**Proposta de um Sistema de Avaliação de Risco de Crédito Utilizando Machine Learning**

**Maycon Brandão Bruzolato¹, Marcelo dos Reis²**

¹ Aluno de Graduação do Curso de Ciência da Computação da PUC Minas Campus de Poços de Caldas - MG, Brasil

² Orientador e Docente do Departamento de Ciência da Computação da PUC Minas Campus de Poços de Caldas - MG, Brasil

[maycon.brandao@sga.pucminas.br](mailto:maycon.brandao@sga.pucminas.br), [marceloreis@pucpcaldas.br](mailto:marceloreis@pucpcaldas.br)

***Abstract****.* *Credit risk assessment is a way to prevent and reduce default, in order to make decision-making more secure. This work presents the proposal of a prototype system that assists in credit risk analysis through machine learning by handling and manipulating data using the Python language and the Scikit-Learn library.*

***Resumo****. A avaliação de risco de crédito é uma forma de se precaver e reduzir a inadimplência, de modo a tornar a tomada de decisão mais segura. Este trabalho apresenta a proposta de um modelo que auxilia na análise de risco de crédito por machine learning tratando e manipulando dados utilizando a linguagem python e da biblioteca do Scikit-Learn.*

1. **Introdução**

No início do século XX a escolha quanto aos proponentes de crédito eram baseadas na avaliação de analistas de crédito (THOMAS, et al., 2017). Assim, as decisões quanto aos limites ou aprovações e negações acabavam por serem subjetivas, que as vezes divergiam dentro da própria instituição dependendo do analista, sendo que uma mesma solicitação pode ou não aprovada.

Uma das maneiras de tornar essas decisões mais precisas, é utilizando do aprendizado de máquina, que consiste em um modelo que elimina a subjetividade humana. Ele é constituído por “um conjunto de regras e procedimentos, que permite que os computadores possam agir e tomar decisões baseadas em dados ao invés de serem explicitamente programados para realizar uma determinada tarefa. Programas de *machine* *learning* também são projetados para aprender e melhorar ao longo do tempo, quando expostos a novos dados” (MATOS, 2015).

Nesse contexto, a *machine* *learning* é uma tecnologia que cabe muito bem para a construção de modelos analíticos, fazendo com que as máquinas aprendam com os dados e efetuem análises preditivas de maneira rápida e concisa. Assim, a criação de uma pontuação de risco em crédito automatizada e eficaz por meio da *machine* *learning* ajudaria na previsão da capacidade de crédito com precisão diminuindo a taxa de inadimplência por parte de uma carteira.

1. **Análise de crédito**

A análise de crédito é o processo que envolve a reunião de informações a respeito de um tomador de crédito, com o objetivo de decidir sobre a concessão do crédito, sendo o objetivo da análise identificar os riscos e predizer a possibilidade dos futuros pagamentos, evitando a inadimplência (SÁ, 1999).

1. **Aprendizado de máquina**

Tradicionalmente, a única maneira de fazer o computador realizar alguma tarefa é escrever um algoritmo que descreva essas ações de forma detalhada, porém com os novos códigos de *machine* *learning*, conhecidos como aprendizes, tudo é diferente. Eles descobrem sozinhos por meio de inferências a partir de dados (DOMINGOS, 2015).

Modelos de aprendizado de máquina podem ser um fator competitivo e de decisão para desenvolver formas ágeis de processar dados (FRANCO, 2014).

Além disso, o aprendizado de máquina envolve o desenvolvimento de algoritmos que aprendem usando a generalização oriunda de experiências de treinamento ligada a tarefas dentro de um domínio analisado (LUGER, 2013).

1. **Proposta de um sistema de avaliação de risco de crédito**

O objetivo deste trabalho é desenvolver um modelo de *machine* *learning* capaz de prever a probabilidade de determinado cliente se tornar inadimplente, utilizando *Scikit*-Learn por meio da linguagem *python*.

1. ***Dataset***

Um *dataset* consiste em um conjunto de dados, geralmente organizados de forma a favorecer a manipulação e extração de informações relevantes. Neste trabalho será usado um *dataset* elaborado a partir de bases de dados de uma *factoring* situada no sul do estado de Minas Gerais, nele constam informações de seus cedentes coletadas pela empresa.

Pelo *dataset* ser extenso e possuir algumas informações irrelevantes, ou que não agregam tanto valor, foi necessário a análise de todas as colunas para avaliar quais são pertinentes ou não. Essa analise previa ocorreu utilizando o software Microsoft Excel, uma vez que foi necessário realizar a padronização dos dados por meio da categorização, além disso a classificação de risco teve de ser adaptada para um padrão numérico, possibilitando a manipulação da variável, sendo a última etapa a criação de um documento em formato *CSV (Comma-Separated Values).*

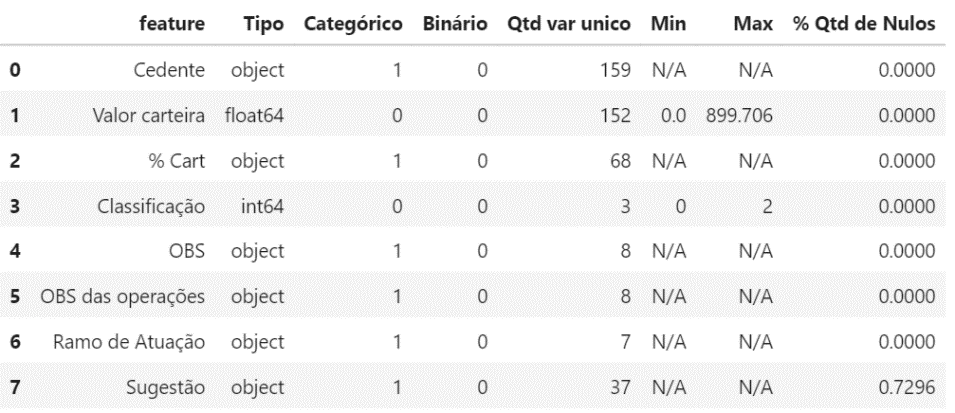
A coluna “Classificação” o principal alvo, que consiste em uma variável no valor possível: 0, 1 ou 2; são elas responsáveis pela separação de um cedente com classificação de alto, médio ou baixo risco, respectivamente. Outras colunas também são importantes para o modelo de aprendizado de máquina proposto como: “OBS”, “OBS das operações” e “Ramo de Atuação” que correspondem na devida ordem, observações relevantes sobre os sacados daqueles cedentes, o segundo sobre os títulos do cedente e o terceiro é referente ao ramo de atuação do cedente. A variáveis “Sugestão” não é tão relevante, por fugir do intuito do trabalho.

Ao final da elaboração do *dataframe* foi constatado a presença de 159 registros e 8 variáveis que serão utilizadas ao decorrer do desenvolvimento.

1. **Limpeza e tratamento dos dados**

Após inspecionar os dados do *dataset*, organizando e padronizando todos os registros é necessário a estruturação das *features* em forma de um dicionário para recebê-las e simplificar a visualização.

Assim, foi estabelecido a criação do dicionário utilizando oito parâmetros que recebem as *features*, são eles os representados na figura 1 abaixo.



**Figura 1: Dicionário de parâmetros**

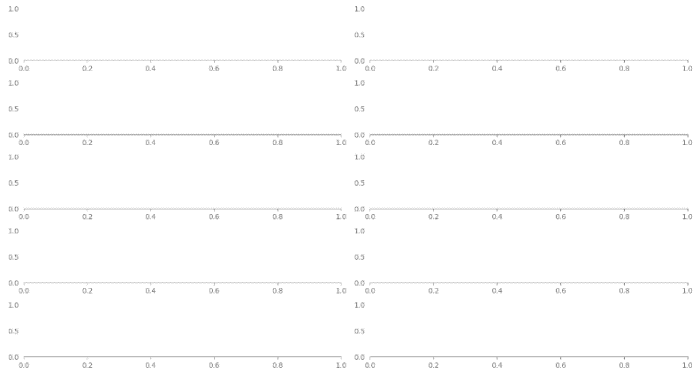
Logo após, foi necessário excluir os valores ausentes dentro das variáveis, excluir as que não serão relevantes e remover os valores negativos.

Desta feita, com o propósito de deslindar ainda mais o *dataset*, foi preciso retirar os valores ausentes da variável alvo “Classificação” e excluir a variável “Sugestão” para ser possível ter um resultado ainda mais claro sobre as *features* do *dataframe*, garantindo a ausência de nulos, que é o objetivo da limpeza dos dados.

1. ***Outliers***

Com a retirada de todos os registros com nulos a próxima etapa a ser seguida foi a análise das *outliers* onde através da ferramenta *boxplot* é possível visualizar a distribuição e os valores discrepantes entre os dados, possibilitando perspectivas gráficas, além de facilitar a representação da variação de dados de uma variável numérica por meio de quartis.

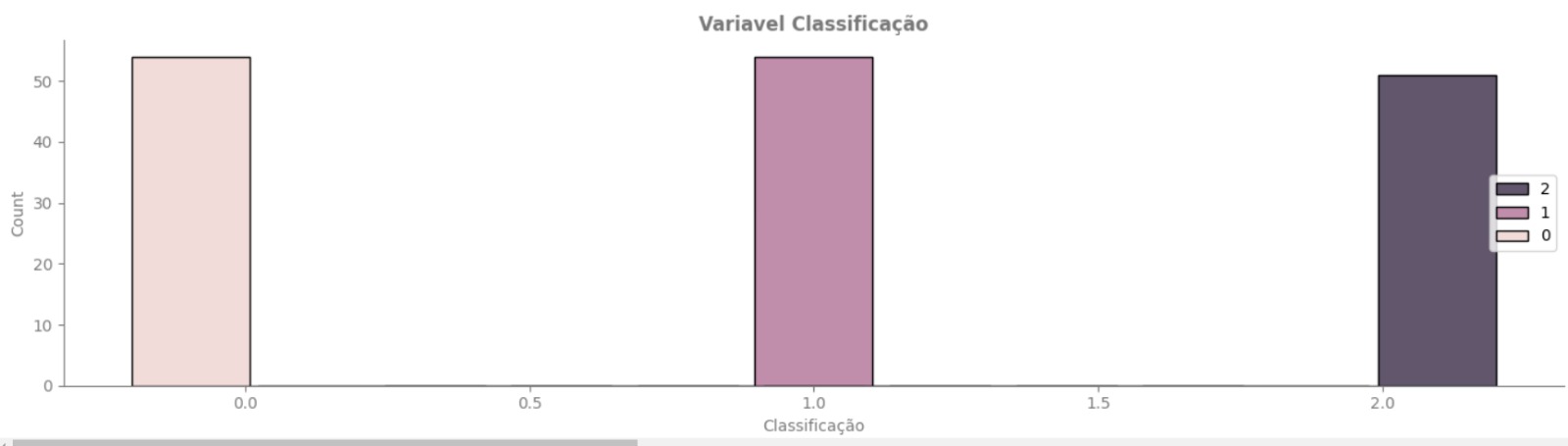
Realizada as análises plotando as *outliers* foi possível verificar que nenhum dos registros do *dataset* possui um valor diferente de 0,1 e 2 já que um *outlier* é um dado que se distancia radicalmente dos demais que compõem a amostra analisada, resultando na ausência de qualquer *outlier* como o gráfico vazio da figura 2.



***Figura 2: Representação dos outliers***

1. **BalanceamentodeDados**

O processo de harmonização de dados é importante para excluir uma possível variação entre as amostras, no caso da predição de um modelo de *machine* *learning* o algoritmo tende a classificar a maior parte dos casos conforme a distribuição dos dados. Caso a ocorrência do alvo não seja balanceada o modelo opta pela que apresente maior constância.

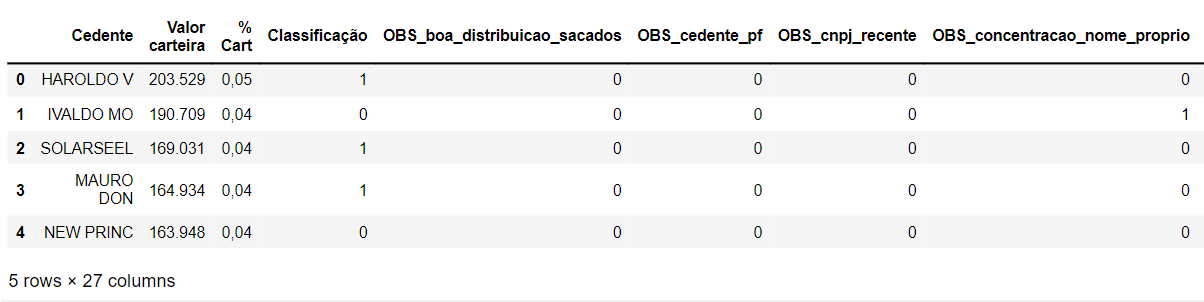
Na variável “Classificação”, são possíveis dois valores. Durante a análise foi notado que a incidência de cada uma delas acontecer é próxima, descartando ações que eliminariam tendências na base de dados, conforme no gráfico aludido pela figura 3.

**Figura 3: Gráfico da variável alvo Classificação**

1. **Preparação dos dados**

Realizadas as etapas necessárias para filtrar e extrair o melhor dos dados do *dataset*, é preciso transformar esses dados em valores numéricos e assim ser possível aplicar um modelo de *machine* *learning*. Nas variáveis categóricas foi necessário o método *get*\_*dummies* da biblioteca Pandas para conseguir convertê-las.

Em variáveis numéricas como o “Valor da carteira” e “Classificação”, a conversão não é necessária, já que se encontram no formato adequado, a variável “% Cart” foi desconsiderado para a construção do modelo, já que é baseada em outra variável “Valor carteira”.

 Segue na figura 4 a tabela com os cinco primeiros registros para exemplificar essa conversão.

**Figura 4: Conversão de variáveis**

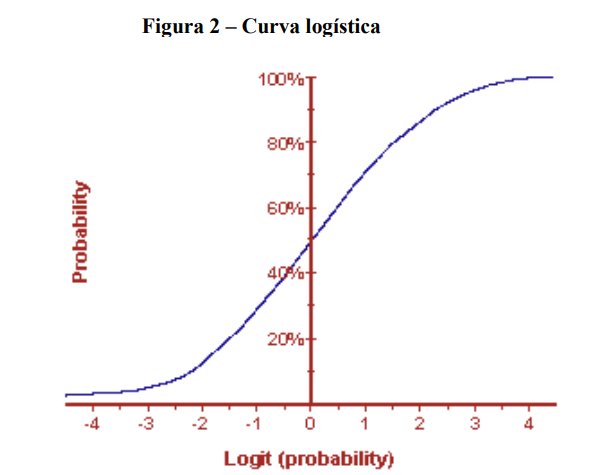
1. **Escolha do melhor modelo de aprendizado de máquina**

Finalizadas as preparações dos dados, é possível aplicar vários modelos de *machine* *learning*, em busca de performance e do qual atenda da melhor forma.

A métrica inicialmente escolhida para avaliar e auxiliar a definir o melhor modelo de acordo com a performance do *dataset* é a *ROC-AUC*, já que esse modelo é indicado quando é necessário haver distinção entre as classes, sucintamente eles representam uma curva que é capaz de distinguir entre as classes. Quanto maior a *AUC* indubitavelmente melhor será a capacidade do modelo em distinguir um alto ou baixo risco de crédito.

Entretanto, no trabalho não foi utilizado a métrica *ROC-AUC* para a escolha do melhor modelo devido a dificuldades na implementação em avaliar a performance de diferentes modelos diante do *dataset*.

Dessa forma, o modelo definido foi de regressão logística, uma vez que é a técnica mais utilizada no mercado para o desenvolvimento de modelo de *credit scoring* (ROSA, 2000), esse tipo de modelo tem por finalidade prever a probabilidade de ocorrência ou não de um evento, no caso de ser ou não inadimplente. O gráfico formado pela relação entre as variáveis independentes ou *preditoras* e a dependente forma a curva logística apresentado na figura 5.



**Figura 5: Modelo de curva logística. Adaptado de Adaptado de SHARMA, 1996**

1. **Treino e Teste**

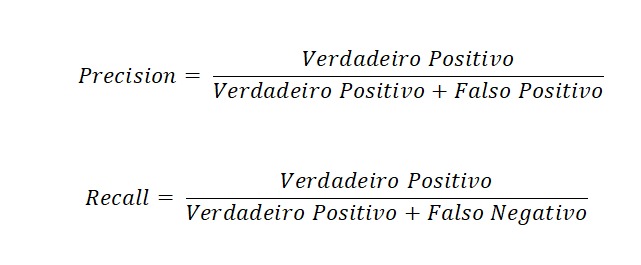
Concluídas as conversões, é necessário dividir a base entre testes e treinos para que seja possível ajustar os modelos de aprendizado de máquina, conforme o *dataset* de treino, selecionar o modelo de acordo com a performance e realizar as otimizações necessárias para validar o nível de generalização do modelo em dados desconhecidos, os de testes.

Isso é feito dividindo os valores em duas variáveis, sendo X as variáveis independentes e Y a variável “Classificação”, que é o alvo.

1. **Métricas para avaliação do modelo**

Para mensurar os resultados do modelo é necessário definir uma métrica principal; algumas das métricas mais utilizadas são: acurácia, *precision*, *recall* e *f1-score*. A acurácia consiste no percentual de acertos do modelo, a *precision* é recomendada quando o custo de um falso positivo é alto, já o recall é recomendado quando o custo de um falso negativo é alto. Quando se deseja medir o desempenho geral do modelo, a *f1-score* é a métrica recomendada e, seu cálculo ocorre através da média harmônica entre recall e *precision*.

Para avaliar a precisão do modelo desenvolvido, foi definido além da acuracidade e recall representada, pela figura 6.

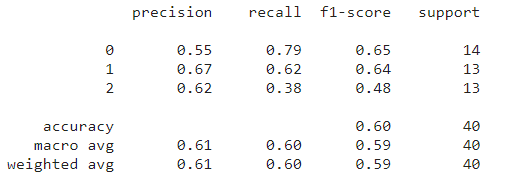


**Figura 6: Cálculo de recall e precision. Adaptado de SHUNG, 2018**

1. **Resultados**

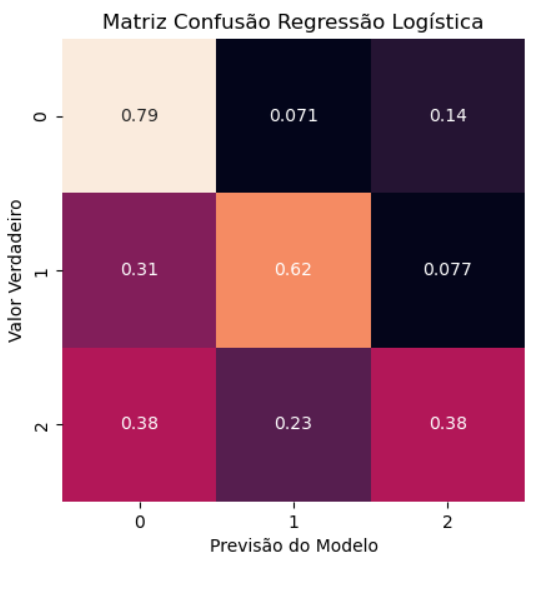
Concluída a aplicação do modelo de aprendizado de máquina foi constatada uma acurácia de 60%, que indica a probabilidade da taxa de acertos no modelo, entretanto essa medida pode ser tendenciada pelo não balanceamento das classes, o que não ocorre no *dataset* do trabalho, por isso é importante levar em consideração também outras métricas, como a *precision, recall.*

No contexto de risco de credito, uma vez que o custo de falsos negativos é alto a métrica a ser utilizada para análise foi o *recall*, pois a classificação de uma solicitação de alto risco como risco baixo seria prejudicial ao negócio. Foi obtido como resultado para o modelo um *recall* de 79% em casos de alto risco, 62% em de médio e 38% em baixo risco, tendo como média um valor de 60%, conforme a figura 7.



**Figura 7: Resultados do modelo**

Na matriz de confusão representado pela figura 8, é possível observar as frequências de classificação para cada classe do modelo, sendo que os resultados presentes na diagonal são referentes a porcentagem de assertividade do modelo em cada classe.



**Figura 8: Matriz de confusão do modelo**

1. **Conclusão**

A utilização de técnicas como o tratamento e o balanceamento dos dados, complementados pela definição de métricas como o caso da acuracidade e *recall,* garantiram um treinamento para o algoritmo sem tendências, o que promove a acuracidade e boa capacidade de separabilidade entre as classes de decisão em um modelo de *machine learning* supervisionado.

O tópico que mais gerou dificuldade no desenvolvimento do trabalho foi quanto a escolha do modelo de aprendizado de máquina, sendo que em sua concepção inicial o pressuposto de testar diferentes modelos e comparar qual alcançaria a melhor medida *ROC-AUC*, dando assim melhor performance para o modelo. Entretanto devido à erros no código não foi possível a implementação dos testes.

Dessa forma a escolha da regressão logística se deu através de referenciais de outros trabalhos correlatos, que afirmam ser a técnica mais adequada para modelos de *credit scoring.*

Assim o modelo de *machine* *learning* proposto neste trabalho, utilizando a regressão logística foi capaz de prever, dentre dados de teste de cedentes, qual o nível de risco ao serem inadimplentes, em 60% dos casos, minimizando possíveis perdas quanto ao ceder de crédito.

É importante ressaltar que, seria interessante a verificação da existência de mais dados, que poderiam ser aproveitados para o treinamento do aprendizado de máquina, ou a utilização de um *dataset* com um número maior de registros, além da realização dos testes entre os modelos, podendo assim obter um maior desempenho com outras formas de regressão.

**Referências**

DOMINGOS, Pedro. The Master Algorithm. Novatec editora LTDA, 2015.

FRANCO, C. R. Inteligência Artificial. Londrina: Educational Publisher and Distributor S.A, 2014.

LUGER, G. F. Artificial intelligence. Pearson education of Brazil. 6ª edição, 2013.

MATOS, D. Conceitos Fundamentais de Machine Learning. Ciência e Dados. 2015. Disponível em: <<http://www.cienciaedados.com/conceitosfundamentais-de-machine-learning/>>. Acesso em: 28 de abril de2023

ROSA, P. T. M. Modelos de Credit Scoring: Regressão Logística, CHAID e REAL. 2000.125 f. Dissertação (Mestrado em Estatística) – Curso de Pós-graduação em Estatística, Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo, 2000.

SÁ, Carlos Alexandre. Apostila de determinação de limites de crédito n°41. São Paulo: 1999.

SHARMA, S. Applied multivariate techniques. New York: John Wiley and Sons, 1996.

SHUNG. Medium, Stay curious. Accuracy, Precision, Recall or F1. 2018. Disponível em: <<https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>>. Acesso em: 07 de maio de 2023.

THOMAS, L., CROOK, J., & EDELMAN, D. (2017). “Credit Scoring and Its Applications, Second Edition”. SIAM News. ISBN 978-1-611974-55-3