 mil registros de validação, foi obtida a seguinte matriz de confusão (Tabela 1):

TABELA 1 – Matriz de confusão para regressão logística inicial.

Regressão Logística Inicial		
Previsão \ Real	**Adimplente (1)**	**Inadimplente (0)**
Adimplente (1)	21731	2734
Inadimplente (0)	225	310

Fonte: Elaborada pelos autores (2021).

Em termos de acurácia, o modelo acertou 88,16% de suas previsões. Sua sensibilidade ficou em torno de 98,98%. Já no que se refere à precisão: 88,82% das classificações “adimplente” estavam corretas. Isso significa que, caso os pedidos de empréstimo fossem filtrados por esse modelo, a proporção de adimplentes cresceria em relação à proporção da amostra utilizada (87,54%).

4.2 Aperfeiçoamento dos modelos

Para aperfeiçoar os modelos de predição, iniciou-se um processo de análise do impacto das variáveis utilizadas sobre os resultados obtidos. O objetivo era identificar aquelas variáveis que pouco influenciavam as respostas (ou as influenciavam de forma negativa) para removê-las e melhorar, dessa forma, tanto os resultados quanto o tempo de processamento.

4.2.1 Correlações parciais

A princípio, pensou-se em utilizar os coeficientes de correlação parcial entre as variáveis para ordená-las quanto ao seu poder preditivo, método comumente utilizado em modelos de regressão linear. Entretanto, os resultados rapidamente demonstraram que, devido à natureza logística do modelo de regressão utilizado neste trabalho, os coeficientes de correlação linear eram inadequados.

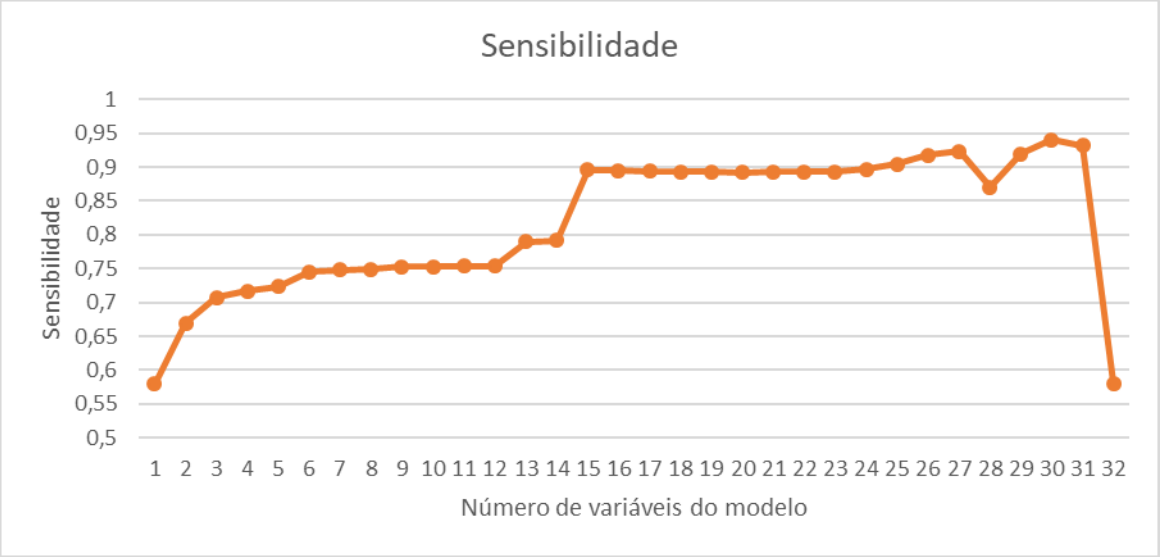
4.2.2 Adição passo a passo de variáveis

Em seguida, foi utilizada uma abordagem mais prática de adicionar as variáveis ao modelo, uma a uma, e escolher aquela que fornecesse o melhor resultado. Foram utilizados dois critérios para essa escolha: uma condição para a precisão do modelo gerado e outra para a sensibilidade. Primeiro, o modelo deveria apresentar precisão de pelo menos 90%. Além disso, o modelo deveria ter a maior sensibilidade dentre aqueles que respeitassem o critério de precisão. Em casos em que nenhum dos modelos atingiu 90% de precisão, optou-se por escolher a variável cujo modelo apresentava o valor mais alto desse indicador.

O critério de precisão foi estabelecido para garantir que os modelos aumentassem significativamente a proporção de adimplentes ao filtrarem uma amostra de pedidos (no arquivo todo, 87,55% são adimplentes). Já o critério de sensibilidade foi escolhido para que o modelo obtido descartasse o mínimo possível de registros adimplentes nesse processo de filtragem.

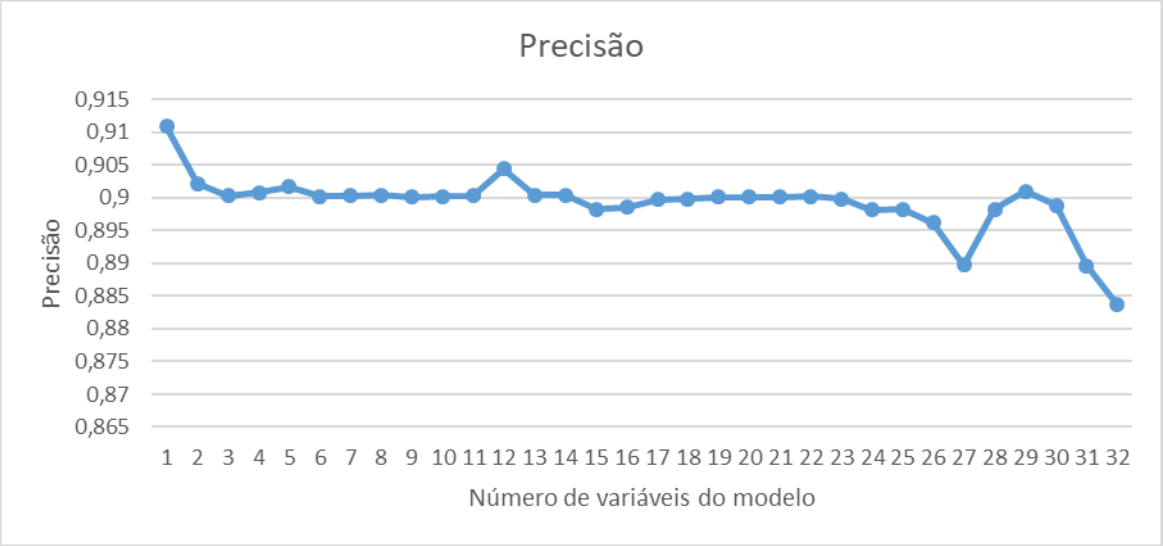
Plotando a sensibilidade e a precisão em função do número de variáveis dos modelos, obtiveram-se os seguintes resultados (Figuras 1 e 2):

FIGURA 1 – Sensibilidade dos modelos construídos pelo método passo a passo.



Fonte: Elaborada pelos autores (2021).

FIGURA 2 – Precisão dos modelos construídos pelo método passo a passo.



Fonte: Elaborada pelos autores (2021).

Pela análise do gráfico de sensibilidade (Figura 1), nota-se um crescimento até o patamar de 75%, onde o índice fica estagnado até o modelo de número 13, quando volta a subir e chega a cerca de 90%. A partir disso, a sensibilidade sofre pequenas oscilações, alcançando 94%, e, finalmente, cai para 58%.

A precisão (Figura 2), por sua vez, cai inicialmente e então fica estagnada em torno de 90% até o modelo de número 23. Depois, passa a sofrer oscilações até atingir 88,4% no último modelo, com 32 variáveis.

Segundo o critério adotado, o modelo ótimo seria aquele com 29 variáveis, visto que apresenta a maior sensibilidade (91,94%) dentre os modelos com precisão maior ou igual a 90% (sua precisão é de 90,1%). Isso significa que, nos testes realizados, o modelo conseguiu identificar corretamente 91,94% dos pedidos de empréstimo adimplentes, tendo 90,1% de precisão em suas classificações do tipo “adimplente”.

Com isso, o modelo foi capaz de transformar uma lista com 5000 pedidos de empréstimo (dos quais 87,88% eram adimplentes) em uma lista com 4484 pedidos, sendo 90,1% deles adimplentes. Aplicando-se um teste de hipótese a esses dados, pode-se afirmar que a proporção aumentou com mais de 99,999% de confiança.

No modelo com 29 variáveis, apenas *fico_range_low* (pontuação de crédito do indivíduo no momento em que pediu o empréstimo), *disbursement_method_cash* (coluna que indica se o indivíduo receberá o empréstimo em dinheiro) e *home_ownership_rent* (indica se a

casa do indivíduo é alugada) ficaram de fora. É possível notar nas Figuras 1 e 2, que, de fato, a sensibilidade e a precisão dos modelos decrescem consideravelmente com a adição das três últimas variáveis.

4.2.3 Adição passo a passo de variáveis pelo AIC

Também foi desenvolvido um método para análise do impacto das variáveis sobre os modelos a partir do AIC (Akaike Information Criterion). De forma análoga ao método anterior, as variáveis foram adicionadas ao modelo, uma a uma, escolhendo-se a variável com melhor resultado para o próximo modelo. Neste caso, porém, o critério de escolha é o AIC mais baixo, o que indica uma boa relação entre acurácia e complexidade.

Com esse método, foram obtidos os seguintes resultados (Tabela 2):

TABELA 2 – Resultados do método passo a passo pelo AIC.

Método passo a passo pelo AIC		
Variável adicionada	**Número de variáveis**	**AIC**
Fico_range_low	1	3703,65
Debt_settlement_flag_Y	2	3419,21
Int_rate	3	3369,88
Initial_list_status_w	4	3352,79
Purpose_moving	5	3389,60

Fonte: Elaborada pelos autores (2021).

Pela Tabela 2, nota-se que o AIC cai gradativamente conforme novas variáveis são adicionadas, até a variável de número quatro. Com a adição da quinta variável (*purpose_moving*, que indica se o indivíduo quer usar o dinheiro do empréstimo para se mudar), o AIC volta a crescer. Isso indica que, a partir de 5 variáveis, o ganho de acurácia não compensa mais o aumento da complexidade do modelo, por isso o método foi interrompido.

O modelo de 4 variáveis, considerado o melhor pelo método passo a passo do AIC, apresenta precisão de 89,13% e sensibilidade de 100%. Assim, de uma lista de 5000 pedidos (dos quais 87,88% eram adimplentes), restaram 4394 pedidos, sendo 89,12% adimplentes. Assim, é possível afirmar que a proporção de adimplentes cresceu com confiança de cerca de 99,995%. Ademais, nessa filtragem, nenhum pedido adimplente foi perdido.

5. CONSIDERAÇÕES FINAIS

O objetivo deste trabalho era construir um modelo de predição que fosse capaz de filtrar os pedidos de empréstimo disponíveis na plataforma da LendingClub, de forma a reduzirem-se as chances de se investir no empréstimo de uma pessoa que não conseguirá arcar com suas dívidas. A partir do estudo e da aplicação de diferentes técnicas estatísticas, foram construídos diversos modelos com esse fim.

Pelo método passo a passo com precisão e sensibilidade como critérios, foi possível analisar os pedidos de empréstimo e prever aqueles que viriam a ser adimplentes com precisão de 90,1% e sensibilidade de 91,94%. Com essa filtragem, eleva-se significativamente a proporção de adimplentes dentre os pedidos analisados (de 87,88% para 90,1%).

Ademais, pelo método passo a passo pelo AIC, foi obtido um modelo com 89,13% de precisão e 100% de sensibilidade. Tal modelo, conquanto forneça menor aumento da proporção de adimplentes, descarta um menor número de pedidos adimplentes em sua filtragem.

Portanto, dado o objetivo inicial, foram construídos dois modelos satisfatórios e com aplicações distintas. O primeiro, pode ser usado quando o foco for aumentar a proporção de adimplentes o máximo possível, enquanto o segundo pode ser utilizado quando se deseja descartar o menor número possível de adimplentes durante o processo. Assim, fica evidente o valor de cada um deles para investimentos na plataforma da LendingClub.

Futuramente, seria interessante estudar o desempenho de outros modelos de aprendizado de máquina, como árvores de decisão, na tarefa de identificar bons pagadores para pedidos de empréstimo. Tais modelos poderiam ser comparados aos de regressão logística, para que fosse construído um modelo ótimo.

REFERÊNCIAS

BALAJI, T.K.; ANNAVARAPU, C.S.R.; BABLANI, A. Machine learning algorithms for social media analysis: A survey. **Computer Science Review**, [*s. l.*], 2021.

DÓRIA, Fernanda Guimarães; **Utilização de Big Data para a construção de features no desenvolvimento de modelos preditivos em análise de crédito**. Orientador: Prof. Dr. Ricardo Giglio. 2019. Trabalho de conclusão de curso (Bacharelado em engenharia mecânica com habilitação em produção) - Centro Tecnológico da Universidade Federal de Santa Catarina, [*S. l.*], 2019.

GARTNER (Estados Unidos da América). **Gartner Glossary**: Big Data. [*S. l.*: *s. n.*], 2021. Disponível em: https://www.gartner.com/en/information-technology/glossary/big-data. Acesso em: 13 jul. 2021.

GRUS, Joel. **Data Science do Zero**. 1. ed. Rio de Janeiro: Alta Books, 2016.

MIDDLETON, A. and CHALA, A. (2015) "HPCC Systems: Introduction to HPCC", https://cdn.hpccsystems.com/whitepapers/wp_introduction_HPCC.pdf, December.

MILLER, James D.; FORTE, Rui Miguel. **Mastering Predictive Analytics with R**. 2. ed. rev. [*S. l.*]: Packt Publishing, 2017. 448 p.

ORACLE (Estados Unidos da América). **Big data definido**. [*S. l.*: *s. n.*]. Disponível em: https://www.oracle.com/br/big-data/what-is-big-data/#link5. Acesso em: 13 jul. 2021.

RUIZ-PÉREZ, Iñaki *et al*. A field-based approach to determine soft tissue injury risk in elite futsal using novel machine learning techniques. **Frontiers in Psychology**, [*s. l.*], 5 fev. 2021. Disponível em: https://doi.org/10.3389/fpsyg.2021.610210. Acesso em: 9 jul. 2021.

YASH (2020). Lending Club 2007-2020Q3: Consolidated data of all the years. Disponível em: https://www.kaggle.com/ethon0426/lending-club-20072020q1. Acessado em novembro de 2020.