Proposta de um framework utilizando regressão logística binomial para aumento na conversão de receita: um estudo de caso para empresas prestadora de serviços e soluções tecnológicas

Ivan Ajala[1]; Marcos dos Santos[2]

[1] Pós-graduando em Data Science. Via Transversal Sul, 169, T 01 AP04 - Jardim Novo Osasco; 06045 – Osasco, São Paulo, Brasil
[2] Instituto Militar de Engenharia (IME). Doutor. Praça General Tibúrcio, 80 – Urca; 22290-270 Rio de Janeiro, RJ, Brasil
*autor correspondente: ivan_ajala@hotmail.com

Proposta de um framework utilizando regressão logística binomial para aumento na conversão de receita: um estudo de caso baseado em empresas prestadora de serviços e soluções tecnológicas

Resumo

Esta pesquisa busca prever e identificar quais orçamentos tem menor probabilidade de serem aprovados, permitindo eventuais ajustes nos orçamentos antes mesmo de serem enviados aos clientes e sinalizar para o time de vendas quais dentre os orçamentos deverão direcionar os seus esforços, aumentando as chances de aprovações e consequentemente aumentar a receita e competividade da empresa em relação a seus concorrentes. O "dataset" utilizado contém registros pseudoanonimizadas e correspondentes meses de agosto de 2020 até dezembro de 2022. Após a análise exploratória dos dados, identificou-se quais variáveis mostraram possível dependência com a variável dicotômica binária (y), onde 1 significa que o orçamento (evento) foi aprovado e 0 (não evento) não aprovado pelo cliente. Para identificar as melhores variáveis para o modelo, foi utilizadas técnicas como Árvore de Decisão, "Random Forest" e "Gradient Boosting". Para verificar os possíveis efeitos devido o desbalanceamento da base de dados, utilizou-se algoritmos como "NearMiss" e SMOTE. Por fim foi realizado uma prévia comparação com outro modelo classificador KNN.
Palavras-chave: Aumento de receita, Modelos de classificação, Regressão logística binomial, NearMiss, SMOTE, Machine Learning, Random Forest, KNN.

Introdução

No Brasil, de acordo a última Pesquisa Anual de Serviços – PAS, realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE, 2022), estima-se que existam por volta de 1,4 milhões de empresas prestadoras de serviços não financeiros que geram uma receita operacional líquida de R$ 1,8 trilhões empregando por volta de 12,5 milhões de pessoas. O setor de prestação de serviços em um modo geral, caracteriza-se por atividades heterogêneas desde o porte das empresas, à remuneração média assim como a intensidade no uso tecnológico, sendo que nas últimas décadas, o desempenho geral das atividades destas empresas vem destacando-se tanto pelo dinamismo como pela crescente participação na composição da econômica brasileira. Dentre as 34 atividades de serviços que compõe a PAS, a prestação de serviços em tecnologia da informação, está entre as três atividades mais relevantes do setor de serviços com uma participação na receita operacional líquida de 9,6%, ficando atrás somente para as atividades de transporte de cargas e serviços técnico-profissionais, com 12,1% e 11,4% de participação respectivamente.

Já o comportamento do cliente no mercado atual exige cada dia mais que os serviços sejam realizados no menor tempo, com a melhor qualidade e principalmente com o menor custo possível. Com a globalização e todo desenvolvimento tecnológico, de acordo com Kotler e Keller (2007), as empresas devem conhecer os seus clientes, tornando os seus produtos e serviços adequados a esses por meio de processos criativos e que entreguem também valor ao cliente, administrando uma boa relação se beneficiando, captando e fidelizando seus

clientes. Conhecer o seu cliente não é apenas saber a sua atividade de negócios e atuação no mercado. Deve-se entender o porquê e qual finalidade está sendo solicitado o serviço, qual será o seu publico alvo, qual é o seu budget para a contratação do serviço assim como identificar quais são os seus concorrentes assim como os seus diferenciais quando comparados aos seus negócios.

Com o crescente e real avanço tecnológico em diversas áreas, junto com a proposta de uma nova era industrial, surge cada vez mais a necessidade em planejar novos sistemas com o propósito de adaptar o ambiente de produção e mercadológico a este novo momento da indústria. Momento este conhecido como a Quarta Revolução Industrial ou Indústria 4.0, expressões usadas para definir o englobamento de tecnologias para automação e troca de dados, utilizando conceitos de sistemas ciber-físicos (Internet das coisas e Computação em Nuvem), tendo como foco melhoria da eficiência e produtividade dos processos. Em 2020, em virtude da COVID-19, muitas empresas se viram necessárias em migrar ou se antecipar para este conceito industrial, sendo que em alguns casos, foram até obrigadas para a manutenção da existência de seus negócios. Conforme Ardito et al. (2019), a quarta revolução industrial é um sucesso e não um hype momentâneo e desde a disseminação do termo “Indústria 4.0” em 2011 e a transformação digital exigida pela Indústria 4.0 imediatamente chamou a atenção de industriais e governos em todo o mundo.

Com o comportamento e exigências dos clientes no mercado atual e o crescimento na quantidade de empresas, principalmente na área de serviços, a Industria 4.0 proporciona novas formas de soluções para os problemas coorporativos, proporcionando benefícios, tais como: a redução dos custos, operações em tempo real, manufatura modular podendo se adaptar com maior flexibilidade, operações integradas e otimização. Dentre os conceitos associados à inteligência artificial, existem inúmeras tecnologias que envolvem a Industria 4.0, entre elas, a “machine learning” também conhecida como “aprendizado de máquina”, que possuem excelentes modelos preditivos, onde com eles, as empresas podem entender as oportunidades e principais riscos aos seus negócios e antecipar-se tanto aos concorrentes como até mesmos aos seus próprios clientes, ou seja, uma excelente ferramenta para ajudar as empresas no processo de tomada de decisões tanto estratégicas como operacionais.

Material e métodos

Esta pesquisa utilizou a base de dados do sistema proprietário BWX de uma empresa de médio porte situada em São Paulo, Brasil, que presta serviços de soluções tecnológicas tendo como atividade principal do negócio os serviços de traduções. A base dados foi disponibilizada com a finalidade somente para utilização desta pesquisa, onde os dados

passaram por pseudoanonimização de modo que fossem preservados as informações referentes aos clientes. A base de dados esta disponível no diretório GitHub e pode ser acessado através do link https://github.com/IvanAjala/TCC_USP_RegLog.

A pesquisa tem como proposito identificar e prever dentre os orçamentos realizados na plataforma BWX, quais não serão aprovados de modo que o time de vendas possa concentrar os seus esforços e agir mais de perto nesses orçamentos e consequentemente aumentar as chances de aprovação.

Sobre a plataforma BWX

O BWX é uma plataforma em nuvem a qual permite ao cliente realizar toda a gestão de seu conteúdo a ser traduzido de modo compartilhado com toda equipe interna, eliminando toda complexidade de gerenciamento dos projetos em atividades simples, previsíveis e transparente.

As solicitações dos orçamentos podem ser feitas pelos clientes através de e-mail onde o gerente de projetos integrará a solicitação a ferramenta BWX e seguirá com a elaboração do orçamento ou o próprio cliente pode fazer a solicitação pela plataforma BWX realizando o upload do material a ser traduzido, o idioma de origem e destino e a data a qual necessita o trabalho.

Disposição da base de dados

A base de dados que será utilizada para a construção do modelo foi disponibilizada pela área de tecnologia da empresa no formato .xlsx, sendo composta por 50816 orçamentos e 9 variáveis.

O banco de dados foi carregado no “Python” versão 3 utilizando o “Jupyter Notebook”, onde entre as variáveis, duas são quantitativas numéricas e sete qualitativas categóricas, conforme pode ser visto na Tabela 1.

Tabela 1. Informações inicial da base de dados

Column	Non-Null Count	Dtype
idQuote	Non-Null Count	object
idCurrency	Non-Null Count	object
idSource	Non-Null Count	object
idTarget	Non-Null Count	object
creationDate	Non-Null Count	int64
status	Non-Null Count	object
totalCost	Non-Null Count	float64
idClient	Non-Null Count	object
idCreator	Non-Null Count	object

Fonte: Dados originais da pesquisa

Dicionário da base de dados

Antes mesmo de qualquer aplicação de técnicas estatísticas é de suma importância avaliar e entender o que representa cada uma das variáveis que compõe o banco de dados (Palmeira et al., 2020). A Tabela 2 apresenta uma breve descrição de cada uma das variáveis que compõe o banco de dados de acordo como consta no dicionário de variáveis do BWX.

Tabela 2. Dicionário da base de dados

Nome da variável	Tipo da variável	Descrição da variável
idQuote	object	Variável categórica de identificação única para cada orçamento.
idCurrency	object	Variável categórica referente a moeda a qual o orçamento foi criado.
idSource	object	Variável categórica referente ao idioma de origem do material a ser orçado.
idTarget	object	Variável categórica referente ao idioma de destino do material a ser orçado.
creationDate	int64	Variável numérica no formato timestamping contendo as informações referentes a data o qual o orçamento está sendo criado.
totalCost	float64	Variável numérica referente ao valor do orçamento.
idCreator	object	Variável categórica referente contendo o nome do criador do orçamento.
idClient	object	Variável categórica referente contendo o nome do cliente o qual está solicitando o orçamento.
status	Object	Variável categórica referente ao status do orçamento, sendo APPROVED para os orçamentos aprovados e NOT APPROVED para os orçamentos não aprovados.

Fonte: Dados originais da pesquisa

Resumo da base de dados

Após entender o que cada uma das variáveis significa é importante ter uma descrição estatística da base de dados para então iniciar a análise exploratória dos dados. A Tabela 3 demonstra o conjunto estatístico descritivo das variáveis quantitativas numéricas em relação ao total de valores, valores mínimos e máximos auxiliando na evidenciação de possíveis

outliers, os quartis das distribuições, a média e o desvio padrão, assim como o resumo das variáveis categóricas em relação a cardinalidade, correlação, balanceamento, valores ausentes e duplicados.

Tabela 3. Resumo estatístico das variáveis numéricas e categóricas

Descrição	Total
Number of variables	9
Number of observations	50816
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%

Variável	Tipo	Valores	Notas
totalCost	Numerica	γ1 = 24.15	Enviesado
		count	50.816.000
		mean	949.361.474
		std	6.329.862.852
		min	0.010000
		0,25	4.568.000
		0,5	36.052.800
		0,75	215.865.895
		max	388.773.000.000
idQuote	Categorica	50673	Alta Cardinalidade
idSource	Categorica	56	Alta Cardinalidade, Alta Correlação com idCurrency e Desbalanceado
idTarget	Categorica	1624	Alta Cardinalidade
idClient	Categorica	903	Alta Cardinalidade e Desbalanceado
idCreator	Categorica	186	Alta Cardinalidade e Desbalanceado
idCurrency	Categorica	2	Alta Correlação com idSource
status	Categorica	2	Variável dependente
creationDate	Numerica	46509	min. 27/08/2020 máx. 30/12/2022

Fonte: Dados originais da pesquisa

Análises exploratórias da base de dados

A análise exploratória dos dados nada mais é do que uma abordagem analítica do conjunto de dados visando encontrar geralmente de forma gráfica informações ocultas, as variáveis mais importantes e as suas tendencias, comportamentos anômalos, testar a validade das hipóteses assumidas, auxiliando na escolha das melhoras variáveis para o modelo o qual será utilizado. De acordo com Motta et al. (2022), todos os pesquisadores deveriam iniciar a sua análise pela exploração dos dados para posteriormente definir qual melhor modelo se aplica ao determinado problema de pesquisa.

Para esta pesquisa a análise exploratória dos dados será realizada no “Python” com auxílio das bibliotecas “Pandas” e “Seaborn”, sendo a primeira uma biblioteca específica para trabalhar com dados e a segunda específica para visualizacao gráfica de dados. Assim como o banco de dados utilizado nesta pesquisa, o código fonte desenvolvido no “Python” e está disponível no diretório GitHub no link https://github.com/IvanAjala/TCC_USP_RegLog.

De acordo com a Tabela 3, a base de dados não possui dados duplicados e também dados faltantes e isso é um bom sinal uma vez que a ausência de dados pode ter um alto impacto na inferência estatística. Segundo Silva Júnior et el. (2019), a duplicação dos dados pode ocasionar em uma análise enviesada ocasionando superestimação da importância de certas informações. Por outro lado, a ausência deles pode levar o pesquisador a uma análise incompleta e subestimada, uma vez que a análise pode não refletir a realidade reduzindo a precisão dos resultados. Por isso é de suma importância que o pesquisador garanta que os dados estejam completos e corretos antes de seguir com a análise estatística dos mesmos.

A variável dependente originalmente é dicotômica, ou seja, com duas categorias distintas, sendo APPROVED o evento que será posteriormente codificada como 1 e NOT_APPROVED como não evento que será codificado como 0. De acordo com a Figura 1, dentre os 50816 orçamentos da base de dados, 44060 são APPROVED e 6756 NOT_APPROVED, ou seja, 86,7% dos orçamentos foram aprovados, evidenciando um possível desbalanceamento o que pode fazer com que no caso desta pesquisa, o modelo aprenda mais com a classe majoritária APPROVED.

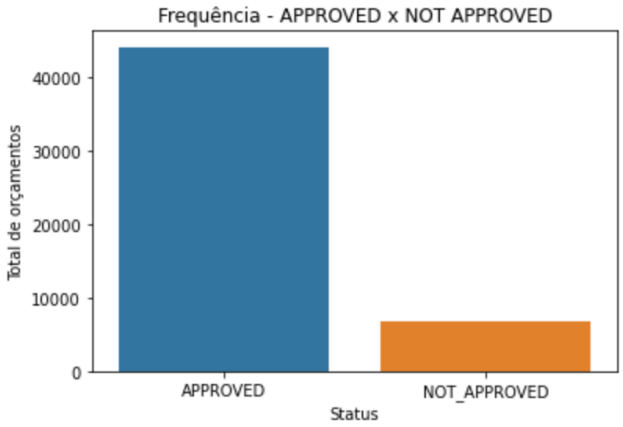

Figura 1 – Total orçamentos aprovados e não aprovados
Fonte: Dados originais da pesquisa

Ainda na Tabela 3, nota-se que as variáveis qualitativas categóricas idSource, idTarget, idClient e idCreator possuem uma alta cardinalidade, ou seja, uma grande parte dos valores em cada uma dessas variáveis são representados por poucas ou uma única categoria.

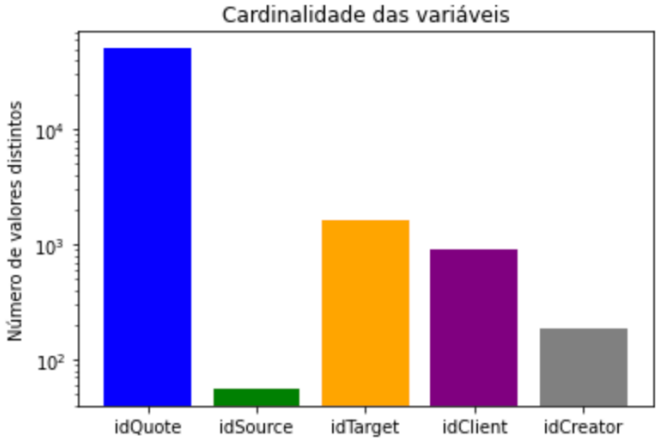

Figura 2 – Cardinalidade entre as variáveis qualitativas categóricas
Fonte: Dados originais da pesquisa

Já as variáveis idSource e idCurrency mostraram uma correlação positiva moderada de 0.673, ou seja, à medida que uma dessas variáveis aumenta a outra consequentemente tende a aumentar e vice-versa o que pode causar multicolinearidade, o que segundo Gonçalves et al. (2022) pode ocasionar em problemas de interpretação dos resultados consequente da imprecisão na estimativa dos coeficientes.

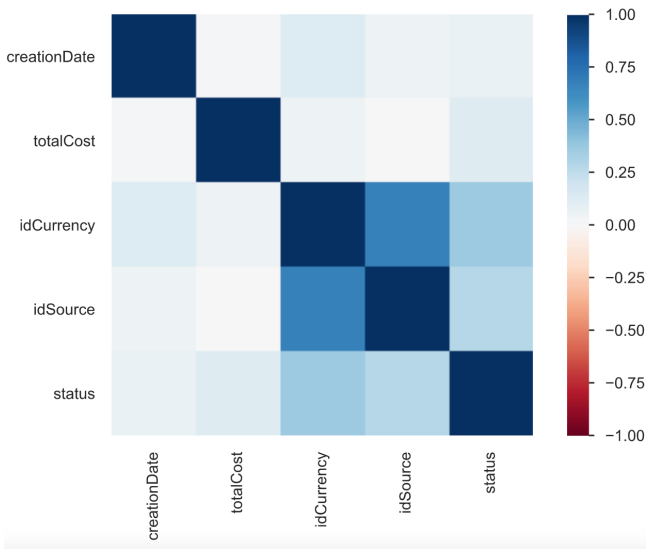

Figura 3 – Correlação entre as variáveis
Fonte: Dados originais da pesquisa

A variável quantitativa totalCost tem um coeficiente de assimetria (gamma 1) positiva de aproximadamente 24,15 (vide Tabela 3), indicando que os dados estão distribuídos de forma desigual em relação à média contendo a maioria dos dados concentrados em uma das extremidades da distribuição. Neste caso por ser positivo, há uma alta distribuição dos valores baixos e uma calda longa de valores altos conforme a Figura 4.

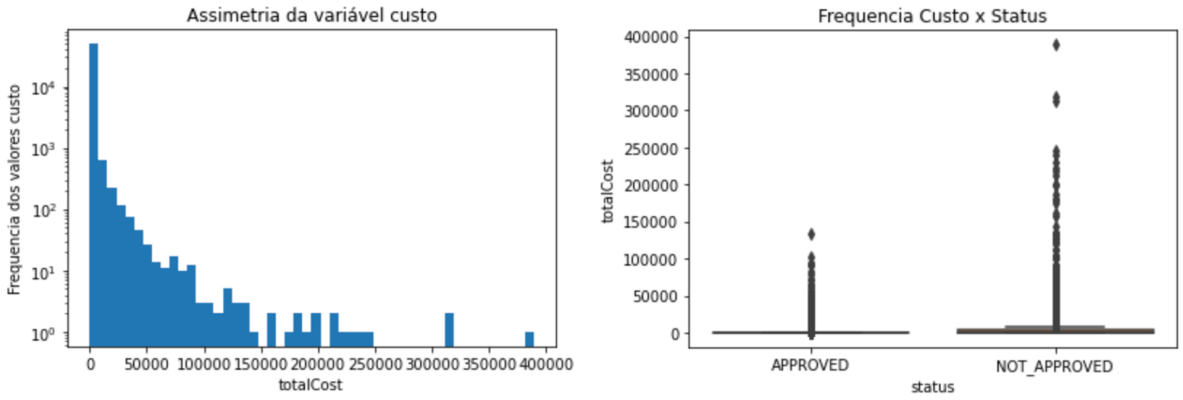

Figura 4 – Distribuição assimétrica do custo
Fonte: Dados originais da pesquisa

Tratamento e modelagem da base de dados

Pelo fato da base de dados conter sete variáveis qualitativas categóricas, faz-se necessário que estas variáveis sejam transformadas em “dummies”, uma prática bem comum usada frequentemente em analises estatísticas nas mais variáveis áreas, o que para Oliveira et al. (2019) esta transformação se faz necessário de modo que incorpore informações qualitativas em modelos quantitativos que consequentemente melhora a precisão dos

resultados permitindo análises comparativas. A Figura 5 contém a visualização parcial do banco de dados original o qual contém 50816 linhas e 9 colunas.

	idQuote	idCurrency	idSource	idTarget	creationDate	status	totalCost	idClient	idCreator
50806	B-21189-8019	USD	LS1	LT1757	1625710000000	NOT_APPROVED	300.20	CLI94	CRE1
50807	B-21162-1218	USD	LS1	LT1757	1623370000000	NOT_APPROVED	52.54	CLI94	CRE1
50808	B-21159-85227	USD	LS1	LT55	1623200000000	APPROVED	217.59	CLI94	CRE1
50809	B-21145-79747	USD	LS1	LT14	1621980000000	APPROVED	10.00	CLI94	CRE1
50810	B-21138-78366	USD	LS1	LT1757	1621370000000	NOT_APPROVED	1289.82	CLI94	CRE1
50811	B-21126-78923	USD	LS2	LT168	1620340000000	APPROVED	897.45	CLI3	CRE1
50812	B-21123-66246	BRL	LS1	LT14	1620070000000	NOT_APPROVED	374.04	CLI50	CRE22
50813	B-21123-57896	BRL	LS1	LT14	1620060000000	NOT_APPROVED	314.50	CLI50	CRE1
50814	B-21118-72176	USD	LS1	LT1767	1619640000000	NOT_APPROVED	848.20	CLI535	CRE1

Figura 5 – Previa visualização de parte do banco de dados original
Fonte: Dados originais da pesquisa

No processo de tratamento dos dados, foram excluídas as variáveis idQuote, visto que é o identificador único de cada orçamento, a variável creationDate por se tratar de um valor timestamping e não agregar significativamente ao modelo e por último a variável idSource devido a sua correlação moderada com a variável idCurrency de modo que se evite uma possível multicolinearidade que pode impactar na eficiência do modelo preditivo.

As demais variáveis qualitativas independentes foram transformadas em “dummies”, assim como a variável qualitativa dependente, onde os valores APPROVED foram transformados em 1 e NOT_APPROVED transformados em 0, ou seja, os eventos serão 1 e não eventos 0.

Após essa transformação em “dummies”, o banco de dados passou a conter 50816 linhas e 2716 colunas.

totalCost	Status	currency_usd	idTarget_LT1	idTarget_LT10	idTarget_LT100	idTarget_LT1000	idTarget_LT1001	idTarget_LT1002	idTarget_LT1003	...	idCreator_CR
897.45	1	1	0	0	0	0	0	0	0	...	
374.04	0	0	0	0	0	0	0	0	0	...	
314.50	0	0	0	0	0	0	0	0	0	...	
848.20	0	1	0	0	0	0	0	0	0	...	
755.78	0	1	0	0	0	0	0	0	0	...	

Figura 6 – Previa visualização banco após a transformação das variáveis em dummies
Fonte: Dados originais da pesquisa

O problema de pesquisa consiste em criar um modelo que baseado nas variáveis independentes idQuote, idCurrency, idSource, idTarget, creationDate, totalCost, idClient e idCreator explique a variável dependente Status, de modo que possa identificar entre todos os orçamentos, os quais tem a maior probabilidade em não serem aprovados pelos clientes,

proporcionando um “framework” para o time de vendas, que indique dentre todos orçamentos, quais de fato deverão ser aumentando a assertividade nos follow-ups junto aos clientes e consequentemente aumentando a conversão de aprovações dos orçamentos os quais hoje não são aprovados.

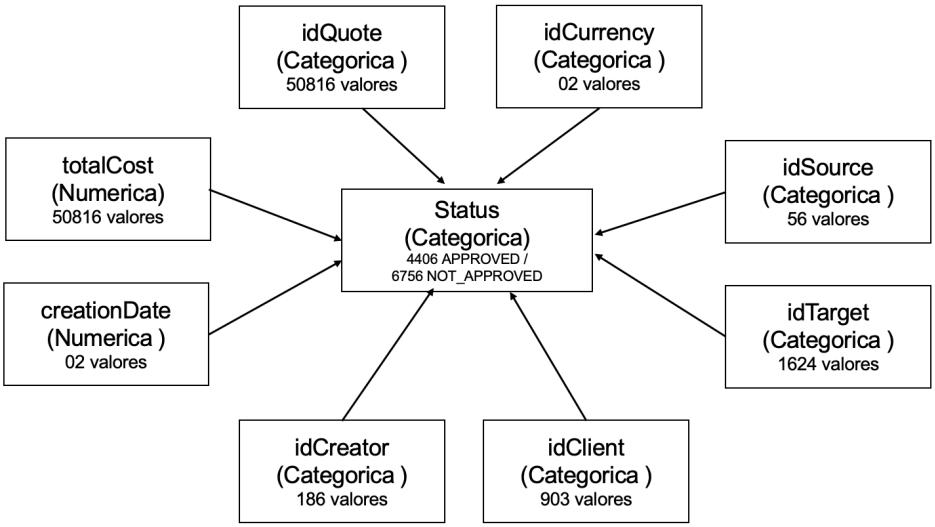

Figura 7 – Ilustração das variáveis explicativas e dependente do problema de pesquisa
Fonte: Criada pelo próprio autor com base nos exemplos das anotações das aulas

Uma vez que neste problema a variável dependente é categórica dicotômica, o modelo a ser proposto para resolução do mesmo é o de regressão logística binomial. Para uma melhor compreensão do que será abordado nesta pesquisa, cabe um adendo de que a regressão logistica tem como base principal a técnica de regressão linear, onde tem-se y como uma variável quantitativa dependente que é explicada pela relação linear das variáveis preditoras explicativas X como apresentado na formula a seguir:

$$Y_i = \alpha + \beta_1.X_{1i} + \beta_2.X_{2i} + \cdots + \beta_k.X_{ki} \quad (1)$$

A regressão linear pode ser Ilustrada graficamente em um plano cartesiano contendo as coordenadas x e y onde uma linha reta que passa o mais próximo possível pelos pontos no gráfico. Esta reta representa a linha de regressão linear, conforme a Figura 8.

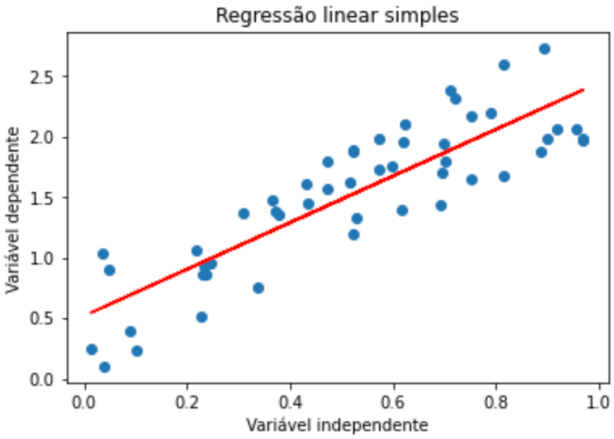

Figura 8 – Ilustração gráfica da regressão linear no plano cartesiano
Fonte: Criada pelo próprio autor no Python utilizando a biblioteca matplotlib

Fávero e Belfiore (2017), explica que diferente da regressão linear a qual a estimativa é por meio dos mínimos múltiplos quadrados onde a variável dependente é quantitativa, na regressão logistica o fenômeno a ser estudado se apresenta de forma qualitativa por uma ou mais variáveis dummy, dependendo da quantidade de possibilidade de respostas para esta variável dependente. Esta técnica quando aplicada corretamente, auxilia em estimar de forma correta a probabilidade de ocorrência de determinado fenômeno e apesar de ainda não ser tão utilizada em muitas das áreas do conhecimento humano, vem crescendo nas áreas de crédito bancário prevendo, por exemplo, a probabilidade de inadimplência para determinada pessoa, assim como na área da medicina prevendo a probabilidade de um individuo ter ou não uma doença.

Como o objetivo desta pesquisa é encontrar a probabilidade de ocorrência dos eventos de interesse, sendo orçamentos aprovados (Y=1) e principalmente a probabilidade da ocorrência do não evento que neste caso são os orçamentos não aprovados (Y=0), que Y representa de forma qualitativa dicotômica o orçamento, o modelo de regressão logístico binário pode ser contextualizado da seguinte forma, conforme descrito no Manual de analise de dados por (Fávero e Belfiore, 2017).

$$Z_i = \alpha + \beta_1.X_{1i} + \beta_2.X_{2i} + \cdots + \beta_k.X_{ki} \quad (2)$$

Na eq. (2), Z é o logito, α representa a constante, β_j (j = 1, 2, 3.... κ) são os parâmetros estimados para cada variável explicativa, X_j são as variáveis explicativas podendo ser métricas ou dummies e o subscrito i que inicia em 1 e vai até n, representa cada observação da amostra, sendo n é o tamanho da amostra. Para definir a expressão da probabilidade p a partir da ocorrência do evento de interesse para cada uma das observações em função do

logito Z_i se faz necessário definir o conceito de chance de ocorrência para um evento, ou $odds$ conforme a eq. (3):

$$chance\ (odds)_{\mathrm{Y_i=1}} = \frac{p_i}{1-p_i} \tag{03}$$

Apesar dos termos soarem ser a mesma coisa, o sentido de ambos são distintos, visto que a probabilidade é uma medida matemática expressa como um número entre o intervalo de 0 e 1 da chance de um evento ocorrer. Por outro lado, a chance pode ser expressa como uma porcentagem ou uma estimativa qualitativa baseada em fatores intangíveis como intuição e conhecimento prévio.

Na regressão logistica binaria o logito Z é definido como um logaritmo da chance, onde:

$$\ln\left(chance_{\mathrm{Y_i=1}}\right) = Z_i \tag{4}$$

então:

$$\ln\left(\frac{p_i}{1-p_i}\right) = Z_i \tag{5}$$

Como a ideia é definir a expressão para a probabilidade de ocorrência do evento de estudo em função do logito, matematicamente pode-se isolar o p_i a partir da eq. (5) resultando em:

$$\frac{p_i}{1-p_i} = e^{Zi} \tag{6}$$

$$p_i = (1 - pi).e^{Zi} \tag{7}$$

$$p_i.(1 + e^{Zi}) = e^{Zi} \tag{8}$$

ou seja, a probabilidade de ocorrência do evento é:

$$p_i = \frac{e^{Zi}}{1+e^{Zi}} = \frac{1}{1+e^{-Zi}} \tag{9}$$

e a probabilidade de não ocorrência do evento é:

$$1 - p_i = 1 - \frac{e^{Zi}}{1+e^{Zi}} = \frac{1}{1+e^{Zi}} \tag{10}$$

Então as equações (1) e (9) definem conseguem definir a expressão geral da probabilidade de ocorrência de um determinado evento que se apresente de forma dicotômica para uma observação i como pode ser visto a seguir:

$$p_i = \frac{1}{1 + e^{-(\alpha + \beta_1 . X_{1i} + \beta_2 . X_{2i} + \cdots + \beta_k . X_{ki})}} \quad (11)$$

A regressão logistica pode ser Ilustrada graficamente em um plano cartesiano contendo as coordenadas x que representa as variáveis independentes e o eixo y representando a probabilidade da variável dependente pertencer a uma categoria especifica, onde a curva em formato de S representa a relação não linear entre as variáveis, conforme ilustrado na Figura 9.

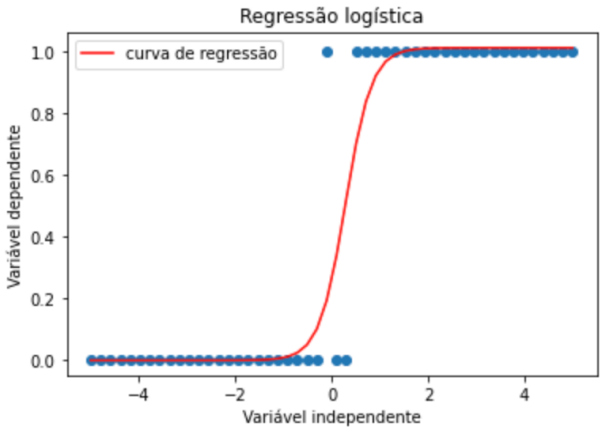

Figura 9 – Ilustração gráfica da regressão logistica no plano cartesiano
Fonte: Criada pelo próprio autor no Python utilizando a biblioteca matplotlib

Na Figura 10, fica claro o comportamento e natureza da relação entre as variáveis independentes e dependentes para cada modelo. Enquanto no de regressão modelo linear uma reta é traçada para modelar a relação linear entre as variáveis, no modelo de regressão logístico é plotado uma linha para modelas as relações não lineares e categóricas. Ao verificar como os pontos de dados são distribuídos e comparar qual das linhas melhor se ajusta com essa distribuição ajuda a compreender qual modelo mais apropriado tanto descrever como entender a natureza da relação entre as variáveis.

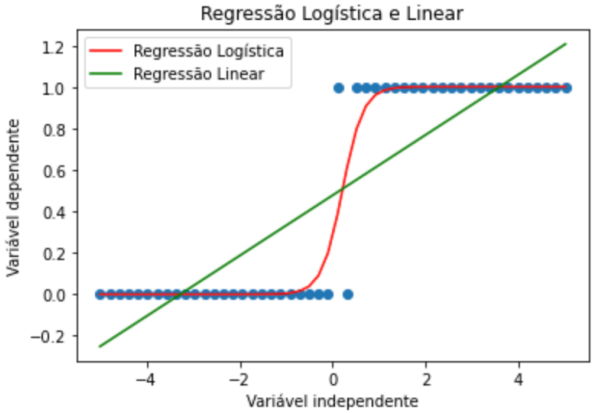

Figura 10 – Ilustração gráfica da regressão logistica e linear plano cartesiano
Fonte: Criada pelo próprio autor no Python utilizando a biblioteca matplotlib

Resultados e discussões

Finalizado o processo de tratamento e modelagem, a base de dados passou de 9 para 2716 variáveis devido a transformação das variáveis categóricas em “dummies”, ocasionando em um aumento consideravelmente a complexidade do modelo, dificultando a interpretação dos resultados e impactando na performance do modelo. Como alternativa para este problema, para avaliar a importância dos atributos e identificar dentre as 2716 variáveis, quais são as com maior importância, permitindo selecionar somente os recursos mais eficientes e consequentemente reduzindo a complexidade do modelo, fez-se uso das técnicas como, árvore de decisão, que segundo Miranda et al. (2018) o método consiste em dividir os dados em grupos cada vez menores, com base em variáveis preditoras, formando grupos homogêneos em relação à variável dependente, “Random Forest” que de acordo com Mantas et al. (2019), demonstra-se útil para problemas contendo muitas variáveis preditoras e dados não balanceados, que com base nas árvores de decisões, ajuda a identificar a importância relativa das variáveis por meio da combinação das previsões de todas as árvores através de uma votação majoritária ou média e por último a utilização da técnica “Gradient Boosting”, definido por Georganos et al. (2018) como um algoritmo de aprendizado de máquina que realiza a construção de modelos preditivos baseado por uma construção sequencial de árvores de decisão que minimizam os erros residuais do modelo anterior.

Conforme a Tabela 4, as três técnicas usando a biblioteca “sklearn”, praticamente classificaram as mesmas variáveis, porém com pesos e ordem de importância diferentes.

Tabela 4. Resumo das variáveis com maior importância para o modelo

Decision Tree		Random Forest		Gradient Boosting	
Feature	Importance	Feature	Importance	Feature	Importance
totalCost	0.346405	totalCost	0.399558	totalCost	0.498671
currency_usd	0.219512	currency_usd	0.094670	currency_usd	0.214521

idClient_CLI1	0.091381	idCreator_CRE1	0.034615	idClient_CLI57	0.045935
idCreator_CRE1	0.071337	idClient_CLI1	0.030127	idClient_CLI17	0.041434
idClient_CLI17	0.041010	idClient_CLI57	0.023804	idCreator_CRE1	0.034021
idCreator_CRE9	0.038381	idClient_CLI17	0.019293	idClient_CLI27	0.029539
idClient_CLI57	0.035391	idTarget_LT65	0.017361	idCreator_CRE9	0.028921
idClient_CLI18	0.033560	idClient_CLI27	0.016876	idCreator_CRE7	0.022538
idClient_CLI27	0.029316	idCreator_CRE7	0.014663	idClient_CLI50	0.021956
idCreator_CRE7	0.029244	idClient_CLI7	0.013493	idClient_CLI18	0.016012
idClient_CLI50	0.027501	idCreator_CRE9	0.012279	idClient_CLI7	0.013374
idClient_CLI94	0.013916	idClient_CLI50	0.009533	idClient_CLI94	0.010068
idCreator_CRE59	0.009578	idClient_CLI18	0.008132	idCreator_CRE59	0.007703
idClient_CLI7	0.003950	idClient_CLI20	0.007654	idClient_CLI1	0.005413
idClient_CLI20	0.002918	idCreator_CRE59	0.007217	idClient_CLI20	0.001809
idCreator_CRE86	0.001642	idClient_CLI44	0.006911	idClient_CLI46	0.001008
idTarget_LT9	0.000826	idClient_CLI94	0.006530	idTarget_LT1809	0.000925
idTarget_LT1809	0.000806	idTarget_LT14	0.006049	idClient_CLI134	0.000852
idCreator_CRE184	0.000396	idClient_CLI96	0.005890	idClient_CLI16	0.000826
idClient_CLI81	0.000363	idTarget_LT146	0.005820	idCreator_CRE86	0.000717

Fonte: Dados originais da pesquisa

O modelo foi executado com as variáveis selecionadas pelas três técnicas separadamente e dentre elas, o “Random Forest” demonstrou um desempenho melhor no “Log-Likelihood” e Pseudo R^2 e estas serão as variáveis a serem utilizadas no modelo de regressão logistica.

Tabela 5. Valores LL e Pseudo R^2 para as melhores variáveis selecionadas

	DecisionTree	RandomForest	GradientBoosting
Log-Likelihood	-12438	-10725	-11679
Pseudo R^2	0.4489	0.5248	0.4825

Fonte: Dados originais da pesquisa

Uma vez que a variável dependente se apresenta na forma dicotômica, não é possível estimar os parâmetros da equação de probabilidade por meio da minimização da somatória dos quadrados dos resíduos como acontece nos modelos tradicionais de regressão. Portando nesta pesquisa será utilizada a estimação por máxima verossimilhança, a qual é a mais utilizada na estimação em modelos de regressão logistica.

Fávero e Belfiore (2017), explicam que todo pesquisador deve se preocupar somente com o pressuposto da ausência de multicolinearidade das variáveis explicativas quando utilizar modelos de regressão logistica. Como na analise exploratoria dos dados foi identificado que as variáveis independente idSource e idCurrency mostraram uma correlação positiva moderada de 0.673, durante o processo de tratamento dos dados a variável idSource foi retirada da analise de modo que não impacte na qualidade preditiva do modelo.

A estimação por máxima verossimilhança, conhecido também pelo termo “Log-Likelihood” pode ser encontrada através do somatório logaritmo da função de verossimilhança:

$$LL = \sum_{i=1}^{100}\{[(\Upsilon_i).ln(p_i)] + [(1 - \Upsilon_i).ln(1 - p_i)]\} \quad (12)$$

No modelo do estudo, através da biblioteca statsmodels foi gerado um sumário o qual é possível identificar o LL_0, $LL_{max,}$ o pseudo R^2 constante, coeficientes e ver o quanto e onde o modelo pode ser melhor ajustado.

Tabela 6. Resultados da regressão logistica

Logit Regression Results						
No. Observations:						50816
Method:						MLE
Pseudo R^2:						0.5248
Log-Likelihood:						-10726
LL-Null:						-22570.
LLR p-value:						0.000
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.1145	0.040	-2.885	0.004	-0.192	-0.037
totalCost	-0.0007	1.74e-05	-41.956	0.000	-0.001	-0.001
currency_usd	42.834	0.070	60.805	0.000	4.145	4.421
idCreator_CRE1	-111.70	0.915	-12.206	0.000	-12.964	-9.377
idClient_CLI1	98.947	0.911	10.861	0.000	8.109	11.680
idClient_CLI57	81.969	0.390	21.035	0.000	7.433	8.961
idClient_CLI17	93.391	0.873	10.698	0.000	7.628	11.050
idTarget_LT65	0.1525	0.058	2.640	0.008	0.039	0.266
idClient_CLI27	-444.28	2.59e+07	-1.71e-0	1.000	-5.09e+0	5.09e+07
idCreator_CRE7	40.770	0.231	17.613	0.000	3.623	4.531
idClient_CLI7	-26.467	0.409	-6.475	0.000	-3.448	-1.846
idCreator_CRE9	-618.62	6.75e+11	-9.16e-1	1.000	-1.32e+1	1.32e+12
idClient_CLI50	62.046	0.305	20.314	0.000	5.606	6.803
idCreator_CRE59	-21.731	0.108	-20.042	0.000	-2.386	-1.961
idClient_CLI96	-0.9337	0.254	-3.669	0.000	-1.432	-0.435
idClient_CLI70	-12.011	0.275	-4.375	0.000	-1.739	-0.663
idClient_CLI47	-0.2399	0.362	-0.662	0.508	-0.950	0.470
idClient_CLI174	-17.857	0.745	-2.395	0.017	-3.247	-0.325
idClient_CLI22	-0.6636	0.226	-2.941	0.003	-1.106	-0.221

Fonte: Dados originais da pesquisa

Nota-se na Tabela 6 que o valor da somatória do logaritmo da função de verossimilhança para o modelo nulo é “LL-Null”: -22570. Já o valor máximo possível da somatória de logaritmo da função de verossimilhança é “Log-Likelihood” = -10726, a qual seria a solução ótima para este modelo.

Porém na coluna do valor p da estatística de teste para o coeficiente estimado, as variáveis idCreator_CRE9, idClient_CLI27, idClient_CLI47 e idClient_CLI22 são maiores que 0.05, indicando que estes coeficientes não são estatisticamente significativos.

Continuando o método de seleção de variáveis foi aplicado no modelo o “stepwise” da biblioteca “statsmodels”, onde Da Silva (2022) explica que este processo consiste em uma técnica estatística a qual identifica e seleciona as variáveis mais relevantes para o modelo, que expliquem a resposta ou variável dependente de forma progressiva e sistemática, proporcionando de forma objetiva uma conclusão ou resultado final. Este procedimento é realizado em duas etapas, sendo que a primeira é a de seleção das variáveis, sendo adicionadas ao modelo uma de cada vez, tendo como critério selecionar as variáveis que tem o maior F. A segunda etapa consiste na exclusão das variáveis independentes que não contribuem significativamente para explicar a variável dependente.

Tabela 7. Resultados da regressão logistica após procedimento Stepwise

Logit Regression Results						
No. Observations:						50816
Method:						MLE
Pseudo R^2:						0.4903
Log-Likelihood:						-11504.
LL-Null:						-22570.
LLR p-value:						0.000
	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.1377	0.039	-3.529	0.000	-0.214	-0.061
totalCost	-0.0007	1.57e-05	-42.142	0.000	-0.001	-0.001
currency_usd	37.162	0.059	63.454	0.000	3.601	3.831
idCreator_CRE1	-104.721	0.866	-12.088	0.000	-12.17	-8.774
idClient_CLI1	97.823	0.864	11.324	0.000	8.089	11.475
idClient_CLI57	75.754	0.365	20.759	0.000	6.860	8.291
idClient_CLI17	87.062	0.809	10.761	0.000	7.120	10.292
idTarget_LT65	0.1151	0.057	2.037	0.042	0.004	0.226
idCreator_CRE7	37.066	0.198	18.753	0.000	3.319	4.094
idClient_CLI7	-57.899	0.225	-25.683	0.000	-6.232	-5.348
idClient_CLI50	57.311	0.289	19.834	0.000	5.165	6.297
idCreator_CRE59	-17.961	0.099	-18.082	0.000	-1.991	-1.601
idClient_CLI96	-0.9843	0.253	-3.890	0.000	-1.480	-0.488
idClient_CLI70	-12.023	0.273	-4.401	0.000	-1.738	-0.667

idClient_CLI174	-18.030	0.744	-2.422	0.015	-3.262	-0.344

Atributes discarded on the process...:

'idCreator_CRE9', p-value: 0.9929349010215119
'idClient_CLI27', p-value: 0.9735860719167109
'idClient_CLI47', p-value: 0.9422870089667222
'idClient_CLI22', p-value: 0.39583218258248587

Fonte: Dados originais da pesquisa

Após aplicar o procedimento “stepwise”, houve aumento no “Log-Likelihood”, assim como nos indices "Akaike Information Criterion" [AIC] e o "Bayesian Information Criterion" [BIC], os quais Fernandes et al. (2020) aborda como medidas estatisticas de suma importancia, que auxiliam na seleção do melhor modelo dentre um conjunto de modelos candidatos, sendo que o AIC tem como base de calculo a função de verossimilhança e o número de parâmetros no modelo e o BIC realiza o calculo usando a função de verossimilhança, o número de parâmetros e o tamanho da amostra. Já o Pseudo R^2, uma medida importante, o qual Fernandes et al. (2020) define como útil para avaliar a qualidade do ajuste do modelo, nota- uma queda o que sugere que o modelo com “stepwise” tem um ajuste pior quando comparado ao modelo sem “stepwise”, sendo este último considerado o de melhor ajuste e portanto o estudo seguirá com o modelo sem “stepwise”.

Tabela 8. Comparativo entre Modelo 1 e Modelo 2

Indicadores	Modelo 1	Modelo 2
Pseudo R^2:	0.5248	0.4903
Log-Likelihood:	-10725.59272	-11812.1511
AIC:	21489.185451	23654.30223
BIC:	21657.068815	23786.84173

Fonte: Dados originais da pesquisa

Portanto a expressão de probabilidade estimada para o modelo final é:

$$p_i = \frac{1}{1 + e^{-\left(\begin{array}{c} -0.1145 + -0.0007_1 . totalCost_i + 4.2834 . currency_usd_i + (-11.1703) . idCreator_CRE1_i \\ +9.8947 . idClient_CLI1_i + 8.1969 . idClient_CLI57i + 9.3391 . idClient_CLI17i + 0.1525 . idTarget_LT65i \\ +(-44.4286) . idClient_CLI27i + 4.0770 . idCreator_CRE7i + (-2.6467) . idClient_CLI7i \\ +(-61.8629) . idCreator_CRE9i + 6.2046 . idClient_CLI50i + (-2.1731) . idCreator_CRE59i \\ +(-0.9337) . idClient_CLI96i + (1.2011) . idClient_CLI70i + (-0.2399) . idClient_CLI47i \\ +(-1.7857) . idClient\_CLI174 + (-0.6636) . idClient\_CLI22 \end{array}\right)}} \quad (13)$$

Estimado o modelo de probabilidade de ocorrência do evento, a Tabela 8 apresenta as métricas de desempenho do modelo de classificação binária elaborado, geradas a partir da biblioteca “sklearn” no “Python”. A partir dele é possível avaliar o quanto o modelo está prevendo corretamente e determinar se o modelo é útil para um determinado contexto, o qual para esta pesquisa é prever quais orçamentos não serão aprovados, ou seja, os valores 0

(não eventos). A base foi dividida em base com 70% dos dados para treinamento e 30% dos dados para teste do modelo.

Os indicadores iniciais que serão utilizados para avaliar o modelo são os de precisão, “recall”, “f1-score” e acurácia. De acordo com DeVries et al. (2021), estes índices são as mais usados na avaliação da qualidade de um modelo de classificação em aprendizado de máquina, as quais fornecem informações da capacidade em que o modelo está classificando corretamente as amostras de dados em diferentes classes, sendo que a precisão "precision" mede o quanto preciso o modelo está corretamente classificando as amostras positivas comparada ao total das amostras classificadas como positivas, o "recall" mede a capacidade do modelo em identificar corretamente as amostras positivas em relação ao total de amostras classificadas como positivas na verdadeira população, o "f1-score" é uma média harmônica entre a precisão e o "recall" e a acurácia demonstra de forma simples e direta a proporção de amostras classificadas corretamente em relação ao total de amostras. Ambos os índices permitem comparar e selecionar o melhor modelo que se ajusta aos dados e técnicas aplicadas.

Na perspectiva em prever os orçamentos não aprovados, utilizando a base de testes com 30% dos dados e com um cutoff de 0.05, a precisão para a classe 0 é de 0,76, o que significa que 76% das predições para a classe 0 foram corretas. O “recall” para a classe 0 é de 0,76, o que significa que o modelo corretamente identificou 76% das observações pertencentes à classe 0.

Verificando a métrica "f1-score", para a classe 0 obteve um valor de 0,76. As métricas macro avg e weighted avg são médias ponderadas das métricas para ambas as classes e são úteis para avaliar o desempenho geral do modelo. No caso deste modelo, a média ponderada da precisão, recall e f1-score para a classe 0 é de 0,92. A acurácia geral do modelo é de 0,92, o que indica que 92% das predições feitas pelo modelo estão corretas.

Tabela 9. Tabela de classificação para os dados de testes do modelo (cutoff = 0.5)

	precision	recall	f1-score	support
0	0.76	0.76	0.76	2067
1	0.95	0.95	0.95	10637
accuracy			0.92	12704
macro avg	0.86	0.86	0.86	
weighted avg	0.92	0.92	0.92	

Fonte: Dados originais da pesquisa

Para De Sá et al. (2020), o "cutoff" pode ser definido como um ponto de corte para separar as amostras em duas categorias distintas em modelos de classificação binária, onde a definição dele deverá ser realizada de acordo com a necessidade do negócio, porém

observando que ao diminuí-lo abaixo do padrão de 50%, pode resultar no aumento do "recall" e o aumento pode aumentar a precisão e consequentemente diminuir o "recall". Para esta pesquisa, o modelo a um “cutoff” de 0,5 apresentou a seguinte matriz de confusão:

Tabela 10. Matriz de confusão para os dados de testes do modelo (“cutoff” = 0.5)

	Valor predito		
Valor real	0	1	Total
0	1566	501	2067
1	489	10148	10637

Fonte: Dados originais da pesquisa
Nota: Não evento = 0; Evento = 1

Ao realizar um ajuste no “cutoff” para 0.7, observou que precisão e a recall para a classe 0 são ligeiramente maiores quando o cutoff é definido como 0,7. A precisão é de 0,75 e o “recall” é de 0,77, o que significa que o modelo acerta 75% das previsões para a classe 0 e que 77% das observações pertencentes à classe 0 foram corretamente identificadas pelo modelo.

Além disso, o “f1-score” para a classe 0 é de 0,76, o que é o mesmo que o obtido com o “cutoff” de 0,5. No entanto, a acurácia geral do modelo e as métricas “macro avg” e “weighted avg” são praticamente iguais em ambas as tabelas, indicando que a mudança do cutoff não teve um impacto significativo no desempenho geral do modelo.

Tabela 11. Tabela de classificação para os dados de testes do modelo (“cutoff” = 0.7)

	precision	recall	f1-score	support
0	0.75	0.77	0.76	2067
1	0.96	0.95	0.95	10637
accuracy			0.92	12704
macro avg	0.85	0.86	0.86	
weighted avg	0.92	0.92	12704	

Fonte: Dados originais da pesquisa

Já um “cutoff” ajustado igual a 0,7, resultou na seguinte matriz de confusão:

Tabela 12. Matriz de confusão para os dados de testes do modelo (“cutoff” = 0.7)

	Valor predito		
Valor real	0	1	Total
0	1592	475	2067
1	536	10101	10637

Fonte: Dados originais da pesquisa

O desempenho do modelo pode ser avaliado também pela área sob a curva “Receiver Operating Characteristic” [ROC] e o coeficiente de Gini. Ambas as medidas são usadas para avaliar o desempenho de um modelo de classificação em relação a uma classificação aleatória. Polo e Miot (2020), área sob a curva ROC avalia a capacidade do modelo de classificação em distinguir entre classes positivas e negativas, ajudando a visualizar o desempenho do modelo em diferentes níveis de "cutoff", permitindo ao pesquisador escolher o melhor valor de "cutoff" para o modelo de acordo com a necessidade de seu negócio. Já o coeficiente de Gini fornece uma medida numérica da capacidade discriminatória do modelo, o que permite comparar o desempenho de diferentes modelos e nortear o pesquisador na escolha entre os modelos, o melhor para realizar a classificação binária (Demenech et al., 2020). Ou seja, ambos são altamente recomendados em estudos científicos que envolvem modelos de classificação binária em que a classificação correta é essencial para a tomada de decisões precisas.

Neste estudo, obteve-se um valor de 0.847 para a área sob a curva ROC que varia de 0 a 1, indicando que o modelo tem um bom desempenho em distinguir as classes positivas e negativas. Por sua vez, o coeficiente de Gini, foi de 0.693040 indicado que o desempenho do modelo é bom em relação a uma classificação aleatória.

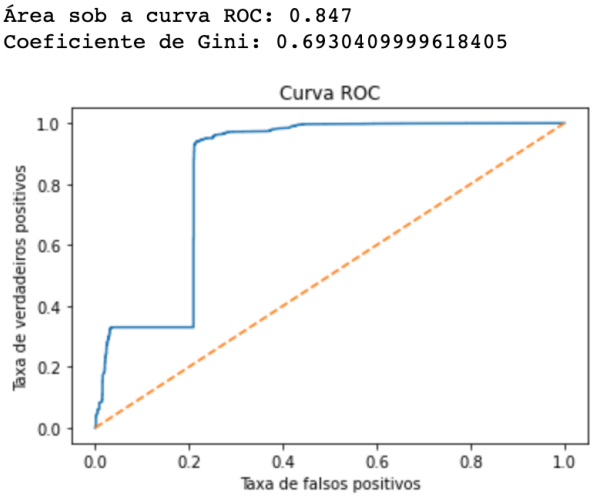

Figura 11 – Area da curva ROC e coeficiente Gini
Fonte: Dados originais da pesquisa

Apesar de para modelos de regressão logistica não ser indicado o balanceamento de classes, esta pesquisa teve como preocupação em verificar se há possíveis efeitos provenientes devido ao desbalanceamento das classes e o quando ajudaria em melhorar o modelo. Entre os algoritmos para esta finalidade, foi adotado o "NearMiss" que de acordo com Rubaidi et al. (2022), este algoritmo pode melhorar o desempenho de modelos de

classificação em problemas de desbalanceamento de classe, uma vez que ele fornece um conjunto de dados mais equilibrado, o que consequentemente evita um vies tendencioso em direção à classe majoritária. Para esta pesquisa, este algoritmo foi aplicado através da biblioteca “imblearn” com o objetivo em equilibrar as classes majoritária (APROVED) de modo que as classes fiquem distribuídas de forma mais equilibrada. A aplicação deste algoritmo, resultou no balanceamento de 2068 amostras como APPROVED (evento) e 2067 NOT_APPROVED (não evento).

Comparando o modelo inicial com o modelo com o “NearMiss” (Tabela 13), observando apenas na ótica desta pesquisa que são os não-eventos, todos os indicadores aumentaram exceto a acurácia obteve uma queda considerável, uma vez que o balanceamento das classes fez que que o modelo começasse a aprender melhor com a classe minoritária.

Tabela 13. Comparativo dos indicadores entre o modelo inicial e com “NearMiss”

	modelo inicial	modelo NearMiss
precision	0.76	0.92
recall	0.76	0.78
f1-score	0.76	0.84
accuracy	0.92	0.85
coef. ROC	0.847	0.845
indice GINI	0.69304	0.6907

Fonte: Dados originais da pesquisa

Como comportaria o modelo caso igualasse a classe minoritária com o mesmo número de amostras da classe majoritária? Para responder esta pergunta, utilizou-se o "Synthetic Minority Over-sampling Technique [SMOTE]", que seleciona uma amostra da classe minoritária e identificando suas k amostras mais próximas na classe minoritária, gera amostras sintéticas a fim de igualar com o total de amostras da classe majoritária (Zhang et al, 2022). Aplicado o algoritmo, resultou no balanceamento de 10637 amostras para cada classe, ou seja, foram adicionadas 8570 amostras sintéticas como não eventos.

Comparando com os modelos anteriores (Tabela 14), o modelo SMOTE obteve uma precisão maior, e os outros índices praticamente iguais ao modelo “NearMiss”, exceto o índice GINI o qual teve um resultado abaixo.

Tabela 14. Comparativo dos indicadores adicionando o modelo SMOTE (cutoff = 0.5)

	modelo inicial	modelo NearMiss	modelo SMOTE
precision	0.76	0.92	0.93
recall	0.76	0.78	0.78
f1-score	0.76	0.84	0.85

accuracy	0.92	0.85	0.86
coef. ROC	0.847	0.845	0.832
indice GINI	0.69304	0.6907	0.6633

Fonte: Dados originais da pesquisa

Adicionando um “cutoff” de 0.7 para os três modelos, ambos tiveram uma queda na precisão, porém um aumento no recall. Os demais indicadores permaneceram os mesmos, conforme pode ser visualizado na Tabela 15. Já os coeficientes ROC e índice GINI não mudaram, visto que eles independem do “cutoff”.

Tabela 15. Comparativo dos indicadores aumentado o cutoff (cutoff = 0.7)

	modelo inicial	modelo NearMiss	modelo SMOTE
precision	0.75	0.90	0.92
recall	0.77	0.79	0.78
f1-score	0.76	0.84	0.84
accuracy	0.92	0.85	0.86
coef. ROC	0.847	0.845	0.834
indice GINI	0.69304	0.6907	0.6670

Fonte: Dados originais da pesquisa

A validação do desempenho do modelo foi realizada também usando a base de dados referente aos orçamentos do mês de janeiro de 2023, a fim de verificar o comportamento de novos dados para os cenários anteriormente apresentados. Nesta nova base 136 orçamentos são APPROVED (eventos) e 52 NOT_APPROVED (não eventos), sendo estes o qual o objetivo são identificar.

Tabela 16. Comparativo dos indicadores aplicando os modelos nos dados janeiro 2023

“cutoff” = 0.5			
	modelo inicial	modelo NearMiss	modelo SMOTE
precision	0.83	0.92	1.00
recall	0.77	0.85	0.91
f1-score	0.80	0.88	0.95
accuracy	0.89	0.88	0.96
coef. ROC	0.847	0.870	0.838
indice GINI	0.80542	0.73964	0.6766
“cutoff” = 0.7			
	modelo inicial	modelo NearMiss	modelo SMOTE
precision	0,79	0.69	0.71
recall	0,85	0,85	1.00
f1-score	0,81	0.76	0.83

accuracy	0,89	0.73	0.79
coef. ROC	0.847	0.870	0.974
indice GINI	0.80542	0.73964	0.94723

Fonte: Dados originais da pesquisa

Conforme demonstrado na Tabela 16, ao comparar os indicadores, todos demonstram resultados similares e em alguns casos até mesmo superiores ao modelo inicial. Isso evidencia que o modelo inicial obteve um bom desempenho na identificação dos eventos e principalmente os não eventos os quais a identificação são o principal objetivo desta pesquisa.

Para fins ilustrativo e didático, a Figura 12 permite visualizar a taxa de verdadeiro positivo e taxa de falso positivo para vários pontos de corte em cada modelo.

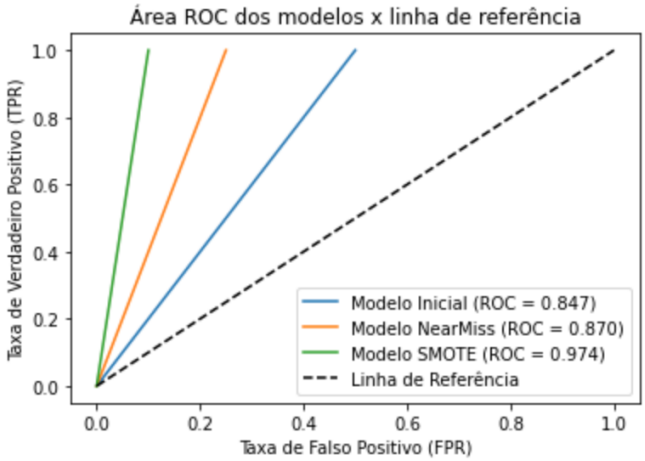

Figura 12 – Visualização das áreas ROC em relação a linha de referencia
Fonte: Dados originais da pesquisa

O modelo desta pesquisa também foi comparado a outro modelo de classificação conhecido como "K-Nearest Neighbors" [KNN], algoritmo muito utilizado para resolução de problemas em classificação, o qual se baseia em distâncias de k amostras mais próximas no conjunto de treinamento para classificar uma nova amostra (Zhang, 2022). Apesar de terem a mesma acurácia, o modelo de regressão logistica apresentou um resultado “recall” para os (não eventos) significativamente melhor.

Tabela 17. Comparativo modelo de regressão logistica e KNN

	regLog	KNN
precision	0.76	0.77
recall	0.76	0.69
f1-score	0.76	0.73
accuracy	0.92	0.92

Fonte: Dados originais da pesquisa

Considerações Finais

Os modelos de regressão logísticas são bastante usados nas áreas de créditos predizendo os riscos e se é viável realizar o empréstimo bancário para determinada pessoa, assim como na área da medicina predizendo a probabilidade de um individuo de ter ou não uma doença. Porém com esta pesquisa, ficou claro que estes modelos podem também em empresas que vendem produtos e serviços e sendo um bom framework de suporte para a área de vendas atuar para conversão de receitas.

O método stepwise, uma ótima técnica a adiciona e remove uma a uma as variáveis em uma base métrica de desempenho, seja por valor de R^2 ajustado pelo erro quadrado médio, por fim selecionando as melhores variáveis, para o modelo de pesquisa não foi tão eficaz de acordo com os indicadores “LogLik”, AIC e BIC. Provavelmente o método não considerou a interação entre as variáveis que podem não ser importantes quando consideradas individualmente, porém podem ter um efeito significativo no resultado quando interagem entre si.

Apesar de não ser indicado o balanceamento da base de dados em modelos de regressão logistica, para esta pesquisa foram testados dois algoritmos que realizam balanceamento das classes para comparativo com o modelo generalizado com as classes desbalanceadas. O primeiro deles foi o “NearMiss” que ao realizar o balanceamento igualando a classe majoritária com a classe minoritária obteve um resultado bem similar ao modelo inicial. Apesar de ser um método simples e de fácil implementação, ele pode ser mais eficaz em casos mais extremos, onde há um maior desequilíbrio entre as classes ou para diminuir o número de exemplos a serem processados e reduzindo o tempo de treinamento do modelo. O outro algoritmo utilizado foi o SMOTE, a qual ao comparar todos os indicadores não demonstrou ser tão eficiente, provavelmente pelo fato de ao criar novos exemplos sintéticos da classe minoritária com uma combinação de exemplos existentes dessa classe com os seus vizinhos mais próximos, pode ter introduzido vieses no conjunto de dados.

Para este problema de pesquisa, o qual a ideia é buscar os potenciais orçamentos que não serão aprovados, ou seja, os não eventos, de modo que o time de vendas possa agir de forma mais efetiva, o modelo inicial generalizado obteve um bom desemprenho. Na base de teste com 12704 orçamentos, direcionaria o time de vendas atuar em 2067 orçamentos, ou seja, em aproximadamente 16% dos orçamentos, ao invés de focar de forma aleatória e intuitiva, sendo que entre esses, 1592 realmente não foram aprovados. Já nos 188 orçamentos feitos no mês janeiro de 2023, o modelo indicaria para o time de vendas priorizar 46 orçamentos que representaram 24% dos orçamentos feitos no mês, sendo que dentre estes, 39 realmente não foram aprovados. Notou-se que ao aumentar o cutoff de 0.5 para 0.7,

obteve-se um aumento significativo em identificar os não eventos para os orçamentos referentes a janeiro de 2023, passando de 77% para 85%, sem afetar de forma drástica os demais indicadores e a acurácia sendo mantida em 89% em ambos. Fato que corrobora que para o problema desta pesquisa, o cutoff igual a 0.7 se adequa a realidade e necessidade do negócio.

Para estudos futuros, fica a recomendação para novas pesquisas visando outros modelos classificatórios, como por exemplo, Árvore de Decisão, KNN, entre outros, sendo que este último, apesar de não explorado nesta pesquisa, obteve resultados bem próximos ao modelo de regressão logistica, podendo ser também um bom modelo para este tipo de problema de pesquisa.

Por fim, a pesquisa mostrou-se relevante para a sociedade como um todo, uma vez que com a crescente mercadológica nesse nicho, o modelo do estudo pode propiciar grandes benefícios aos mesmos.

Agradecimentos

Primeiramente a Deus, pois tudo acontece somente com a permissão dele. Aos meus pais, João "in memoriam" e Neusa, os quais me deram o direcionamento corretos na vida proporcionando chegar até aqui. A minha esposa Ariell, por todo companheirismo, suporte e apoio, principalmente nos mais difíceis momentos. Aos meus filhos, Enzo Benjamim e Louise os quais na simplicidade e ingenuidade de seus sorrisos me motiva cada dia mais. Ao meu professor e orientador Marcos dos Santos, por toda a sua inspiradora história de vida, mostrando o caminho das pedras e tranquilizando para que eu conseguisse desenvolver este trabalho dentro do meu melhor e por fim, a instituição USP por propiciar e tornar real um sonho que tenho desde criança.

Referências

ARDITO, L., Petruzzelli, A.M., Panniello, U., Garavelli, A.C., 2019. Towards Industry 4.0: mapping digital technologies for supply chain management-marketing integration. Bus. Process Manag. J. 25 (2), 323e346. Barata, J., Rupino Da Cunha, P., Stal, J., 2018. Mobile supply. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/BPMJ-04-2017-0088/full/html. Accesso em: 19 set. 2021.

Da Silva, B.M. 2022. "MODELO PREDITIVO APLICADO AO FUTEBOL BRASILEIRO /Predictive Model Applied to Brazilian Football." Revista Brasileira De Futsal E Futebol 14.58: 291. Disponível em: https://rnp-primo.hosted.exlibrisgroup.com/permalink/f/vsvpiv/TN_cdi_gale_infotracmisc_A729756735. Acesso em: 21 fev. 2023.

De Sá, C.P.N., Jiménez, M.F., Rosa, M.W., Arlindo, E.M, Ayub, A.C.K,. Cardoso, R.B., Kreitchmann, R., Beitune, P.E. 2020. "Evaluation of Angiogenic Factors (PlGF and SFlt-1) in Pre-eclampsia Diagnosis." Revista Brasileira De Ginecologia E Obstetrícia 42.11: 697-704. Disponível em: https://www-thieme-connect-de.ez67.periodicos.capes.gov.br/products/ejournals/abstract/10.1055/s-0040-1713916. Acesso em: 11 nov. 2022.

Demenech, L.M., Dumith, S.D.C., Vieira, M.E.C.D., Silva, L.N. 2020. "Desigualdade Econômica E Risco De Infecção E Morte Por COVID-19 No Brasil." Revista Brasileira De Epidemiologia 23: Revista Brasileira De Epidemiologia, 2020, Vol.23. Disponível em: https://rnp-primo.hosted.exlibrisgroup.com/permalink/f/vsvpiv/TN_cdi_scielo_journals_S1415_790X2020000100209. Acesso em: 11 nov. 2022.

DeVries, Z., Locke, E., Hoda, M., Moravek, D., Phan, K., Stratton, A., Kingwell, S., Wai, E K., Phan, P. 2021. "Using a National Surgical Database to Predict Complications following Posterior Lumbar Surgery and Comparing the Area under the Curve and F1-score for the Assessment of Prognostic Capability." The Spine Journal 21.7: 1135-142. Disponível em: https://doi.org/10.1016/j.spinee.2021.02.007. Acesso em: 9 set. 2022.

FÁVERO, P. F., BELFIORE, P. 2017. Análise de Dados. Estatística e Modelagem Multiderivada com Excel, SPSS e Stata. 7. ed. Elsevier Editora Ltda. 2017.

Fernandes, A.V.T; FILHO, D.B.F; ROCHA, E.C; NASCIMENTO, W.S. 2020. Leia este artigo se você quiser aprender regressão logística. (2020). Rev. Sociol. Polit. 28 (74). Disponível em https://doi.org/10.1590/1678-987320287406en. Acesso em: 11 jan. 2023.

Gonçalves, B.D.O., Lerner, A.F., Souza, R.B.D.L.. 2022. "Relação Entre a Estrutura De Capital E a Política De Dividendos Nas Empresas Brasileiras Que Negociam American Depositary Receipts (ADR's)." Reunir : Revista De Administração, Ciências Contábeis E Sustentabilidade 12.1: 87-98. Disponível em: https://rnp-primo.hosted.exlibrisgroup.com/permalink/f/vsvpiv/TN_cdi_crossref_primary_10_18696_reunir_v12i1_1256 https://rnp-primo.hosted.exlibrisgroup. Acesso em: 06 dez. 2022.

Georganos, S., Grippa, T., Vanhuysse, S., Lennert, M., Shimoni, M., Wolff, E. 2018. "Very High Resolution Object-Based Land Use-Land Cover Urban Classification Using Extreme Gradient Boosting." IEEE Geoscience and Remote Sensing Letters 15.4: 607-11. Disponível em: https://rnp-primo.hosted.exlibrisgroup.com/permalink/f/vsvpiv/TN_cdi_proquest_journals_2174546710. Acesso em: 11 dez. 2022.

IBGE, Diretoria de Pesquisas, Coordenação de Estatísticas Estruturais e Temáticas em Empresas, Pesquisa Anual de Serviços 2022. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/periodicos/150/pas_2020_v22_informativo.pdf> Acesso em: 27 set. 2022.

KOTLER, P.; KELLER, K. L. Administração de marketing: a bíblia do marketing. 12. ed. São Paulo, 2007.

Mantas, C.J., Javier G. , Castellano, S.M., Abellán, J.2019. "A Comparison of Random Forest Based Algorithms: Random Credal Random Forest versus Oblique Random Forest." Soft Computing (Berlin, Germany) 23.21: 10739-0754. Disponível em: https://link-springer-com.ez67.periodicos.capes.gov.br/article/10.1007/s00500-018-3628-5. Acesso em: 08 dez. 2022.

Miranda, W.D.A., Medeiros, L.B.D., Nascimento, J.A.D., Ribeiro, K.S.Q.S., Nogueira, J.D.A., Leadebal, O.D.C.P. 2018. "Modelo Preditivo De Retenção No Cuidado Especializado Em HIV/aids." Cadernos De Saúde Pública 34.10: Cadernos De Saúde Pública, 2018, Vol.34 (10). Disponível em: https://www.scielo.br/j/csp/a/RWwSLpbLGzVzNyXVgYgndJv/?lang=pt. Acesso em: 06 dez. 2022.

Motta, D.B.A., Brito, B.B.D., Neves, L.L.V., Almeida, R.L.D., Santos, L.D., Barauna, V.G., Haraguchi, F.K. 2022. "Body Fat Estimated by Equations Based on Anthropometric Parameters Correlates with Bioelectrical Impedance in Patients Undergoing Bariatric Surgery." Revista Brasileira De Crescimento E Desenvolvimento Humano 32.3: 185-92. Disponível em: https://doi.org/10.36311/jhgd.v32.13776. Acesso em: 11 nov. 2022.

Oliveira, W.F., Lima, E.M., Gomes, D.I., Alves, K.S., Santos, P.M., Azevedo, G.S., Mezzomo, R. 2019. "Agronomic Performance of Marandu Grass Treated with Plant Growth Biostimulants in the Amazon Biome." Arquivo Brasileiro De Medicina Veterinária E Zootecnia 71.2: 603-12. Disponível em: https://rnp-primo.hosted.exlibrisgroup.com/permalink/f/vsvpiv/TN_cdi_doaj_primary_oai_doaj_org_article_dffa6c51563542eb8758b212c8796d96. Acesso em: 06 dez. 2022.

Palmeira, L.L.D.L, Cordeiro, C.P.B.S, Prado, E.C.D. 2020. "A Análise De Conteúdo E Sua Importância Como Instrumento De Interpretação Dos Dados Qualitativos Nas Pesquisas Educacionais. Disponível em: https://rnp-primo.hosted.exlibrisgroup.com/permalink/f/vsvpiv/TN_cdi_crossref_primary_10_5585_cpg_v19n1_17159. Acesso em: 07 out. 2022.

Polo, T.C.F., Miot, H.A. 2020. "Aplicações Da Curva ROC Em Estudos Clínicos E Experimentais." Jornal Vascular Brasileiro 19: Jornal Vascular Brasileiro, 2020, Vol.19. Disponível em: https://www.scielo.br/j/jvb/a/8S8Pfqnz8csmQJVqwgZT8gH/?lang=pt. Acesso em: 04 fev. 2023.

Rubaidi, Z.S., Ammar. B.B., Aouicha, M.B. 2022. "Fraud Detection Using Large-scale Imbalance Dataset." International Journal on Artificial Intelligence Tools 31.8 (): International Journal on Artificial Intelligence Tools, 2022, Vol.31 (8). Disponível em: https://www-worldscientific-com.ez67.periodicos.capes.gov.br/doi/abs/10.1142/S0218213022500373. Acesso em: 17 fev. 2023.

Silva Júnior, A.A.D., Gomes, R.D.S.R, Ventura, T.M., Rodrigues, T.R., Nogueira, J.D.S., Oliveira, A.G.D., Figueiredo, J.M.D. 2019. "Visão Geral Sobre O Tratamento De Dados

Meteorológicos No Brasil." Natural Resources (Aquidabã) 9.2: 59-66. Disponível em: https://rnp-primo.hosted.exlibrisgroup.com/permalink/f/vsvpiv/TN_cdi_crossref_primary_10_6008_CBPC2237_9290_2019_002_0006. Acesso em: 11 nov. 2022.

Zhang, A.H.Y., Huan, Z., Yang, X., Zheng, S., Gao, S. 2022. "SMOTE-RkNN: A Hybrid Re-sampling Method Based on SMOTE and Reverse K-nearest Neighbors." Information Sciences 595: 70-88. Disponível em: https://www-sciencedirect.ez67.periodicos.capes.gov.br/science/article/pii/S0020025522001736?via%3Dihub. Acesso em: 23 fev. 2023.

Zhang, S. 2022. "Challenges in KNN Classification." IEEE Transactions on Knowledge and Data Engineering 34.10: 1. Disponível em: https://rnp-primo.hosted.exlibrisgroup.com/permalink/f/vsvpiv/TN_cdi_scielo_journals_S1415_790X2020000100209. Acesso em: 23 fev. 2023.