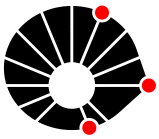

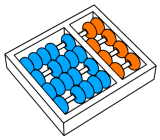

Análise de dados para modelagem da probabilidade de default para crédito de pessoa física

Guilherme Tezoli Bakaukas *Luiz Fernando Bittencourt*

Relatório Técnico - IC-PFG-23-26
Projeto Final de Graduação
2023 - Dezembro

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Análise de dados para modelagem da probabilidade de default para crédito de pessoa física

Guilherme Tezoli Bakaukas Luiz Fernando Bittencourt

Resumo

Este trabalho visa estudar o processo de análise de risco no mercado de crédito, cujo foco principal é estimar a probabilidade de default de uma pessoa física. Para tanto, o projeto utilizou uma base de dados de crédito dos Estados Unidos para treinar e avaliar o desempenho do modelo. Com base nesse estudo, entende-se que através de informações pessoais tanto básicas, quanto relacionadas ao comportamento financeiro, é possível gerar um modelo capaz de estimar, de maneira razoável, a probabilidade de uma pessoa física ser inadimplente com o pagamento de um empréstimo. Ademais, através da identificação de quais informações são consideradas importantes para o modelo, é possível avançar no entendimento do comportamento do inadimplente de maneira geral, além de otimizar a coleta de informações utilizadas para esse propósito.

1 Introdução

Com a flexibilização da política de crédito por parte do Banco Central e a evolução das Sociedades de Crédito Direto (SCDs), o crédito passou a ser um recurso cada vez mais acessível, tanto para as pessoas físicas, quanto para os credores. Logo, com novas empresas ingressando no mercado de crédito, a necessidade de uma análise precisa de risco vem assumindo uma maior relevância.

No entanto, para que a instituição credora determine o nível de risco daquela operação e, consequentemente, a taxa de juros, é importante entender sobre a probabilidade de default (inadimplência) daquele indivíduo.

Assim, este projeto tem como objetivo criar um modelo capaz de calcular essa métrica, através de informações do candidato ao crédito, utilizando recursos estatísticos, análise de dados e aprendizado de máquina.

Atualmente, a definição desse parâmetro envolve diversas técnicas, como, por exemplo, a modelagem de um score, baseado em informações pessoais e histórico de crédito, assim como a elaboração de uma estimativa, através da similaridade com casos de inadimplência anteriores. Frequentemente, as instituições credoras utilizam sistemas chamados Bureaus de dados, que caracterizam agências de crédito que realizam a coleta de informações através de instituições autorizadas a armazenar dados relacionados ao histórico financeiro de pessoas físicas. Neles, podem ser consultadas diversas informações básicas de pessoas físicas, como data de nascimento e endereço, por exemplo, além de informações sobre o comportamento financeiro, como histórico de pagamentos, limites de crédito, entre outros dados. Assim,

torna-se possível realizar uma análise ainda mais robusta sobre o candidato ao crédito, fornecendo mais insumos para a modelagem da probabilidade de default.

Dessa forma, a fim de acompanhar essa tendência, este trabalho utiliza uma parcela de dados abrangente, incluindo informações pessoais, bem como aquelas relacionadas ao comportamento financeiro do candidato ao crédito.

Nesse sentido, ressalta-se que um movimento muito comum dentre as instituições credoras, é a utilização de uma variável chamada Credit Score, gerada através de informações do histórico de crédito de um indivíduo e disponibilizada no mercado por empresas especializadas na área.

Logo, muitas probabilidades de default são fortemente vinculadas a esse valor. No entanto, optou-se pela não utilização desta informação já processada, justamente para estimar a probabilidade de default diretamente das informações disponíveis em mais baixo nível.

Assim, este projeto será fundamentado em um modelo de aprendizagem, utilizando uma base de dados históricos, com casos de inadimplência e pagamento completo do empréstimo. De modo geral, essa operação costuma ser realizada através de modelos de aprendizagem que retornam uma probabilidade como resultado, como regressão logística e redes neurais por exemplo. Por fim, para o contexto do projeto, será utilizada uma ferramenta de gradient boosting chamada XGBoost com a finalidade de cumprir com o objetivo descrito.

2 Conceitos básicos

A fim de introduzir o a área referente ao risco de crédito de uma pessoa física, é necessário apresentar alguns conceitos importantes. Primeiramente, deve-se analisar onde a probabilidade de default se encaixa nesse contexto.

Basicamente, o processo de avaliação do risco de um dado empréstimo baseia-se nos seguintes conceitos:

- EL = Expected Loss
- PD = Probability of Default
- LGD = Loss Given Default
- EAD = Exposure at Default

Que se relacionam da seguinte maneira:

$$EL = PD * LGD * EAD \tag{1}$$

Assim, percebe-se que a perda esperada (EL) é definida pela probabilidade do devedor ser inadimplente (PD), multiplicada pelo valor do empréstimo (EAD), multiplicado pelo valor que define a proporção do EAD que o credor não poderá recuperar em caso de default (LGD).

Tal relação é amplamente reconhecida e recomendada por padrões regulatórios internacionais para instituições bancárias, como os chamados Basel Accords [1], que definem uma série de regulamentações para supervisionar e padronizar os serviços bancários de diversos países.

Desse modo, a instituição credora será capaz de balancear a relação entre os parâmetros de valor do empréstimo e da probabilidade de inadimplência, com o objetivo de encontrar um equilíbrio saudável para estimar o Expected Loss.

Além disso, a probabilidade de default também é importante para a definição dos juros daquele empréstimo, chamado de 'interest rate', uma vez que, é importante que a instituição estime uma taxa de juros compatível com o risco daquela operação de crédito, ou seja, para compensar uma probabilidade de default elevada, entende-se que o ganho deva ser intensificado.

Contudo, por mais que esses parâmetros estejam interligados, a definição do 'interest rate' inclui também uma combinação de diversos outros fatores, como, por exemplo, o cenário do mercado e outras informações do devedor.

Portanto, por estar intimamente relacionada às projeções de risco e definições de valores como juros e quantia dos empréstimos, a criação de uma boa estimativa da probabilidade de default é de extrema importância para a instituição.

3 Metodologia

Um sistema de modelagem de previsão de default segue, de maneira geral, os passos definidos pelo fluxograma da Figura 1, que demonstra um pipeline de treinamento de um modelo de aprendizado desde a coleta das informações até a avaliação dos resultados.

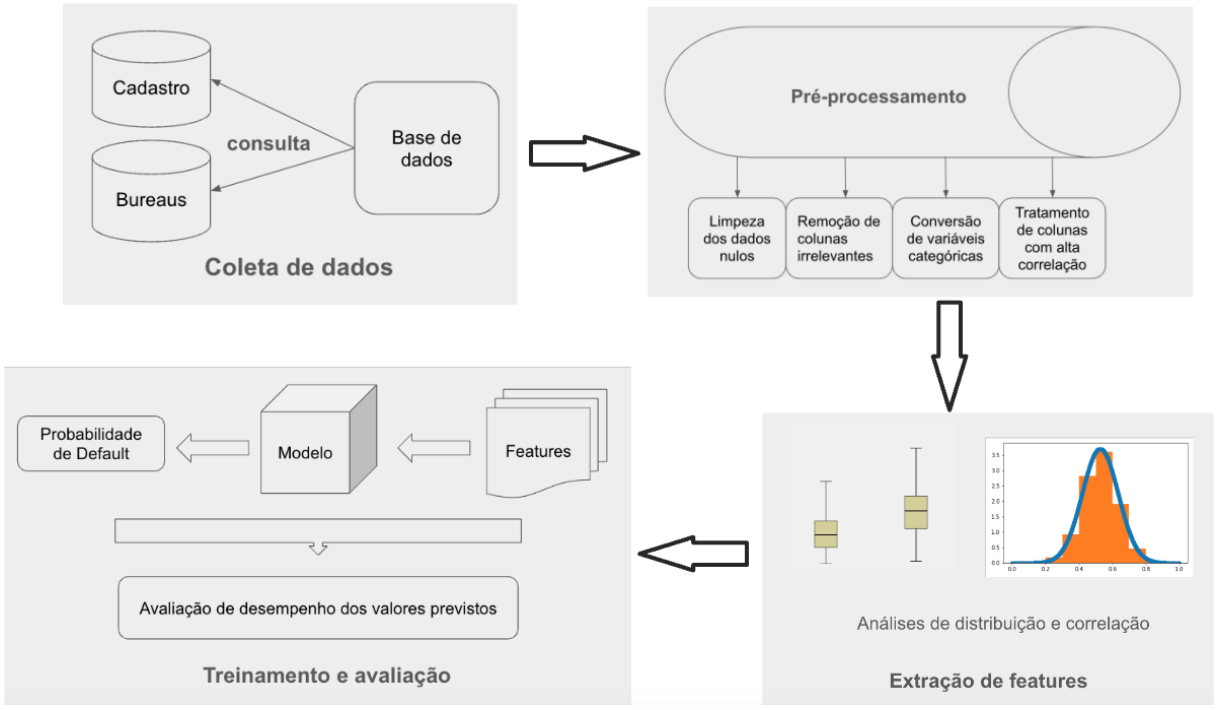

Figura 1: Fluxograma da metodologia

No entanto, para o contexto do projeto, abstraiu-se a etapa de coleta de dados, uma vez

que utilizou-se uma base de dados pronta, com as informações de bureaus já incluídas.

Assim, fundamentado na metodologia descrita, o projeto passou a ser segregado nas seguintes etapas:

(A): Seleção de uma base de dados no contexto de crédito;
(B): Limpeza e tratamento dos dados da base;
(C): Extração das features relevantes para o treinamento;
(D): Treinamento e avaliação de um modelo de aprendizado de máquina.

3.1 Base de dados

A base de dados escolhida para a análise de risco é de autoria de uma empresa chamada Lending Club [2], que fornece o serviço de empréstimo peer-to-peer, situada nos Estados Unidos. Os dados foram coletados no período de 2007 a 2015 e contemplam informações sobre pagamento, status do empréstimo, histórico no cenário de crédito e informações pessoais como endereço e renda, por exemplo.

Essa base está estruturada em 2260668 linhas e 145 colunas e acompanha uma descrição das colunas bem documentada, permitindo uma análise mais detalhada da aplicação de cada informação no contexto estudado.

3.2 Tratamento dos dados

Para tornar a base viável para treinamento, fez-se necessário entender a distribuição e o significado de cada coluna, para torná-las compatíveis com a necessidade do modelo. Desse modo, 4 etapas foram estabelecidas para esse processo:

1. Limpeza dos dados nulos;
2. Remoção de colunas irrelevantes para o problema;
3. Conversão de variáveis categóricas para numéricas;
4. Tratamento de colunas com alta correlação.

Entretanto, antes de iniciar os passos descritos, é essencial analisar a coluna com a informação alvo, 'loan_status'.

Sua distribuição, descrita na Figura 2, indica uma diversidade de valores para definir o status de empréstimo de um indivíduo, logo, será necessário descartar os valores irrelevantes para o nosso contexto.

Para o projeto, colocou-se em foco apenas os dados que indicam a conclusão de pagamento ou inadimplência. Desta forma, para o primeiro cenário, coletou-se apenas o status 'Fully Paid', posto que indica, de fato, o pagamento completo do empréstimo.

Com relação ao cenário de inadimplência, foram reconhecidos os status de 'Default', 'Charged Off' e 'Late (31-120 days)'. Entende-se que seja razoável adicionar os casos muito atrasados, pois caracteriza um forte indicativo de inadimplência.

Portanto, foram excluídos os demais dados da base para seguir com os próximos passos descritos.

loan_status	
Fully Paid	1041952
Current	919695
Charged Off	261655
Late (31-120 days)	21897
In Grace Period	8952
Late (16-30 days)	3737
Does not meet the credit policy. Status:Fully Paid	1988
Does not meet the credit policy. Status:Charged Off	761
Default	31

Figura 2: Distribuição dos valores de loan_status

3.2.1 Limpeza dos dados nulos

Inicialmente, buscou-se entender a distribuição desses casos na base e, para isso, foi calculada a porcentagem de dados nulos para cada coluna e definido um threshold de tolerância de 70%. De modo que, as colunas com valores acima dessa porcentagem foram removidos, como indicado na Figura 3, reduzindo o número de colunas de 145 para 104.

	Column Name	NaN Percentage
0	id	100.000000
1	member_id	100.000000
2	url	100.000000
3	desc	94.423551
4	mths_since_last_record	84.112837
5	mths_since_last_major_derog	74.309585
6	annual_inc_joint	94.660428
7	dti_joint	94.660605
8	verification_status_joint	94.880717
9	mths_since_recent_bc_dlq	77.011175
10	revol_bal_joint	95.221766
11	sec_app_earliest_cr_line	95.221722
12	sec_app_inq_last_6mths	95.221722
13	sec_app_mort_acc	95.221722
14	sec_app_open_acc	95.221722
15	sec_app_revol_util	95.302981
16	sec_app_open_act_il	95.221722
17	sec_app_num_rev_accts	95.221722
18	sec_app_chargeoff_within_12_mths	95.221722
19	sec_app_collections_12_mths_ex_med	95.221722

(a) Primeiras 20 colunas

	Column Name	NaN Percentage
20	sec_app_mths_since_last_major_derog	98.410116
21	hardship_type	99.530537
22	hardship_reason	99.530537
23	hardship_status	99.530537
24	deferral_term	99.530537
25	hardship_amount	99.530537
26	hardship_start_date	99.530537
27	hardship_end_date	99.530537
28	payment_plan_start_date	99.530537
29	hardship_length	99.530537
30	hardship_dpd	99.530537
31	hardship_loan_status	99.530537
32	orig_projected_additional_accrued_interest	99.627278
33	hardship_payoff_balance_amount	99.530537
34	hardship_last_payment_amount	99.530537
35	debt_settlement_flag_date	98.537777
36	settlement_status	98.537777
37	settlement_date	98.537777
38	settlement_amount	98.537777
39	settlement_percentage	98.537777
40	settlement_term	98.537777

(b) Restante das colunas

Figura 3: Colunas removidas pela análise de dados nulos

3.2.2 Remoção de colunas irrelevantes para o problema

Nessa etapa, foi necessário estudar o significado de cada coluna [2] para mensurar seu impacto na classe a ser prevista. Para isso, foi avaliado se a informação estaria atrelada ao comportamento posterior à concessão do crédito, isto é, em essência, informações sobre o

pagamento das parcelas do empréstimo.

Nesse sentido, para ilustrar esse procedimento, na Figura 4 estão elencadas as colunas que foram removidas do conjunto, com as suas respectivas descrições.

Vale ressaltar que as informações que envolvem o valor do empréstimo como 'loan_amnt' foram descartadas pois o objetivo é calcular a probabilidade de default de maneira isolada, isto é, como descrito anteriormente, a probabilidade de default pode ser utilizada para mensurar o valor a ser cedido no empréstimo, logo, para impedir que essa correlação seja considerada na modelagem, optou-se por remover essa informação.

O mesmo vale para os dados de 'grade' e 'sub_grade', pois caracterizam um score de crédito indesejado, que, portanto, devem ser desvinculados da análise.

Por fim, o mesmo foi realizado com a coluna 'int_rate', que indica a informação dos juros concedidos para o empréstimo, e, como essa informação tende a estar associada ao parâmetro objetivo, ela foi removida da base.

Name	Description
collection_recovery_fee	post charge off collection fee
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
last_credit_pull_d	The most recent month LC pulled credit for this loan
last_pymnt_amnt	Last total payment amount received
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
out_prncp	Remaining outstanding principal for total amount funded
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors
recoveries	post charge off gross recovery
sub_grade	LC assigned loan subgrade
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
total_rec_int	Interest received to date
total_rec_late_fee	Late fees received to date
total_rec_prncp	Principal received to date
debt_settlement_flag	Flags whether or not the borrower, who has charged-off, is working with a debt-settlement company.

Figura 4: Colunas irrelevantes para o contexto do projeto

3.2.3 Conversão de variáveis categóricas para numéricas

O processo de conversão pode ser realizado de maneiras diferentes. No contexto do projeto, utilizou-se algumas técnicas como label encoding, one hot encoding e conversão para uma variável de risco para tratar esses casos.

Na base, as colunas que apresentaram um tipo categórico estão elencadas e atreladas às suas respectivas quantidades de categorias na Figura 5.

Uma vez identificadas, iniciou-se o processo de conversão, em que foram adotadas as técnicas de label encoding para colunas com número de categorias reduzidas com uma relação de valor, one hot encoding para aquelas sem uma relação clara de valor e com um número reduzido de categorias e, para as demais, a conversão para variável de risco.

	Coluna	Qtd de categorias
0	term	2
1	emp_title	512694
2	emp_length	11
3	home_ownership	6
4	verification_status	3
5	issue_d	139
6	loan_status	9
7	pymnt_plan	2
8	purpose	14
9	title	63154
10	zip_code	956
11	addr_state	51
12	earliest_cr_line	754
13	initial_list_status	2
14	application_type	2
15	hardship_flag	2
16	disbursement_method	2

Figura 5: Colunas categóricas

As seguintes colunas foram convertidas com label encoding:

1. loan_status:
   0: 'Fully Paid'
   1: 'Default', 'Charged Off', 'Late (31-120 days)'

2. emp_length

   A conversão seguiu o padrão do intervalo 0 - 10 para os valores '< 1 year' - '10+ years'

3. verification_status

   0: 'Not Verified'
   1: 'Source Verified'
   2: 'Verified'

4. initial_list_status

   0: 'w'
   1: 'f'

Já para a técnica de one-hot encoding, apenas a coluna home_ownership foi processada, pois ela possui apenas 5 classes que não compartilham uma relação clara de linearidade. De modo que as colunas 'MORTGAGE', 'RENT', 'OWN', 'ANY' e 'OTHER' foram adicionadas ao conjunto de dados.

Por fim, a técnica de conversão por um variável de risco foi utilizadas nas seguintes colunas:

1. emp_title
2. purpose
3. title
4. zip_code
5. addr_state

Esse procedimento implica em calcular um valor que represente o risco baseado na incidência daquela categoria no banco de dados e pode ser descrito da seguinte maneira:

$$risk\_variable = (Defaut_x/Total_x)/(Default/Total) \tag{2}$$

Seja:

Default_x = Quantidade de dados inadimplentes da categoria x
Total_x = Quantidade de dados da categoria x
Default = Quantidade total de dados inadimplentes
Total = Quantidade total de dados

Assim, para cada categoria da coluna esse valor foi calculado. Porém, não seria adequado utilizar esse valor para casos com incidência muito reduzida, pois indicaria um cenário muito específico e poderia gerar resultados inadequados. Logo, as categorias com incidência menor que um certo threshold foram agrupadas para gerar uma nova distribuição de categorias, e, através dela, novos valores de risco foram calculados.

Entretanto, algumas das colunas categóricas restantes não foram convertidas para valores numéricos por indicarem uma distribuição reduzida, como mostrado no exemplo da Figura 6. Nesta imagem, é notável a existência de uma assimetria na distribuição dos valores da coluna, pois a proporção de valores 'Joint App' é relevantemente menor que os valores 'Individual'. Logo, devido à sua baixa diversidade, ela não será capaz de entregar valor relevante ao modelo. Desse modo, essa coluna, em conjunto com as colunas enumeradas a seguir, foram descartadas por apresentarem um comportamento semelhante.

1. pymnt_plan
2. application_type
3. hardship_flag
4. disbursement_method

Por fim, para as colunas categóricas restantes, as seguintes medidas foram tomadas:

1. issue_d e earliest_cr_line:
   Foram transformadas em um valor, indicando o intervalo de tempo entre a primeira linha de crédito cadastrada (earliest_cr_line) e a data do pedido de empréstimo (issue_d), convertido para escala logarítmica, com o objetivo de incluir a vida ativa do indivíduo no mercado de crédito.

   Ademais, a utilização do logaritmo visou dar mais relevância e visibilidade para as mudanças ocorridas nos intervalos de tempo mais baixos. Assim, um intervalo de 1

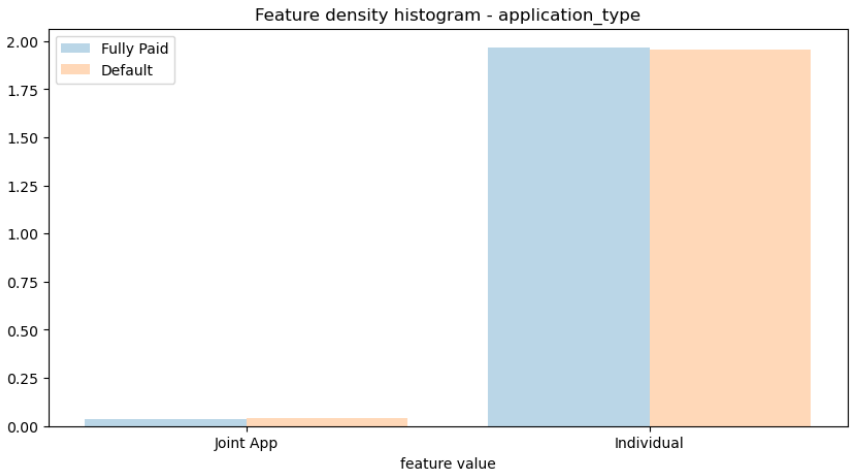

Figura 6: Exemplo de distribuição reduzida: 'application_type'

ano para 2 anos em escala logarítmica será melhor representado com relação a esse mesmo intervalo entre 50 para 51 anos.

$$log(2) - log(1) = 0.301 \tag{3}$$

$$log(51) - log(50) = 0.0086 \tag{4}$$

3.2.4 Tratamento de colunas com alta correlação

Com as colunas todas convertidas em valores numéricos, torna-se possível analisar métricas como a correlação na base completa. Este passo é importante para eliminar features que representem o mesmo fenômeno, a fim de evitar a inserção de informações replicadas no modelo.

Para tal propósito, considerou-se a correlação entre as colunas. Assim, bastou calcular esse valor para cada possível par de colunas e filtrá-los baseado em um threshold, definido em 0.9. Portanto, uma vez definidos os pares de colunas intensamente correlacionadas, manteve-se apenas a coluna com maior correlação com a classe 'loan_status', para preservar a informação que mais gera valor para a predição.

Com esse processo, as seguintes colunas foram eliminadas:

1. num_sats
2. total_il_high_credit_limit
3. total_bal_il
4. num_rev_tl_bal_gt_0
5. tot_cur_bal

3.3 Extração de features

Esta etapa consiste em selecionar as features mais correlacionadas com a coluna 'loan_status' a ser prevista. Sendo assim, o projeto se baseou no seguinte procedimento para realizar essa

análise:

Primeiramente, a fim de entender a distribuição dos dados daquela coluna, foi gerado um histograma de densidade segregado pela classe alvo, com o intuito de analisar tanto a distribuição dos dados, quanto a correlação. Além disso, criou-se um boxplot dos valores de cada coluna, também segregados pela classe, para visualizar a diferença entre as distribuições de maneira mais clara, e detectar possíveis anomalias.

As Figuras 7 e 8 ilustram o processo descrito para a coluna 'annual_inc'.

Nesse caso, é evidente que ela caracteriza uma boa feature, pois é capaz de segregar, de maneira razoável, os dados categorizados como 'Fully Paid' dos dados 'Default', uma vez que, existe uma tendência dos casos com valores reduzidos de 'annual_inc' agruparem mais casos de inadimplência, proporcionalmente. Portanto, essa feature é mantida no conjunto.

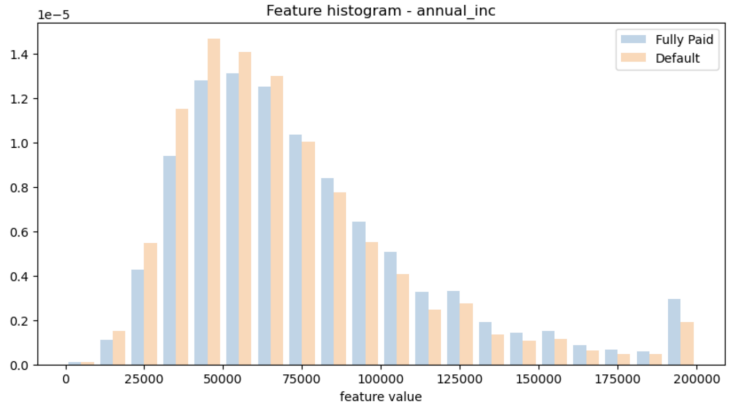

Figura 7: Histograma de densidade: 'annual_inc'

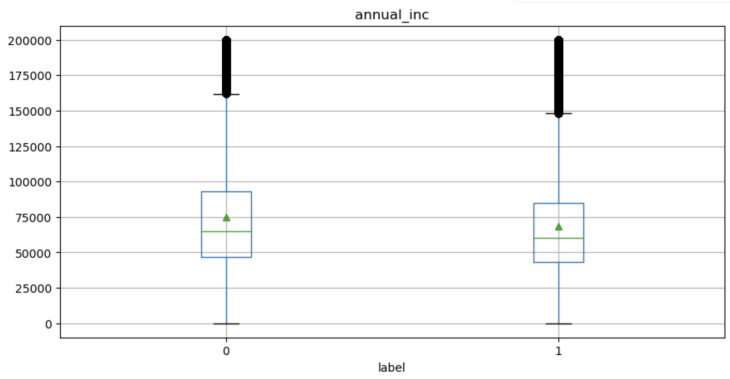

Figura 8: Boxplot da distribuição dos dados: 'annual_inc'

Em contrapartida, encontrou-se cenários como o demonstrado nas Figuras 9 e 10, em que a coluna não foi capaz de diferenciar o comportamento de inadimplência de maneira mais efetiva, ou seja, apresenta pouca correlação com o 'loan_status'.

Isso se deve ao fato da distribuição dos valores da coluna não apresentar nenhuma dis-

crepância entre as classes, visto que o histograma demonstra valores parecidos de 'Default' e 'Fully Paid' para os intervalos indicados. Em adição, os boxplots gerados indicaram uma distribuição muito similar entre os dois conjuntos. Portanto, colunas com esse mesmo comportamento foram removidas da base.

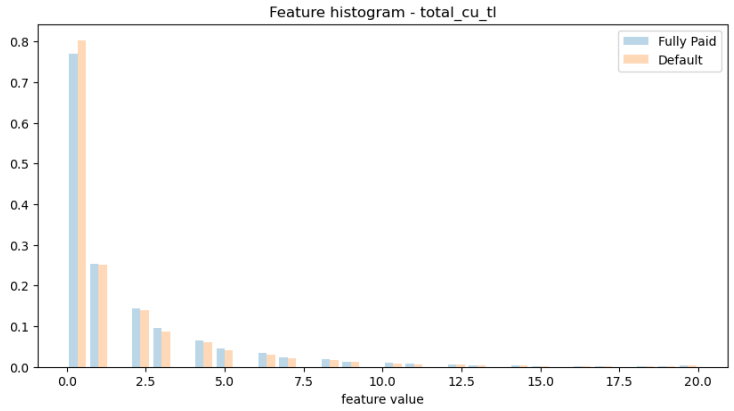

Figura 9: Histograma de densidade: 'total_cu_tl'

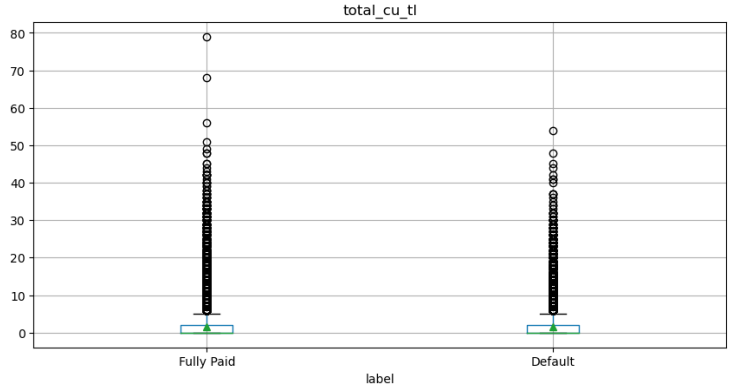

Figura 10: Boxplot da distribuição dos dados: 'total_cu_tl'

A partir desse processo, foram selecionadas apenas 55 das 77 colunas que restavam na base de dados.

Em vista disso, com as features já selecionadas e devidamente tratadas, foi possível seguir com o processo de treinamento do modelo de aprendizado de máquina.

3.4 Treinamento e avaliação de um modelo de aprendizado de máquina

Para a seleção do modelo, foi essencial entender sobre as melhores técnicas de classificação no cenário do mercado financeiro [3]. Como conclusão desse estudo, foi identificado que árvores de decisão, SVMs (Support Vector Machines) e redes neurais poderiam performar de

maneira adequada para o contexto do projeto, uma vez que, retornariam a probabilidade do indivíduo ser classificado como inadimplente, indicando a desejada probabilidade de default.

Portanto, o modelo selecionado para o treinamento foi o XGBoost ("Extreme Gradient Boosting") [4], que consiste em um modelo de classificação que utiliza árvores de decisão combinadas com o processo de Gradient Boosting. Tal processo atua na minimização dos resíduos da última predição com o uso do algoritmo de gradiente descendente [5], acelerando o processo de aprendizagem e otimização dos valores a serem previstos.

A escolha desse modelo se deu devido à sua eficiência na velocidade do treinamento, assim como sua performance em classificação [6]. Quando comparado com outros modelos, como redes neurais e random forest, por exemplo, ele geralmente assume um desempenho superior [7], além de adotar técnicas para minimizar o overfit e alavancar a velocidade de execução.

3.4.1 Treinamento

O modelo selecionado permite definir uma série de hiperparâmetros para que ele execute o treinamento. Assim, visando alcançar o melhor conjunto de parâmetros possível, aplicou-se uma ferramenta chamada GridSearchCV, que atua na otimização dessas variáveis, com o objetivo de identificar a opção que mais favorece uma determinada métrica de avaliação [8].

Nesse cenário, os hiperparâmetros inseridos na otimização foram os seguintes:

1. eta: [0.05, 0.10, 0.30]
2. min_child_weight: [7, 9]
3. gamma: [0.5, 0.7]
4. subsample: [0.6, 0.8]
5. colsample_bytree: [0.8, 1.0]
6. max_depth: [3,4,5]
7. scale_pos_weight: [4]
8. objective: ["binary:logistic"]

Vale ressaltar que a definição do valor de scale_pos_weight esteve atrelado ao desbalanceamento entre os dados 'Default' e 'Fully Paid' descrito na Figura 11, de modo a compensar a proporção existente na base, de 1041952 pagadores para 261686 inadimplentes.

Assim como a métrica de avaliação utilizada foi a F1_score, pois tem boa performance em bases desbalanceadas, como no contexto do projeto, além de garantir uma boa visão sobre o desempenho do modelo tanto para precisão, quanto para o recall.

Desse modo, o processo de otimização dos hiperparâmetros foi finalizado e resultou no seguinte conjunto ótimo, que foi submetido para avaliação:

1. eta: 0.3
2. min_child_weight: 7
3. gamma: 0.5
4. subsample: 0.8
5. colsample_bytree: 0.8
6. max_depth: 5
7. scale_pos_weight: 4
8. objective: "binary:logistic"

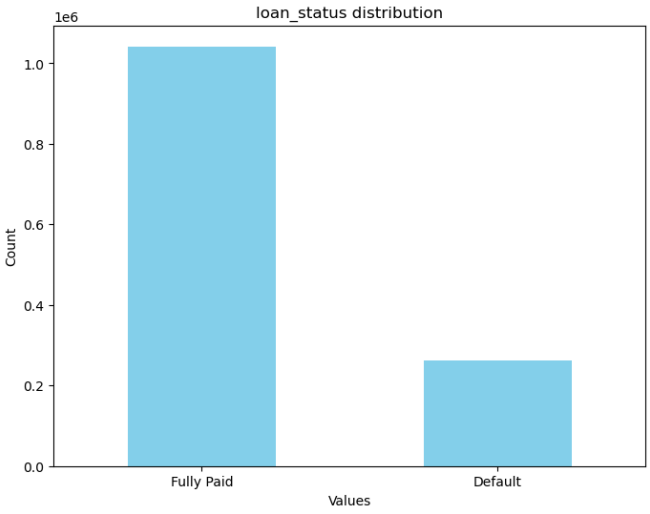

Figura 11: Balanceamento dos valores de loan_status

3.4.2 Avaliação

Para o processo de avaliação, atentou-se aos resultados de precision, recall, F1_score e a matriz de confusão para os conjuntos de treinamento e teste. Assim, foi possível entender o desempenho geral do modelo e identificar divergências entre os resultados de treino e teste, caracterizando um overfit, por exemplo.

Nesse contexto, os resultados obtidos das Figuras 12 e 13 evidenciaram pouca diferença entre as métricas de treino e teste.

Mas, ao analisar os resultados de teste da Figura 12, notou-se um cenário com uma boa precisão para a label 0 (Fully Paid) de 89.07%, acompanhada de uma precisão mais reduzida para a predição da label 1 (Default), de 33.33%. Essa discrepância decorre do desbalanceamento entre as classes.

Contudo, o recall mensurado se encontra balanceado, pois apresentou valores de 65.95% e 67.78% para as labels 0 e 1, respectivamente. Logo, a proporção de acerto na classificação dos casos 0 e 1 se mostrou parecida e razoável, com por volta de 2/3 de acerto para cada classe.

Para complementar a análise, é interessante observar a relação entre as taxas de falso positivo e verdadeiro positivo apontadas pelo modelo e explicitado na Figura 14. Através dela, é possível visualizar o 'trade-off' entre os acertos na classificação de um inadimplente e o impacto no erro da classificação dos casos pagadores.

Nesse sentido, para o propósito do projeto, é válido manter a análise nas taxas determinadas por padrão no classification report da Figura 12, de 67.78% de true positive rate e 34.05% de false positive rate (100% - recall label 0).

Ademais, outro resultado interessante é demonstrado na Figura 15, que ilustra a distribuição, em densidade, da proporção dos casos 'Default' e 'Fully Paid' para os valores de probabilidade previstos. Desse modo, é possível notar que o modelo manifestou capacidade de gerar resultados de probabilidade condizentes com os casos de teste, uma vez que, valo-

```
Train Result:
================================================
------------------------------------------------
CLASSIFICATION REPORT:
                    0.0           1.0  accuracy     macro avg  weighted avg
precision      0.896751      0.342380  0.670751  6.195654e-01  7.854689e-01
recall         0.664578      0.695332  0.670751  6.799547e-01  6.707511e-01
f1-score       0.763402      0.458832  0.670751  6.111170e-01  7.022639e-01
support    833561.000000  209349.000000  0.670751  1.042910e+06  1.042910e+06
------------------------------------------------
Confusion Matrix:
 [[553966 279595]
 [ 63782 145567]]

Test Result:
================================================
------------------------------------------------
CLASSIFICATION REPORT:
                    0.0           1.0  accuracy      macro avg   weighted avg
precision      0.890726      0.333333  0.663201       0.612030       0.778838
recall         0.659525      0.677838  0.663201       0.668681       0.663201
f1-score       0.757885      0.446900  0.663201       0.602392       0.695459
support    208391.000000  52337.000000  0.663201  260728.000000  260728.000000
------------------------------------------------
Confusion Matrix:
 [[137439  70952]
 [ 16861  35476]]
```

Figura 12: Resultados de treino e teste do XGBoost

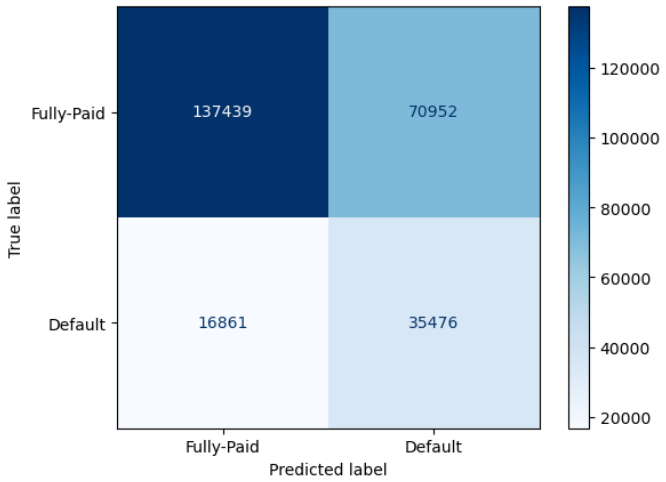

Figura 13: Matriz de confusão da base de teste

res de score maiores 0.5 assumiram, de fato, maior probabilidade de serem inadimplentes, devido a distribuição majoritária desses casos. O contrário também foi validado.

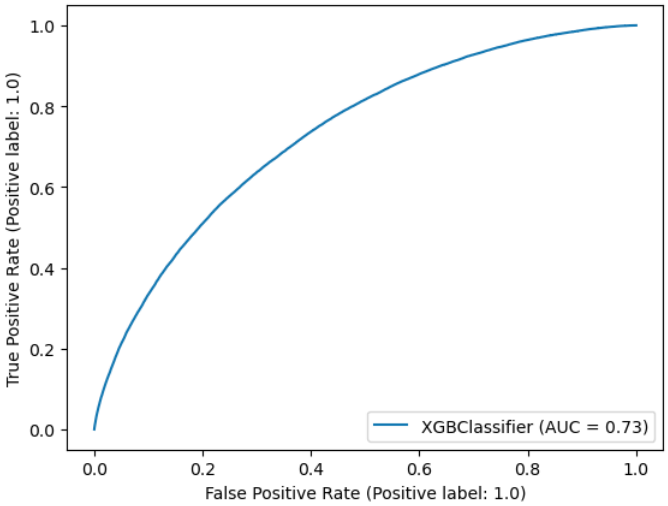

Figura 14: ROC Curve

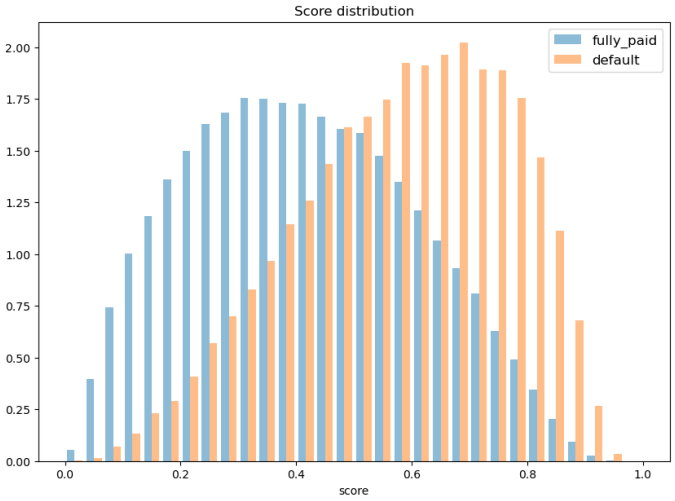

Figura 15: Score distribution

Por fim, o XGBoost nos permite analisar a relação de peso que ele aplica para cada uma das features de entrada.

O gráfico apresentado na Figura 16 confirma a correlação identificada no momento de extração das features mais relevantes. Um exemplo disso está na relevância dada para a coluna 'annual_inc', que foi previamente analisada e classificada como uma boa feature.

Dentre as elencadas na figura, as consideradas mais relevantes pelo modelo foram selecionadas e descritas na Tabela 1.

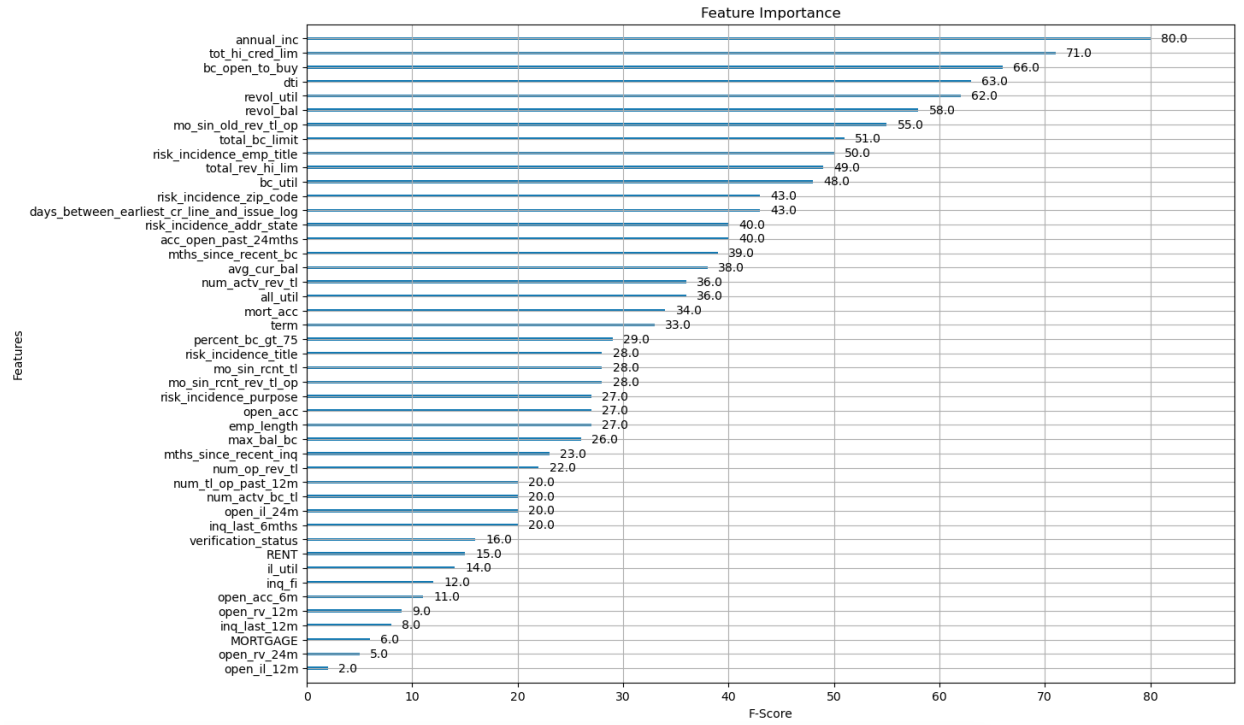

Figura 16: XGBoost importância por feature

Column	**Description**
annual_inc	The selfreported annual income provided by the borrower during registration
tot_hi_cred_lim	Total high credit/credit limit
bc_open_to_buy	Total open to buy on revolving bankcards
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit
revol_bal	Total credit revolving balance

Tabela 1: Descrição das features mais relevantes

Através dela, é possível evidenciar que as informações com maior capacidade de segregar o comportamento inadimplente do bom pagador, envolvem, em especial, dados sobre renda, histórico no mercado de crédito e comportamento financeiro.

4 Conclusão

Com a observação dos resultados, percebe-se que, embora sua performance não demonstre capacidade de segregar completamente os casos inadimplentes dos pagadores, o modelo foi capaz de classificá-los de maneira razoável. De modo que, a distribuição do score evidenciou o fato de valores mais elevados indicaram mais casos confirmados de Default e valores reduzidos validaram o oposto.

Além disso, o modelo foi capaz de identificar quais variáveis assumiram maior relevância na predição dos resultados e, com essa informação, foi possível estender o conhecimento sobre o comportamento do inadimplente, além de pontuar quais informações e áreas são importantes para a análise. Nesse contexto, é possível aprimorar o processo de coleta de informações, tanto para minimizar a consulta de dados menos relevantes para o problema, quanto para direcionar os esforços de seleção de bureaus que possuem as informações mais importantes para a modelagem. Dessa maneira, a instituição pode reduzir seus custos com a coleta de informações e tornar esse processo ainda mais assertivo e direcionado.

Portanto, é razoável afirmar que seu valor de saída poderia ser utilizado como probabilidade de default de uma pessoa física e, consequentemente, associado ao cálculo do 'Expected Loss' (EL) de uma instituição credora. Com esse valor estimado, é possível mensurar, de maneira mais segura, os valores dos empréstimos (EAD), além de associá-lo às taxas de juros, 'interest rate', dessas operações.

Referências

[1] Robert L. Burns. *Economic Capital and the Assessment of Capital Adequacy*. URL: https://www.fdic.gov/regulations/examinations/supervisory/insights/siwin04/siwinter2004-article01.html. (accessed: 13.12.2023).

[2] Lending Club. *Lending Club Loan Data*. URL: https://www.kaggle.com/datasets/adarshsng/lending-club-loan-data-csv?select=loan.csv. (accessed: 20.11.2023).

[3] Akib Mashrur. *Machine Learning for Financial Risk Management: A Survey*. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9249416. (accessed: 22.11.2023).

[4] xgboost developers. *Introduction to Boosted Trees*. URL: https://xgboost.readthedocs.io/en/stable/tutorials/model.html. (accessed: 23.11.2023).

[5] L. Yan X. Wang e Q. Zhang. *Research on the Application of Gradient Descent Algorithm in Machine Learning*. URL: https://ieeexplore.ieee.org/document/9603742. (accessed: 13.12.2023).

[6] Candice Bentéjac. *A Comparative Analysis of XGBoost*. URL: https://www.researchgate.net/publication/337048557_A_Comparative_Analysis_of_XGBoost. (accessed: 23.11.2023).

[7] FINN GUSTAFSSON. *Comparing Random Forest, XGBoost and Neural Networks With Hyperparameter Optimization by Nested Cross-Validation*. URL: https://rucforsk.ruc.dk/ws/portalfiles/portal/64939887/Machine_Learning_Bachelor_2019.pdf. (accessed: 23.11.2023).

[8] Tiago Lima Marinho. *OTIMIZAÇÃO DE HIPERPARÂMETROS DO XGBOOST UTILIZANDO META-APRENDIZAGEM*. URL: https://www.repositorio.ufal.br/bitstream/123456789/9851/1/Otimiza%C3%A7%C3%A3o%20de%20hiperpar%C3%A2metros%20do%20XGBoost%20utilizando%20meta-aprendizagem.pdf. (accessed: 23.11.2023).