PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Robson Luiz de Aquino Costa

ANÁLISE DE RISCO DE CRÉDITO

Rio de Janeiro
2023

Robson Luiz de Aquino Costa

ANÁLISE DE RISCO DE CRÉDITO

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Rio de Janeiro
2023

SUMÁRIO

1. Introdução

1.1. Contextualização

Análise de crédito e risco é uma solução financeira que procura investigar a capacidade de uma pessoa ou empresa de pagar pela compra de um produto ou pela prestação de algum serviço. E mesmo que a pessoa ou empresa analisada se mostre com condições de saldar a dívida que quer contrair, a avaliação do risco verifica se o histórico do comportamento dela no mercado é bom.

Dentre as principais instituições financeiras, o Nubank é uma das que mais tem se destacado no uso de Inteligência Artificial e times de *Data Science*.

O conjunto de dados a ser utilizado neste Projeto de *Data Science* parte de uma competição realizada pela Startup Nubank a fim de revelar talentos e potenciais contratações pela Fintech.

1.2. O problema proposto

Neste problema, o objetivo é prever qual a probabilidade de um cliente da Startup Nubank não cumprir com suas obrigações financeiras e deixar de pagar a sua fatura do Cartão de Crédito.

OBJETIVO: Criar um modelo que forneça a probabilidade de um cliente virar inadimplente.

Espera-se que um modelo seja capaz de minimizar as perdas financeiras do Nubank, porém minimizando também os falsos positivos.

2. Coleta de Dados

Os dados que serão utilizados nesta análise estão disponíveis para download por meio deste link. Consiste basicamente em um arquivo csv contendo 45.000 entradas e 43 colunas.

Este arquivo foi importado para uma estrutura DataFrame utilizando a biblioteca pandas a fim de possibilitar sua manipulação e análise.

Identificador	Nome da Coluna	Tipo de Dados	Descrição do campo
1	ids	String	Identificador
2	target_default	String	Identifica se pagou ou não pagou empréstimo.
3	score_1	String	Classificam a pontuação de crédito do cliente
4	score_2	String	Classificam a pontuação de crédito do cliente
5	score_3	float	Classificam a pontuação de crédito do cliente
6	score_4	float	Classificam a pontuação de crédito do cliente
7	score_5	float	Classificam a pontuação de crédito do cliente
8	score_6	float	Classificam a pontuação de crédito do cliente
9	risk_rate	float	Avaliação de risco
10	last_amount_borrowed	float	Último valor de emprestimo
11	last_borrowed_in_months	float	Último valor de emprestimo no mês
12	credit_limit	float	Limite de crétido
13	reason	String	Motivo do empréstimo
14	income	float	Valor do salário
15	facebook_profile	String	Perfil do facebook
16	state	String	Estado
17	zip	String	cep
18	channel	String	Canal
19	job_name	String	Trabalho
20	real_state	String	Estado
21	ok_since	float	Tempo de situação regular
22	n_bankruptcies	float	Falência
23	n_defaulted_loans	float	Empréstimos não pagos

24	n_accounts	float	Número de contas
25	n_issues	float	Número de casos de problema
26	application_time_applied	String	Tempo de investimento
27	application_time_in_funnel	int	Tempo de aplicação
28	email	String	E-mail
29	external_data_provider_credit_checks_last_2_year	float	Dados de empréstimo nos últimos 2 anos
30	external_data_provider_credit_checks_last_month	int	Dados de empréstimo no último mês
31	external_data_provider_credit_checks_last_year	float	Dados de empréstimo no último ano
32	external_data_provider_email_seen_before	float	E-mail anterior
33	external_data_provider_first_name	String	Primeiro nome
34	external_data_provider_fraud_score	int	Classificação de fraude
35	lat_lon	String	Latitude e longitude
36	marketing_channel	String	Canal de marketing
37	profile_phone_number	String	Número de telefone
38	reported_income	float	Salário informado
39	shipping_state	String	Estado
40	shipping_zip_code	int	CEP
41	profile_tags	String	Perfil
42	user_agent	String	Usuário
43	target_fraud	String	Informações sobre fraude

3. Processamento/Tratamento de Dados

Primeiro a verificação de dados ausentes.

```
ids                                                0
target_default                                  3259
score_1                                          562
score_2                                          562
score_3                                          562
score_4                                            0
score_5                                            0
score_6                                            0
risk_rate                                        562
last_amount_borrowed                           29956
last_borrowed_in_months                        29956
credit_limit                                   13800
reason                                           566
income                                           562
facebook_profile                                4458
state                                            562
zip                                              562
channel                                          562
job_name                                        3336
real_state                                       562
ok_since                                       26545
n_bankruptcies                                   697
n_defaulted_loans                                574
n_accounts                                       562
n_issues                                       11544
application_time_applied                           0
application_time_in_funnel                         0
email                                              0
external_data_provider_credit_checks_last_2_year   22628
external_data_provider_credit_checks_last_month        0
external_data_provider_credit_checks_last_year     15124
external_data_provider_email_seen_before            2233
external_data_provider_first_name                     0
external_data_provider_fraud_score                    0
lat_lon                                            1363
marketing_channel                                  3578
profile_phone_number                                  0
reported_income                                       0
shipping_state                                        0
shipping_zip_code                                     0
profile_tags                                          0
user_agent                                          722
target_fraud                                      43478
dtype: int64
```

Em relação à porcentagem de valores ausentes identificados neste dataset:

- Diversas variáveis como ['target_fraud', 'last_amount_borrowed', 'last_borrowed_in_months', 'ok_since', 'external_data_provider_credit_checks_last_2_year'] possuem mais da metade dos valores ausentes.

- As variáveis ['external_data_provider_credit_checks_last_year', 'credit_limit', 'n_issues'] possuem entre 25-34% do seus valores ausentes.

Em relação à porcentagem de valores preenchidos identificados neste dataset:

- As variáveis ['shipping_zip_code','score_4','score_5','profile_tags','score_6','application_time_in_funnel','shipping_state','reported_income','application_time_applied','profile_phone_number','external_data_provider_fraud_score','external_data_provider_first_name', 'external_data_provider_credit_checks_last_month',email',ids'] estão totalmente preenchidas.

- A variável alvo **target_default** contém valores nulos que serão eliminados do dataset.

Neste projeto, o caso mais extremo (target_fraud) não representa um problema, pois é uma variável alvo que não interessa para a análise de risco de inadimplência. Já as demais features deverão ser usadas com o devido cuidado.

Uma outra análise interessante de se fazer diz respeito à contagem de valores únicos por features. Muitas vezes, variáveis numéricas podem esconder classes/categorias que melhor representam uma feature, ou revelar uma quantidade elevada de classes para "variáveis categóricas".

4. Análise e Exploração dos Dados

Análise dos dados e verificação de valores únicos nos campos:

```
[33] print(base_credit.nunique().sort_values())

external_data_provider_credit_checks_last_2_year        1
channel                                                 1
target_fraud                                            2
target_default                                          2
external_data_provider_credit_checks_last_year          2
facebook_profile                                        2
last_borrowed_in_months                                 2
external_data_provider_credit_checks_last_month         4
n_defaulted_loans                                       5
real_state                                              5
email                                                   6
n_bankruptcies                                          6
score_1                                                 7
marketing_channel                                       9
shipping_state                                         25
score_2                                                35
n_issues                                               44
n_accounts                                             44
state                                                  50
external_data_provider_email_seen_before               62
risk_rate                                              82
score_3                                                88
ok_since                                              100
user_agent                                            297
application_time_in_funnel                            501
zip                                                   823
external_data_provider_fraud_score                   1001
last_amount_borrowed                                14325
reason                                              14874
credit_limit                                        20928
lat_lon                                             22412
profile_tags                                        26131
shipping_zip_code                                   28263
job_name                                            32265
external_data_provider_first_name                   32886
application_time_applied                            35543
reported_income                                     40025
income                                              41211
score_4                                             45000
score_5                                             45000
score_6                                             45000
profile_phone_number                                45000
ids                                                 45000
dtype: int64
```

A contagem de valores únicos mostra que as colunas **external_data_provider_credit_checks_last_2_year** e **channel** apresentam um

único valor possível. Como não há maiores informações sobre cada feature, as duas colunas serão descartadas para o modelo de Machine Learning.

A coluna **ids** é anônima e representa o identificador único do cliente. Normalmente essa coluna não influencia no modelo de machine learning.

A coluna **target_default** é o nosso alvo. Essa coluna representa no dataset se o cliente cumpriu ou não com as obrigações junto a instituição financeira.

As colunas **score_1** e **score_2** estão codificadas de alguma forma. As colunas **score_3**, **score_4**, **score_5** e **score_6** são numéricas. Essas variáveis classificam o cliente quanto a pontuação de crédito e iremos verificar a quantidade de códigos para analisar a sua transformação para categorias.

Existem outras variáveis que apresentam algum tipo de codificação, como [‘reason’, ‘state’, ‘zip’, ‘channel’, ‘job_name’, ‘real_state’] que estão codificadas e também precisarão de alguma análise mais aprofundada para saber se é possível extrair alguma informação das mesmas.

A coluna **lat_lon** está em formato string contendo uma tupla com as coordenadas. A coluna shipping_zip_code é referente ao CEP do canal de comunicação indicado pelo cliente, assim como a coluna zip provavelmente representa o CEP do imóvel onde o empréstimo foi realizado.

As colunas **last_amount_borrowed**, **last_borrowed_in_months**, **credit_limit** indicam a existência de empréstimos, quando o último empréstimo foi realizado e o limite de crédito para o cliente.

Como não temos informações sobre todas as colunas (features), vamos assumir como verdade que :

1. Algumas não são obrigatórias (exemplo perfil no Facebook);
2. Por não conhecermos em detalhes o dataset, para dados que se referem às últimas ocorrências, com informação null, vamos considerar zero , exemplo ['last_amount_borrowed', 'last_borrowed_in_months', 'n_issues'].

3. Para as variáveis numéricas, com valor NaN será adotado o valor da mediana;
4. Para as categóricas o valor mais frequente.

5. Criação de Modelos de Machine Learning

Utilizamos a estratégia de separação da variável alvo das demais e dividir o dataset entre dados de treino e teste com a função train_test_split.

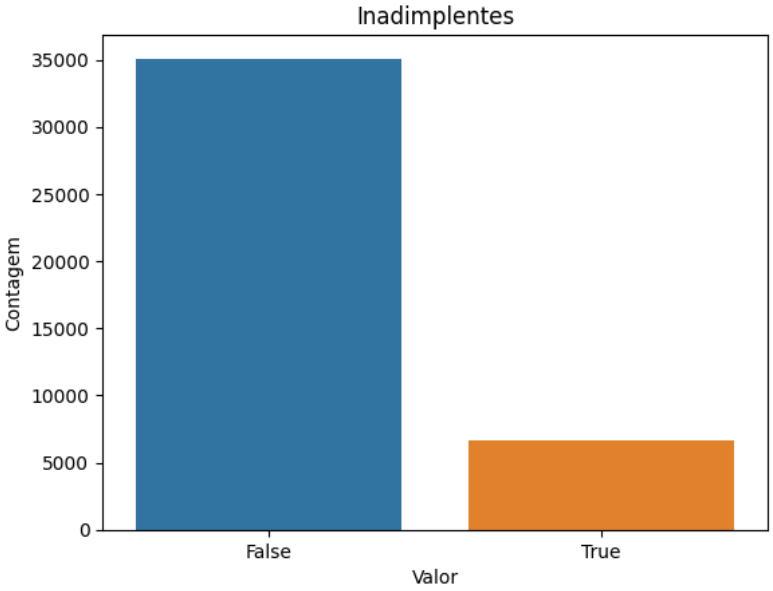
```
      # separando as variáveis independentes da variável alvo
      # X todas as colunas sem a coluna 'target_default
      # y apenas a coluna 'target_default

      X                =                encoded_df.drop('target_default',
axis=1).select_dtypes(exclude='object')
      y = encoded_df['target_default']
```

Utilizamos a Validação Cruzada com KFold.

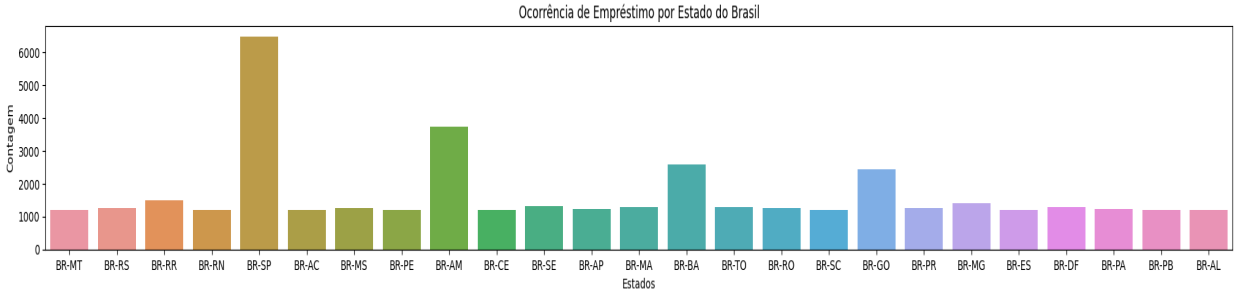
```
# importando o modelo de seleção - KFold
from sklearn.model_selection import KFold

# importando nossas metricas
from sklearn.metrics import accuracy_score

# importando nosso modelo de machine learning - XGBoost
from xgboost import XGBClassifier

# importando nossa biblioteca de funções matemáticas
import numpy as np
```

6. Apresentação dos Resultados

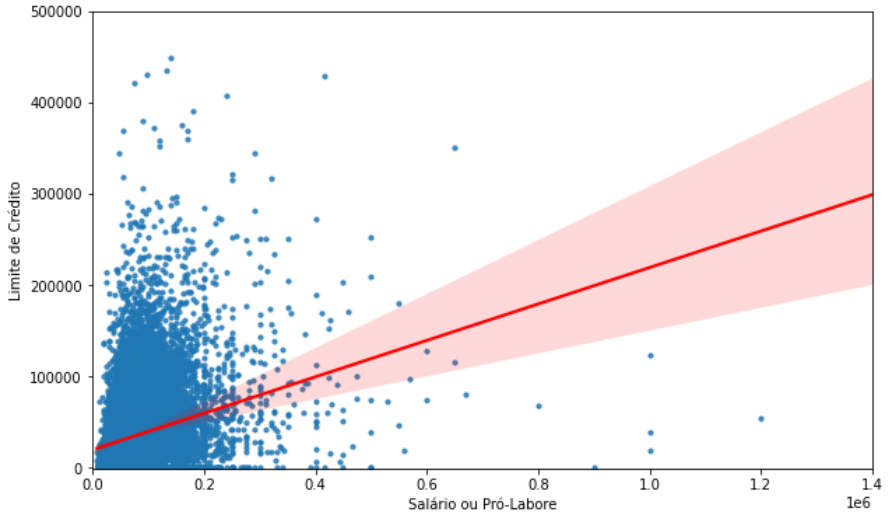
A primeira análise foi sobre proporção de ocorrências de inadimplentes de empréstimos - conforme a figura abaixo, com a proporção de aproximadamente 16% de inadimplentes de acordo com os dados analisados.



Abaixo a distribuição de quantidade de empréstimos nos estados brasileiros, sem entrar no mérito de quantitativo de inadimplentes, e sim, considerando apenas números de registro de empréstimos:

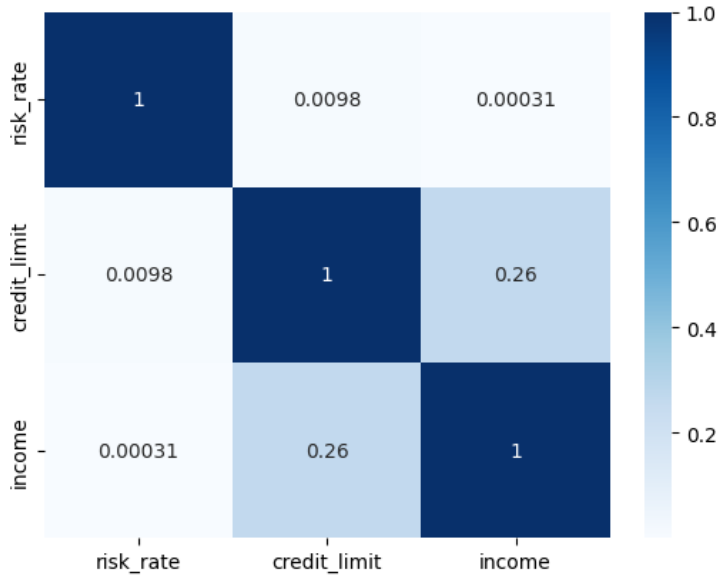


Na sequência observamos a dispersão entre os limites de créditos fornecidos aos clientes, de acordo com o valor salarial informado.



Do gráfico anterior podemos observar pela linha vermelha que conforme o ***income*** (ou salário / pró-labore) aumenta, o limite de crédito também se eleva (indica uma correlação positiva), porém existem algumas distorções na base de dados, pois alguns limites de crédito são bem elevados em relação ao outro parâmetro.

Na última análise, correlação entre as variáveis: risk_rate, credit_limit e income.



Por fim, implementamos um modelo de Machine Learning XGBClassifier considerando a validação cruzada com algoritmo de modelo de seleção KFold com a implementação de um estratégia de classificação em 5 repetições para cada divisão de dados em 3 splits, no que resultou uma acurácia média de 0.8409381911535425.

7. Links

- Apresentação do trabalho - vídeo Youtube (https://youtu.be/nB2plRWudZ8)
- Repositório de arquivos utilizados no trabalho no Github - https://github.com/raquinods/pos_ds.git

REFERÊNCIAS

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro, RJ: LTC, 2011. xvi, 378 p. ISBN 9788521618805