FACULDADE IMPACTA DE TÉCNOLOGIA

COMPARAÇÃO DE MÉTODOS PREDITIVOS PARA ANÁLISE DE CRÉDITO

Eugênio Lenine Gueiros Diniz

Christian Negrão

Higor Carrião

Pamella Oliveira

Roberto Santos

RESUMO

Este estudo tem como objetivo desenvolver um modelo de pontuação de crédito para a aprovação automatizada de linhas de crédito de home equity, utilizando dados disponíveis no conjunto de dados HMEQ. Para isso, serão analisadas as variáveis do conjunto de dados HMEQ a fim de identificar aquelas que têm maior impacto na probabilidade de um mutuário inadimplir ou entrar em atraso com o pagamento de um empréstimo de home equity. Além disso, serão utilizadas técnicas de modelagem preditiva para criar um modelo de pontuação de crédito que seja estatisticamente robusto e empiricamente derivado. A precisão e a interpretabilidade do modelo desenvolvido serão avaliadas e comparadas com modelos de pontuação de crédito existentes. Por fim, com base nos resultados obtidos, serão propostas melhorias no processo de aprovação de linhas de crédito de home equity.

Palavras-chave: modelo de pontuação de crédito, home equity, modelagem preditiva.

ABSTRACT

This study aims to develop a credit scoring model for automated approval of home equity lines of credit, using data available in the HMEQ dataset. To achieve this, the variables in the HMEQ dataset will be analyzed to identify those with the greatest impact on the likelihood of a borrower defaulting or becoming seriously delinquent on a home equity loan. Predictive modeling techniques will be used to create a credit scoring model that is statistically robust and empirically derived. The accuracy and interpretability of the developed model will be evaluated and compared to existing credit scoring models. Finally, based on the results obtained, improvements in the home equity line of credit approval process will be proposed.

Keywords*: credit scoring model, home equity, predictive modeling*.

1 INTRODUÇÃO

Nos últimos anos, o setor de crédito ao consumidor tem visto um aumento na demanda por empréstimos com garantia de imóvel, como as linhas de crédito de home equity. Com isso, os departamentos de crédito de bancos têm buscado maneiras de agilizar e automatizar o processo de tomada de decisão para aprovação desses empréstimos. Uma das formas de fazer isso é através da criação de modelos de pontuação de crédito empiricamente derivados e estatisticamente sólidos, seguindo as recomendações da Equal Credit Opportunity Act (DACCORONE E PORTO, 2021).

Neste contexto, o conjunto de dados Home Equity (HMEQ) surge como uma opção para treinar tais modelos. Este conjunto de dados contém informações de desempenho de empréstimos e informações de linha de base para 5.960 empréstimos recentes de home equity. O home equity pode ser uma forma muito interessante de obtenção de recursos, pois tem a vantagem de oferecer juros menores em comparação com outras modalidades de crédito, como o cheque especial e o cartão de crédito. Além disso, o prazo de pagamento pode ser bastante longo, o que ajuda a tornar as parcelas mais acessíveis (ARANTES, 2016).

O alvo (BAD) é uma variável binária que indica se um candidato eventualmente entrou em default ou estava seriamente inadimplente. Este resultado adverso ocorreu em 1.189 casos (20%). Para cada candidato, 12 variáveis de entrada foram registradas. O objetivo deste artigo é explorar a aplicação do conjunto de dados HMEQ na criação de um modelo de pontuação de crédito para linhas de crédito de home equity, seguindo as recomendações da Equal Credit Opportunity Act (KAGGLE, 2023).

2 OBJETIVOS

2.1 Objetivo Geral

Como objetivo geral, este estudo busca automatizar o processo de tomada de decisão para aprovação das linhas de crédito, sendo necessário para tal o mapeamento dos tipos de requerentes que o banco possui, seus níveis de inadimplência e quais politicas o banco deve tomar para ter uma máxima eficiência de suas linhas de crédito.

2.1 Objetivo Especifico

Os objetivos específicos deste estudo têm como foco o desenvolvimento de um modelo de pontuação de crédito para a aprovação automatizada de linhas de crédito

de home equity, utilizando dados disponíveis no conjunto de dados HMEQ. Para atingir esses objetivos, serão analisadas as variáveis do conjunto de dados HMEQ a fim de identificar aquelas que têm maior impacto na probabilidade de um mutuário inadimplir ou entrar em atraso com o pagamento de um empréstimo de home equity (KAGGLE, 2023).

Além disso, serão utilizadas técnicas de modelagem preditiva para criar um modelo de pontuação de crédito que seja estatisticamente robusto e empiricamente derivado. A precisão e a interpretabilidade do modelo desenvolvido serão avaliadas e comparadas com modelos de pontuação de crédito existentes. Por fim, com base nos resultados obtidos, serão propostas melhorias no processo de aprovação de linhas de crédito de home equity (KAGGLE, 2023).

3 METODOLOGIA

Desta forma, está estudo tem como objetivo principal mapear os tipos de requirentes que o banco possui para melhorar suas políticas de linhas de crédito. Para sua realização, o objetivo específico se divide em elaborar uma análise descritiva da base disponibilizada para um melhor entendimento das variaveis ali contindas e fazer o tratamento e normalização de seus dados para a aplicação de modelos preditivos (CAMARGO, SANTOS e ARAÚJO, 2012).

Na base de dados deste banco, os requerintes inadimplentes e adimplentes são definidos pela variável binária BAD, sendo 1 (um) para adimplentes e 0 (zero) para inadimplentes. Os requerintes adimplentes são aqueles que estão com todas as suas contas em dia, que pagaram pelos emprestimos tomados. Já os inadimplentes são aqueles que possuem alguma pendencia com suas obrigações na instituição.

Além desta variavel, a variavel JOB se divide em 6 (seis) classes, sendo elas: *Other* para Outros; *Office* para Escritório; *Sales* para Vendedor; *Mgr* (*Menager*) para Chefe, *ProfExe* para Profissional e; *Self* para Autonomo.

Abaixo, a Tabela 01 apresenta o dicionário de dados dessa base com o título do atributo, a informação do atributo (descrição da variavel) e o tipo de dado que o atributo representa no pré processamento e análise de suas informações.

Tabela 01 – Dicionário de Dados da Base HMEQ (Home Equity)

Atributo	Descrição	Tipo do Atributo
YOJ	Anos no emprego atual;	float64
DEROG	Número de principais relatórios depreciativos;	float64
DELINQ	Número de linhas de crédito inadimplentes;	float64
BAD	0 (cliente inadimplente no empréstimo), 1 (empréstimo reembolsado);	int64
LOAN	Montante do pedido de empréstimo;	int64
CLAGE	Idade da linha comercial mais antiga em meses;	float64
NINQ	Número de linhas de crédito recentes;	float64
CLNO	Número de linhas de crédito;	float64
DEBTINC	Razão Dívida / Renda;	float64
MORTDUE	Montante devido na hipoteca existente;	float64
VALUE	Valor da propriedade atual;	float64
JOB	Categorias profissionais	object
REASON	DebtCon (consolidação da dívida), HomeImp (melhoria da casa);	object

Fonte: Elaboração própria a partir do banco de dados disponibilizado pela plataforma Kaggle.

Ao se aprofundar na composição da base é possível observar que a base contém muitos valores nulos. Para seguir com o trabalho de construir um modelo descritivo, é necessário corrigir a base preenchendo as informações faltantes. A base será complementada com as seguintes informações para as variáveis que possuem *null* em sua estrutura: MORTDUE, DELINQ, DEROG, CLNO, NINQ, CLAGE, DEBTINC. YOJ, VALUE – Assumido como 0, REASON – Assumido como *'Other*' (Outro), JOB – Assumido como '*None*' (Nenhum).

Para alcançar o objetivo geral de criar um modelo preditivo para aprovação de home equity lines of credit, foram utilizados quatro algoritmos de aprendizado de máquina: regressão logística, Random Forest e XGBoost.

A primeira técnica utilizada para modelar os dados foi a regressão logística, que é amplamente utilizada em problemas de classificação binária. Segundo Gonçalves e Gouvêa (2013), a regressão logística é uma abordagem popular para prever uma variável binária a partir de um conjunto de variáveis independentes. Essa técnica estima a probabilidade de uma resposta binária ocorrer com base em um conjunto de variáveis preditoras.

A segunda técnica utilizada foi Random Forest,um método de conjunto (ensemble method) que consiste em combinar várias árvores de decisão independentes para produzir um modelo preditivo robusto e preciso. Cada árvore é

construída a partir de uma amostra aleatória dos dados de treinamento e utiliza um subconjunto aleatório das variáveis preditoras para dividir os dados em nós. Foram ajustados modelos de Random Forest utilizando um número crescente de árvores e uma variação aleatória de parâmetros, como o número de variáveis preditoras utilizadas para cada árvore e a profundidade máxima das árvores.

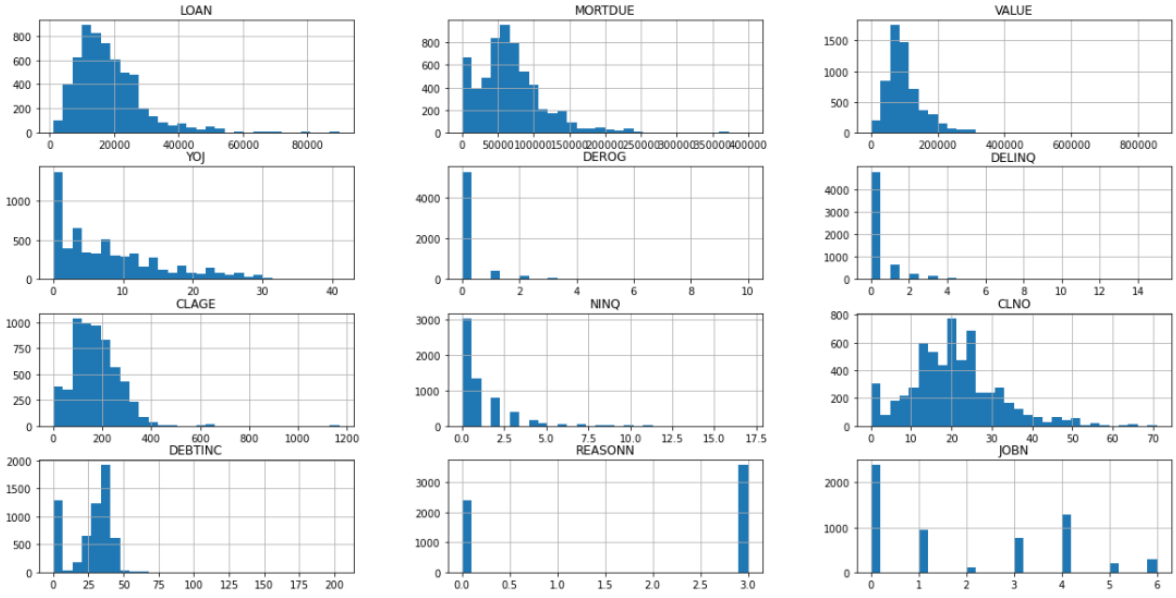
Por fim, o XGBoost, um algoritmo de gradient boosting com uma implementação eficiente e escalável, foi utilizado para modelar os dados. O XGBoost utiliza um conjunto de árvores de decisão para minimizar a função de perda e melhorar a precisão do modelo.

4 DESENVOLVIMENTO

A base inicialmente não possui um bom balancemento, tendo muitos valores em branco ou nulos em sua base conforme análise inicial. A base concentra sua informação nas profissões *Office*, *ProfExe* e *Mgr* (*Other's* não está sendo considerando pois o mesmo pode se dividir em várias outras classes). A distribuição na variavel BAD possui uma distribuição onde 20% dos requirintes são inadimplentes (sendo em sua maioria Profissionais e *Menagers*) e 69% possuem algum tipo de oferta para negociação de dívidas passadas (*Debt Consolidation*).

Na Figura 02, é possível observar também que as variáveis explicativas tem distribuição assimetria a direita e a esquerda e os mesmos não tendência a normalidade.

Figura 02 – Histogramas das variávis explicativa numericas.



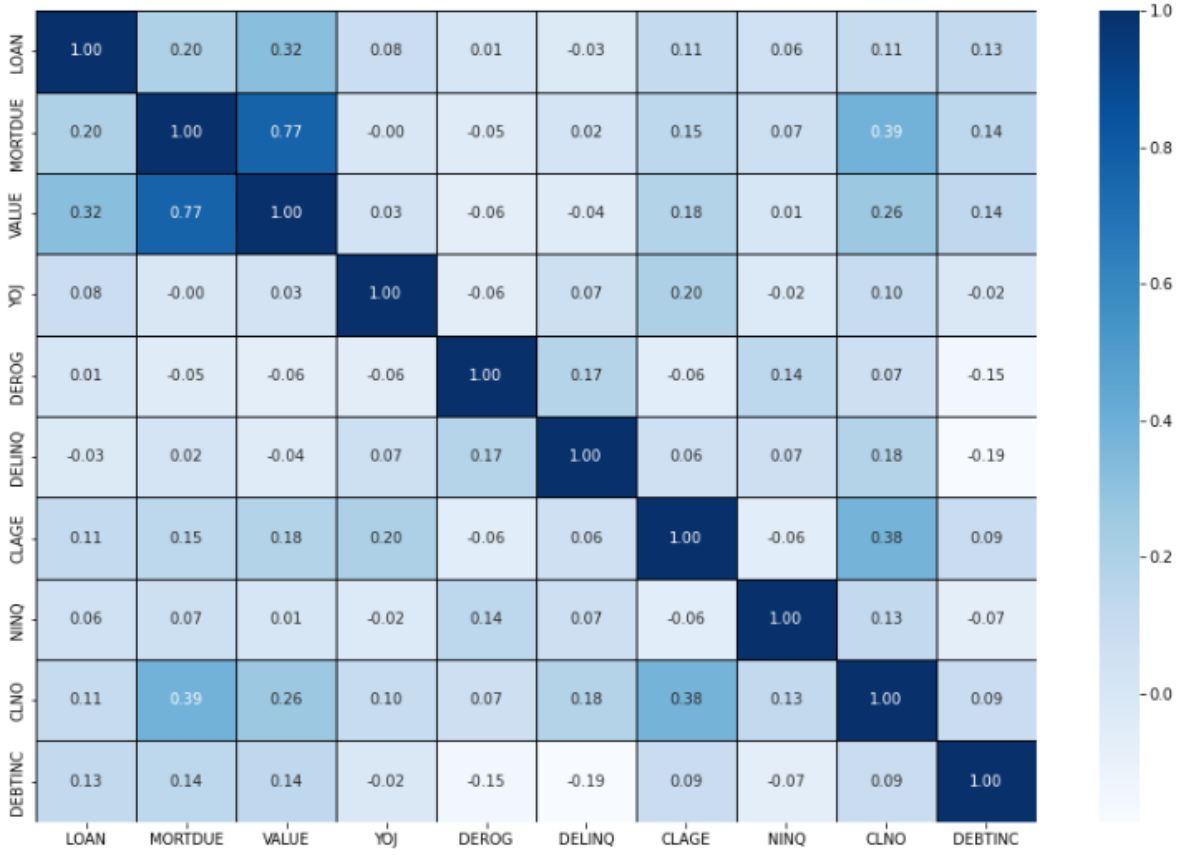
Fonte: Elaboração própria a partir do banco de dados disponibilizado pela plataforma Kaggle.

Com exceção das componentes LOAN, MORTDUE, VALUE e CLAGE por condição do cliente (BAD) que tem menor assimetria, as distribuições das demais apresentam forte assimetria a direita e a esquerda portanto não estão bem distribuídos.

Os modelos de arvore de decisão são modelos simples que não levam em consideração todos os pressupostos de não violação dos modelos de regressão múltipla por OLS como a normalidade, que neste dos métodos OLS muitas vezes se faz necessários transformações nas variáveis target e explicativas.

A Figura 03 apresenta a correlação entre todas as variáveis que foram propostas para o modelo de risco de credito, é possível observar que MARTUDUE é a que tem maior correção com VALUE.

Figura 03 – Correlação entre as 10 variáveis determinantes para o risco de crédito.



Fonte: Elaboração própria a partir do banco de dados disponibilizado pela plataforma Kaggle.

Com o intuito de justificar o uso das variáveis e, também, verificar a existência de multicolinearidade entre as variáveis, foi gerada uma matriz de correlação, como pode ser visto na imagem de modo geral não houve altos coeficientes de correlação

entre pares de variáveis com exceção de MARTUDUE e VALUE. A significância dos coeficientes foi avaliada considerando o nível de 5% como limite de Pvalor.

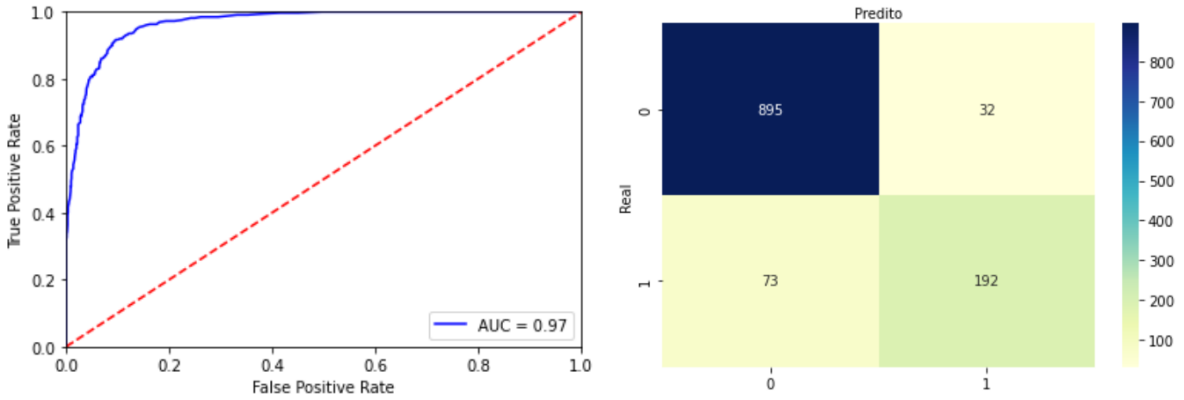
As variaveis MARTUDUE e VALUE foram mantidas no modelo apesar da correlação forte devido a importância individual de suas informações para o modelo.

5 RESULTADOS

O universo de dados desta base foi separado em 75% para treino e 25% para teste para instanciar o primeiro modelo que será o Random Forest. Neste modelo, foram definidos 200 estimadores para aumentar a precisão do modelo, o que obteve por sua vez um resultado de acurácia de 0.9107382550335571. O resultado pode ser considerado um bom resultado, mas deve ser avaliado juntamente com outras métricas de avaliação de modelo, o que para tal será utilizado um método de medida conhecido como AUROC - Area sob a curva ROC (BATISTA, PRATI e MONARD, 2008).

A curva ROC (Receiver Operating Characteristic) é uma curva que representa o desempenho de um modelo de classificação binária em diferentes níveis de threshold (ponto de corte que define a classe positiva e a classe negativa). Ela é construída a partir da relação entre a taxa de verdadeiros positivos (TPR - True Positive Rate) e a taxa de falsos positivos (FPR - False Positive Rate), em diferentes níveis de threshold. Na Figura 04 apresenta não só a sensibilidade da Curva ROC com o modelo Random Forest como também a Matriz de Confusão desses dados (JUNIOR, 2022).

Figura 04 – Curva ROC e Matriz de Confusão para os dados originais.



Fonte: Elaboração própria a partir do banco de dados disponibilizado pela plataforma Kaggle.

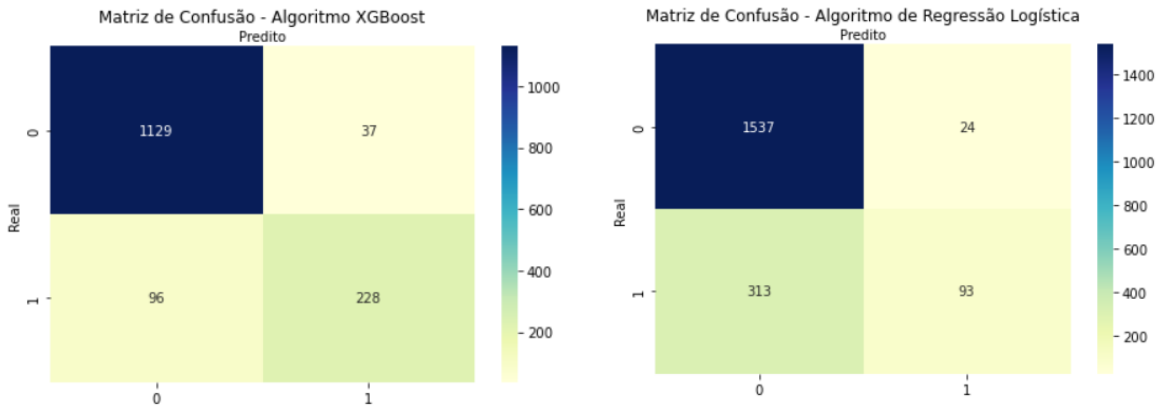
Portanto, a Matriz de Confusão mostra que o modelo de Random Forest teve altos índices de acerto de acordo com a Accuracy, Sensibilidade, Especificidade e

Eficiência. Na prática, a sensibilidade e a especificidade variam em direções opostas. Isto é, geralmente, quando um método é muito sensível a positivos, tende a gerar muitos falso-positivos, e vice-versa.

O modelo avaliado apresentou uma acurácia de 91,07%, o que significa que o modelo acertou cerca de 91% das previsões. A precisão foi de 85,71%, o que significa que das amostras classificadas como positivas, cerca de 86% foram corretamente classificadas. O recall (ou sensibilidade) foi de 72,46%, o que significa que o modelo identificou corretamente cerca de 72% das amostras positivas.

Mantendo a mesma divisão de teste e treino utilizada no modelo Random Forest, a Figura 05 apresenta a Matriz de confusão com base no modelo XGBoost e Regressão Logística.

Figura 05 – Curva ROC e Matriz de Confusão para os dados originais.



Fonte: Elaboração própria a partir do banco de dados disponibilizado pela plataforma Kaggle.

A primeira matriz de confusão foi elaborada com base no algoritmo XGBoost e indica que o modelo teve um bom desempenho na classificação, com 1129 verdadeiros positivos (VP) e 228 verdadeiros negativos (VN), enquanto teve 37 falsos positivos (FP) e 96 falsos negativos (FN). A acurácia do modelo foi de 91,61%, indicando que a grande maioria das classificações foi correta. A precisão das classificações positivas (predições de que o empréstimo seria ruim) foi de 86% e o recall foi de 70,4%. O f1-score foi de 77,7%, que é uma média harmônica da precisão e do recall.

Já a segunda matriz de confusão foi elaborada com base no modelo de Regressão Logística e indica um desempenho menos satisfatório do modelo, com 1537 VP e 93 VN, e 24 FP e 313 FN. A acurácia foi de 82,87%, o que ainda é considerado razoável, mas indica uma queda em relação à primeira matriz. A precisão

foi de 79,5% e o recall foi de 22,9%. O f1-score foi de 35,5%, novamente uma média harmônica da precisão e do recall.

6 CONSIDERAÇÕES FINAIS

O problema principal deste trabalho era entender e tentar prever comportamento de clientes que possam se tornar inadimplentes de uma entidade financeira que concede empréstimos, com base na análise de dados do último ano.

Com base na matriz de confusão gerada através dos três modelos utilizados, se o objetivo for maximizar a taxa de acerto de previsões positivas (VP e FP), o modelo utilizando a Random Forest pode ser a melhor escolha, já que possui um valor de 0,72 de recall para a classe positiva, enquanto a matriz que utiliza o modelo XGBost possui um recall de 0,70 para a mesma classe.

Por outro lado, se o objetivo for minimizar o número de falsos positivos (FP), o modelo XGBoost pode ser mais adequado, já que possui um menor número de FP (37) em comparação com a matriz do Random Forest (32). Considerando o fator acurácia, o modelo com Regressão Logística apresenta um desempenho melhor que os outros modelos.

Por fim, pelo fato do modelo XGBoost possuir um menor grau de falsos positivos, estar próximos de um algo acerto com relação as previsões positivas (tendo uma mínima diferente entre ele e o Random Forest) e possui um alto valor de acurácia, para este problema o melhor modelo que se adequa é o XGBoost.

7 FONTES CONSULTADAS

ARANTES, Rodrigo Diniz ; A influência da alienação fiduciária no crédito imobiliário. 2016.

BATISTA, G. E. A. P. A. ; PRATI, R. C. ; MONARD, M. C. ; Curvas ROC para avaliação de classificadores – Revista IEEE América Latina, 2008 - academia.edu.

CAMARGO, Marcos; CAMARGO, Mirela; ARAÚJO, Elisson; A inadimplência em um programa de crédito de uma instituição financeira pública de Minas Gerais: uma análise utilizando Regressão Logística - REGE-Revista de Gestão, 2012 – Elsevier.

DACCORONE, Fernanda; PORTO, Paola Torneri; A Emissão de Certificados de Recebíveis Imobiliários (CRIs) no Mercado de Home Equity – 2021 - ideas.repec.org.

GONÇALVES, E. B. ; GOUVEA, M. A. ; Análise de risco de crédito com o uso de regressão logística - Revista Contemporânea de Contabilidade, 2013 - redalyc.org.

JUNIO, Viela. Determinação das Métricas Usuais a partir da Matriz de Confusão de Classificadores Multiclasses em Algoritmos Inteligentes nas Ciências do Movimento Humano - Revista CPAQV - Centro de Pesquisas Avançadas em Qualidade de Vida - CPAQV Journal – 2022.

KAGGLE ; HMEQ_Data. Link de acesso : https://www.kaggle.com/datasets/ajay1735/hmeq-data?resource=download. Acessado em fev/2023.