Projeto de Pesquisa e Planejamento de Atividades

Aluno: Gustavo Fernandes Cruz		**Data início curso**: __11/05/2021___
Orientador: Gabrielle Lombardi		**Defesa em:** 06/12/2022
Curso: MBA Data Science e Analytics	**Modalidade**: Distância	Turma: 211

1. Título do projeto *(Inicial)*

Construção de modelo para análise de risco de crédito usando técnicas de machine learning

2. Introdução

(Contextualizar e apresentar a problemática do tema geral, ou seja, a importância do tema proposto e sua relevância. O texto deverá ser escrito de forma impessoal e toda informação utilizada deverá ser embasada por meio de trabalhos de fontes confiáveis com as devidas citações dos autores)

Uma operação de crédito, segundo o Banco Central do Brasil (BACEN), é quando uma pessoa ou empresa recebe dinheiro assumindo o compromisso de pagar, no futuro, o valor disponibilizado acrescido de juros e encargos.

O risco de crédito, possibilidade de ocorrência de perdas associadas ao não cumprimento pelo tomador de crédito, ou contraparte, de suas respectivas obrigações financeiras nos termos pactuados (BACEN, 2009), apresenta crescente uso de métodos estatísticos para classificar se uma pessoa é “bom” ou “mau” pagadora. Esta classificação possui o termo “Score de Crédito” e uma série de problemas particulares desta área vem sendo examinados em que os métodos estatísticos utilizados estão sempre sendo revistos (Hand e Henley, 1997).

Os primeiros modelos de risco de crédito foram elaborados entre 1950 e 1960, e estes eram desenhados a partir de Análise Discriminante, sugerido por Fisher (1936), utilizando funções de discriminação. Com a evolução das metodologias estatísticas, a modelagem é feita a partir de métodos com uma abordagem mais sofisticada como Regressão Logística, Random Forest, XGBoost, Support Vector Machines e Redes Neurais (Ferreira et al. 2015). Com o recente crescimento de 2,9%, em 2020 em relação ao mesmo período em 2019 da concessão de crédito à pessoas físicas (BACEN 2021), as instituições financeiras necessitam de modelos que façam previsões menos arbitrárias, para evitar concessão de crédito a uma pessoa inadimplente evitando assim prejuízo financeiro a instituição. Esses modelos, precisam ser ágeis, consistentes e principalmente assertivos, já que nenhum humano consegue fazer uma série de avaliações rapidamente e com várias variáveis ao mesmo tempo. Esses modelos consistem em efetuar classificação a partir de variáveis que contemplam desde cunho sócio demográfico como região onde reside, idade e renda até variáveis chamadas transacionais como quantidade de crédito solicitado nos últimos 12 meses.

Apesar dos modelos serem métodos matemáticos, podem ocorrer casos em que se recusa um bom pagador e aceita-se um mau pagador. Isto acontece pois nenhum sistema de classificação consegue capturar todas as características necessárias para ter uma classificação perfeita (CHAIA 2003). Estes modelos oferecem, além de uma classificação binária, também um valor de probabilidade do indivíduo ser bom ou mau pagador, chamado Probability of Default (probabilidade de negligência). Segundo Delianedis e colaboradores (2003), ‘default’ é definido como falha em cumprir com obrigação contratual, no caso, o contrato de concessão do crédito entre a instituição financeira e a pessoa física. Tem-se então, como entrega do modelo, uma distribuição de probabilidades de uma pessoa não cumprir com a obrigação de pagar o crédito que lhe foi concedido. Existe a possibilidade da

construção de faixas de crédito, onde são consideradas faixas de probabilidade, possibilitando uma maior flexibilidade para a concessão.

Há cada vez mais direcionamento dos termos de concessão de crédito em tempo real (EXPERIAN, 2021) e os modelos de machine learning estão sendo explorados de diversas maneiras para que essa mecânica tenha dinamismo aliado à assertividade.

3. Objetivo

(Qual o objetivo principal do trabalho, ou seja, qual pergunta deve ser respondida ao final da sua pesquisa)

Construir um modelo de machine learning capaz de identificar se um indivíduo (pessoa física) será adimplente ou inadimplente no ato da concessão de crédito

4. Material e Métodos

(Descrever o(s) método(s) de coleta de dados e a(s) ferramenta(s) de análise a ser(em) utilizada(s) no trabalho de conclusão de curso, ou seja, como será a condução da pesquisa e a forma de obtenção dos resultados, por exemplo, fontes de dados, técnicas, procedimentos, índices, entre outros)

A base de dados que será utilizada neste trabalho foi extraída do repositório de dados online chamado Kaggle e tem como título 'German Credit Risk'. Esta base contém 1.000 observações e 10 variáveis, sendo uma delas a variável resposta. A variável resposta é binária sendo 1 correspondendo ao adimplente e 0 corresponde ao inadimplente. Dentre as variáveis explicativas, 5 são quantitativas e 4 qualitativas. A base está alocada no link https://www.kaggle.com/datasets/uciml/german-credit

A ferramenta utilizada neste trabalho será o software R, em conjunto com o pacote Tidymodels. A modelagem irá seguir as etapas de: tratamento e limpeza dos dados, análise exploratória dos dados, escolha dos modelos, desenho da receita e workflow do modelo, aplicação do cross validation e coleta dos resultados. Os modelos utilizados serão Regressão Logística, Random Forest, XGBoost, Support Vector Machines e Redes Neurais, a fim de efetuar classificação binária sobre a variável resposta e distinguir adimplente e inadimplente.

Se necessário, serão utilizados métodos de machine learning não supervisionados, como Análise de agrupamentos K-means, para criação de variáveis latentes que tenham maior importância do que as originais da base de dados.

Somado aos métodos anteriores, serão utilizados: SMOTE (Synthetic Minority Oversampling Technique) para balanceamento dos dados da variável resposta, K-fold Cross validation para validação cruzada e Grid Search para otimização dos hiperparâmetros dos modelos.

Por fim, será efetuado a coleta e análise das métricas utilizadas para avaliar modelos de machine learning supervisionados de classificação, que servirão de base para escolha do modelo final. Segundo (BALDI et al. 2000) para medir a qualidade de um modelo de classificação de uma variável binária utiliza-se dos indicadores curva ROC (Receiver Operating Characteristics), sensibilidade e especificidade, dependendo diretamente do ponto de corte escolhido. A sensibilidade é a probabilidade de prever corretamente um exemplo positivo (bom pagador) enquanto especificidade é definida como a probabilidade de prever corretamente um exemplo negativo (mau pagador). O gráfico da curva ROC resume os resultados da sensibilidade em função da taxa de falsos positivos (1- especificidade) para todos os possíveis pares de valores. A área abaixo desta curva indica a capacidade do modelo de distinção entre "bom" e "mau" pagadores (FARAGGI et al., 2002). Junto a estas métricas, a fim de encontrar o par de pontos ótimo de sensibilidade e especificidade, a métrica índice J de Youden será explorada (Schisterman, 2007), com o objetivo de encontrar um modelo que preveja bem tanto os 'bons' pagadores quanto os 'maus' pagadores. Segundo (BLÖCHLINGER, 2006), para efetuar a discriminação entre um modelo bom ou

ruim de risco de crédito, estas são as métricas utilizadas, principalmente a curva ROC e a área embaixo da curva (AUC).

5. Resultados Esperados

(Descrever os resultados que são esperados após a realização da coleta e análise dos dados, ou seja, quais resultados são esperados ao final da pesquisa)

Espera-se com este trabalho que o modelo construído faça distinção entre adimplentes e inadimplentes para concessão de crédito

6. Cronograma de Atividades

(Adicionar as "Atividades planejadas", assim como o período (tempo para desenvolver cada atividade) planejado para a realização de cada atividade, sendo que deverá ser adequado ao calendário de entregas das etapas do trabalho de conclusão de curso definido pela Coordenação. Marcar com um "x" a coluna que corresponde ao período planejado para desenvolver cada atividade planeja)

Atividades planejadas	**Mês**									
	MAR	**ABR**	**MAI**	**JUN**	**JUL**	**AGO**	**SET**	**OUT**	**NOV**	**DEZ**
Definição do tema e banco de dados	**12/03**									
Entrega do projeto de pesquisa v1		**02/04**								
Entrega do projeto de pesquisa v2		**09/04**								
Submissão do projeto para o PECEGE		**15/04**								
Análise dos dados		**16-30**	**1-15**							
Resultados Preliminares			**1-15**							
Conversa com Orientadora			**12-15**							
Ajuste do material e métodos				**05-15**						
Descrição dos resultados				**17-24**						
Conversa com Orientadora				**25-30**						
Ajuste final dos Resultados					**1-15**					
Entrega dos resultados preliminares						**15/08**				
Ajuste dos resultados com figuras e tabelas						**18-25**				
Discussão dos resultados						**25-30**	**01-15**			
Conclusão							**16-20**			
Resumo/Abstract/Agradecimentos							**21-30**			
Entrega do TCC v1								**06/10**		
Entrega do TCC v2								**11/10**		
Submeter PECEGE								**17/10**		
Agendar Defesa								**18/10**		
Fazer Apresentação								**25-**	**01-**	

								31	**05**	
Estudar Apresentação									**15-30**	
Defender										**06/12**

Projeto de Pesquisa; Resultados Preliminares; Entrega do Trabalho de Conclusão de Curso; Entrega da Apresentação da Defesa

7. Referências Bibliográficas

(Listagem das bibliografias citadas no projeto de pesquisa, seguindo rigorosamente as Normas do MBA USP ESALQ – Consulte o manual de "Normas para Elaboração do Trabalho de Conclusão de Curso" disponível no Sistema TCC)

Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). *Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation. Management Science, 49(3), 312–329.*

Banco Central do Brasil [BACEN]. 2021. Evolução Recente do Crédito no SFN. Disponível em < https://www.bcb.gov.br/content/acessoinformacao/covid19_docs/Evolucao_Recente_do_Credito.pdf > Acesso em: Abril/2022

Banco Central do Brasil [BACEN] .2021. Relatório de Estabilidade Financeira. Disponível em < https://www.bcb.gov.br/content/publicacoes/ref/202110/RELESTAB202110-refPub.pdf> Acesso em: Abril/2022

Blöchlinger, A., & Leippold, M. (2006). Economic benefit of powerful credit scoring. Journal of Banking & Finance, 30(3), 851–873.

Bruce A., Bruce P.2019. Estatística prática para cientistas de dados: 50 conceitos essenciais Capa comum. Editora Alta Books.

Chaia, Alexandre Jorge. *Modelos de gestão do risco de crédito e sua aplicabilidade ao mercado brasileiro*. Diss. Universidade de São Paulo, 2003.

Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). *Random Forests. Ensemble Machine Learning, 157–175.*

Delianedis, Gordon, and Robert L. Geske. "Credit risk and risk neutral default probabilities: information about rating migrations and defaults." *Available at SSRN 424301* (2003).

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics, 35(5-6), 352–359.

Experian Information Solutions.2021. Navigating a new era of credit risk decisioning. Disponível em < https://www.experian.com.vn/wp-content/uploads/2021/07/Decisioning_Report_2021.pdf >. Acesso em Abril/2022

Faraggi, D., & Reiser, B. (2002). *Estimation of the area under the ROC curve. Statistics in Medicine, 21(20), 3093–3106.*

Fávero, L.P., Belfiore, P. 2021.Manual de Análise de Dados - Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata®. 1ª edição.GEN LTC.Rio de Janeiro. RJ. Brasil.

Ferreira, Paulo H., Francisco Louzada, and Carlos Diniz. "Credit scoring modeling with state-dependent sample selection: A comparison study with the usual logistic modeling." *Pesquisa Operacional* 35.1 (2015): 39-56.

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). *Random Forests for land cover classification. Pattern Recognition Letters, 27(4), 294–300.*

Grolemund,G. 2014. Hands-on Programming with R. 1ª edição. O'Reilly Media.

Hand, D. J., & Henley, W. E. (1997). *Statistical Classification Methods in Consumer Credit Scoring: a Review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523–541.*

Hernández-Orallo, J., Flach, P., & Ferri Ramírez, C. (2012). A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, *13*, 2813-2869.

Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.-E., He-Guelton, L., & Caelen, O. (2018). *Sequence classification for credit-card fraud detection. Expert Systems with Applications, 100, 234–245.*

Provost, F., Fawcett, T. Robust Classification for Imprecise Environments. *Machine Learning* 42, 203–231 (2001)

Refaeilzadeh P., Tang L., Liu H. (2016) Cross-Validation. In: Liu L., Özsu M. (eds) Encyclopedia of Database Systems. Springer, New York, NY

Ruopp, M. D., Perkins, N. J., Whitcomb, B. W., & Schisterman, E. F. (2008). *Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection. Biometrical Journal, 50(3), 419–430.*

Schisterman, E. F., Faraggi, D., Reiser, B., & Hu, J. (2007). Youden Index and the optimal threshold for markers with mass at zero. Statistics in Medicine, 27(2), 297–315. doi:10.1002/sim.2993

Zou Q, Qu K, Luo Y, Yin D, Ju Y and Tang H (2018) Predicting Diabetes Mellitus With Machine Learning Techniques. Front. Genet. 9:515.