Pontifícia Universidade Católica do Rio de Janeiro

Pós Graduação Lato Sensu em Ciência de Dados e Analytics

Sprint: Engenharia de Dados (40530010057_20250_02)

Modelo de Previsão da Probabilidade de Inadimplência em Risco de Crédito

Aluno: Wesley Ramos Neres Santiago

Matrícula: 4052025002507

Objetivo geral:

A partir de uma base de dados contendo informações fictícias de contratos de crédito em instituições financeiras, o objetivo deste projeto é construir e treinar um modelo estatístico de Regressão Logística que possa estimar a probabilidade de inadimplência do contrato observando a capacidade creditícia do cliente, tipo de contrato, taxa de juros e montante do crédito contratado, com o objetivo de fazer a marcação de parâmetro de PD (*probability of default*) para cálculo de Perda Esperada, no contexto de Risco de Crédito.

Perguntas Gerais:

- Quais fatores presentes no perfil dos clientes e nas características das operações de crédito exercem maior influência na probabilidade de inadimplência estimada pelo modelo?

- O modelo desenvolvido apresenta capacidade preditiva satisfatória para identificar contratos que têm maior probabilidade de entrar em inadimplência?

Coleta e Modelagem dos Dados

Os dados utilizados neste projeto foram coletados em 20/11/2025 a partir de um *dataset* disponibilizado na plataforma Kaggle. Os dados foram obtidos pelo link: https://www.kaggle.com/datasets/arunbhuta/credit-analysis-probability-of-default e seu uso é permitido sob a licença Attribution-NonCommercial 4.0 International (CC by 4.0), conforme os temos disponíveis no link https://creativecommons.org/licenses/by-nc/4.0/.

O *dataset* utilizado neste modelo é um arquivo de valores separados por vírgula (CSV) e a modelagem utilizada foi o esquema Flat. O *dataset* foi carregado em um Catálogo de Dados no formato de tabela com nome de “base_credito” na plataforma Databricks Free Edition, possuindo 12 colunas, classificadas conforme o catálogo de dados a seguir:

Catálogo de Dados:

Nome da Coluna	Tipo do Dado	Descrição	Unidade	Valor Mínimo	Valor Máximo	Categoria
person_age	bigint	Idade do Cliente	Anos	20	144	N/A
person_income	bigint	Renda anual declarada	Dólares	4000	6000000	N/A
person_home_ownership	string	Situação de moradia do cliente	N/A	N/A	N/A	“RENT”, “OWN”, “MORTGAGE”,”OTHER”
person_emp_length	bigint	Tempo de emprego	Anos	0	123	N/A
loan_intent	string	Finalidade do empréstimo	N/A	N/A	N/A	“DEBTCONSOLIDATION”, “MEDICAL”,”EDUCATION”,”PERSONAL”,”HOMEIMPROVEMENT”,”VENTURE”
loan_grade	string	Classificação de risco do contrato	N/A	N/A	N/A	“A”,”B”,”C”,”D”,”E”,”F”,”G”
loan_amnt	bigint	Valor do contrato	Dólar	500	35000	N/A
loan_int_rate	double	Taxa de Juros do Contrato	Percentual	5,42	23,22	N/A
loan_status	bigint	Indicador de *default* (0 = adimplente, 1 =	N/A	0	1	N/A

		inadimplente)				
loan_percent_income	double	Proporção entre o valor do empréstimo e renda do cliente	Percentual	0,0	83	NA
cb_person_default_on_file	string	Histórico de inadimplência prévia (Y = sim, N = não)	N/A	Y	N	“Y”,”N”
cb_person_cred_hist_length	bigint	Tempo histórico de crédito no bureau	Anos	2	30	N/A

Arquitetura do Pipeline de Dados

O pipeline de dados desenvolvido neste projeto foi estruturado segundo o conceito de arquitetura em camadas Bronze, Silver e Gold.

A camada Bronze corresponde aos dados brutos importados diretamente da fonte original (*cr_loan.csv*), sem aplicação de regras de negócio ou tratamento estatístico, materializada na tabela *base_credito*.

A camada Silver representa os dados tratados, nos quais foram realizadas etapas de limpeza, tratamento de valores nulos, correção de outliers extremos e padronização de variáveis, resultando na tabela *base_clean*.

A camada Gold corresponde ao conjunto final de dados analíticos, preparado especificamente para a modelagem da probabilidade de default (PD), incluindo seleção de variáveis, transformação de dados categóricos por meio de one-hot encoding e criação de variáveis explicativas adicionais. Essa camada foi utilizada diretamente no treinamento e validação do modelo preditivo com o nome de *base_modelo*.

Essa organização em camadas garante maior rastreabilidade, governança, reprodutibilidade e clareza no fluxo de transformação dos dados ao longo do projeto.

Carga e Tratamento dos dados

A construção do *pipeline* de carga e tratamento dos dados foi realizada no *Databricks Free Edition*, utilizando as ferramentas de Catálogo e a construção de *notebooks* utilizando *Pyspark* para a construção dos códigos, de acordo com os seguintes passos:

1) Acessar o *Databricks Free Edition*
2) Na aba Catálogos, acessar o caminho *workspace/default*
3) Clicar no botão Criar e usar a opção Tabela
4) Carregar o arquivo .csv “cr_loan.csv”
5) Validar os dados inicialmente na prévia dos dados gerados pelo Databricks
6) Clicar em Criar
7) Na aba Espaço de Trabalho, clicar no botão criar e escolher a opção *Notebook*
8) Criar os *notebooks* “Casca”, “Exploração dos Dados”, “Tratamento dos Dados”, “Validação dos Dados”, “Modelo”, “Métricas de Modelo”, “Validação de Modelo”.

A base resultante desse processo é chamada de *base_credito* conforme descrito na seção Arquitetura do Pipeline de Dados.

Qualidade dos Dados

Na etapa de exploração dos dados, deve-se observar métricas estatísticas e também uma análise qualitativa dos dados de cada coluna, levando em conta o objetivo do modelo e as regras de negócio envolvidas no contexto de risco de crédito, buscando ajustar quaisquer inconsistências da base de dados a fim de evitar enviesamento ou predições incorretas do modelo.

O primeiro passo foi catalogar os dados e entender os tipos de variáveis em cada coluna, em seguida executar a contagem de linhas da base de dados para determinar sua volumetria.

Com a função *describe()* do *Pyspark* é possível observar as métricas estatísticas de cada coluna, como a média, contagem, valor mínimo e valor máximo. Na execução desta função, foi possível observar possíveis *outliers* extremos, que provavelmente, indicam inconsistências na base que requerem atenção, a listagem destas inconsistências é descrita a seguir:

1. O valor máximo de *person_age* é 144, indicando que pelo menos um cliente da base tem 144 anos de idade
2. O valor máximo de *person_income* é 6.000.000 (seis milhões), indicando que pelo menos um cliente da base tem como rendimento anual US$ 6.000.000
3. O valor máximo de *person_emp_length* é de 123 anos, indicando que o tempo de emprego de pelo menos um cliente é de 123 anos
4. O valor máximo de *cb_person_hist_length* é de 30 anos, indicando que pelo menos um cliente tem 30 anos de histórico de tomada de crédito

O próximo passo é identificar se há valores nulos na base, pois estes valores podem interferir na predição do modelo e precisam ser observados, para isto, uma função foi construída para selecionar e contar os valores nulos de cada coluna, o resultado desta análise está a seguir:

1. A coluna *person_emp_length* apresenta 895 valores vazios.
2. A coluna *loan_int_rate* apresenta 3116 valores vazios.

Para cada um dos casos, uma estratégia foi adotada a fim de garantir a melhor qualidade da base de dados, visando facilitar a visualização e a tomada de decisão, as métricas de quartil 25% (Q1), mediana (Q2), quartil 75% (Q3), IQR (*interquartile range*), limite inferior e limite superior das colunas numéricas foram salvas na tabela “base_outliers”. Com base nos dados obtidos, as seguintes estratégias foram adotadas:

1. Para a coluna *person_age* foi definido o valor máximo de 100 anos, valores acima como o valor de 144 observado foi substituído pelo valor 100, mantendo a aderência da informação a valores realizáveis dentro da proposta do modelo e do contexto de negócio do problema;
2. Para a coluna *person_income* não foi feito nenhum tratamento, pois o valor de US$ 6.000.000,00 não é um valor irreal, apesar de ser um *outlier* estatístico, portanto, não viola regra de negócio da proposta do modelo;
3. A coluna *person_emp_length* teve o valor máximo definido em 50 anos, valores acima como o valor de 123 observado foi substituìdo, mantendo a aderência da informação para valores realizáveis dentro da proposta do modelo e do contexto de negócio do problema;
4. A coluna *cb_person_hist_length* não sofreu nenhum tratamento, o *outlier* de 30 anos é um valor realizável dentro da proposta do modelo;
5. Os valores vazios de *person_emp_length* foram completos com a mediana da coluna pois dada a assimetria dos valores da coluna, esta métrica se faz mais adequada para evitar enviesamento do modelo por valores elevados;
6. Para *loan_int_rate* usamos a média dos valores agrupados pelo valor *loan_grade*, desta forma, conseguimos obter a taxa de juros média agrupada pelo *rating*, trazendo mais consistência nos valores.

A base resultante deste tratamento é chamada de *base_clean* conforme descrito na seção Arquitetura do Pipeline de Dados.

Treinamento do Modelo

O modelo foi treinado utilizando a técnica de Regressão Logística, método estatístico adequado para determinar a probabilidade de variáveis binárias, como a probabilidade de default.

A *base_clean* sofreu tratamentos adicionais para adequação aos requerimentos da Regressão Logística como a transformação das variáveis categóricas em variáveis binárias de modo que toda a base tenha em seu conteúdo apenas valores numéricos, para isto a função *get_dummies* do Pandas foi utilizada, resultando na base de nível ouro *pdf_clean* descrita na seção de Arquitetura do Pipeline de Dados.

Em seguida, a base foi dividida na proporção de 80%/20% para treino e teste respectivamente, desta forma o modelo pôde aprender como determinar a probabilidade de default observando 80% da base e os 20% restantes foram utilizados para teste.

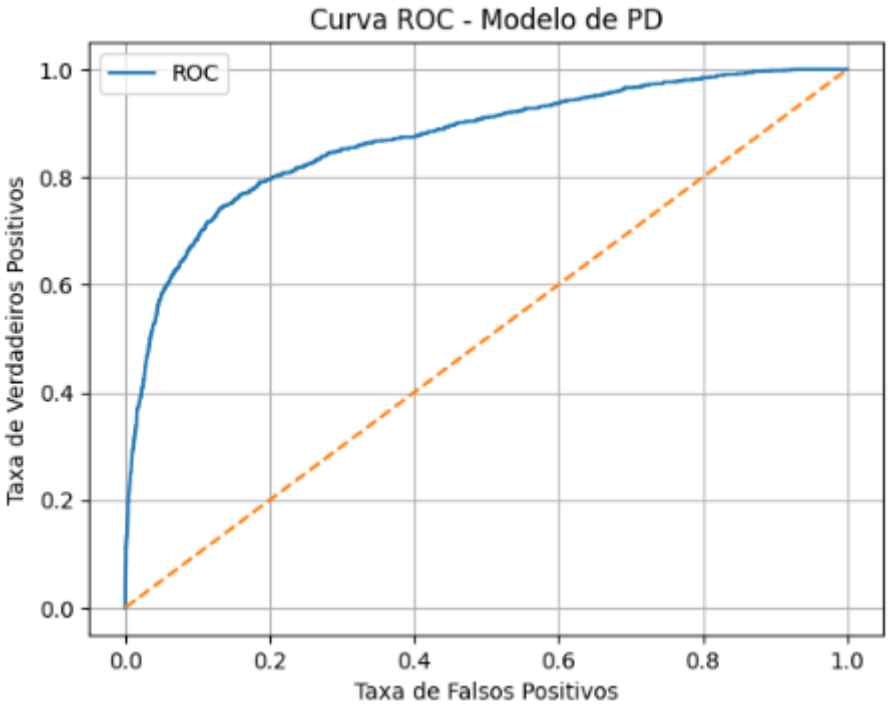
Ao fim do treinamento e aplicação da predição na porção de teste da base, foi possível extrair as métricas estatísticas que apontam o desempenho do modelo em prever a probabilidade de *default*, coeficientes das variáveis e capacidade de separação das classes de resultados.

Resultados do Modelo e Análise de Desempenho

Para determinar a capacidade do modelo de prever corretamente os clientes em *default* foram utilizadas métricas estatísticas de análise da precisão, acurácia e separação de classes, são elas:

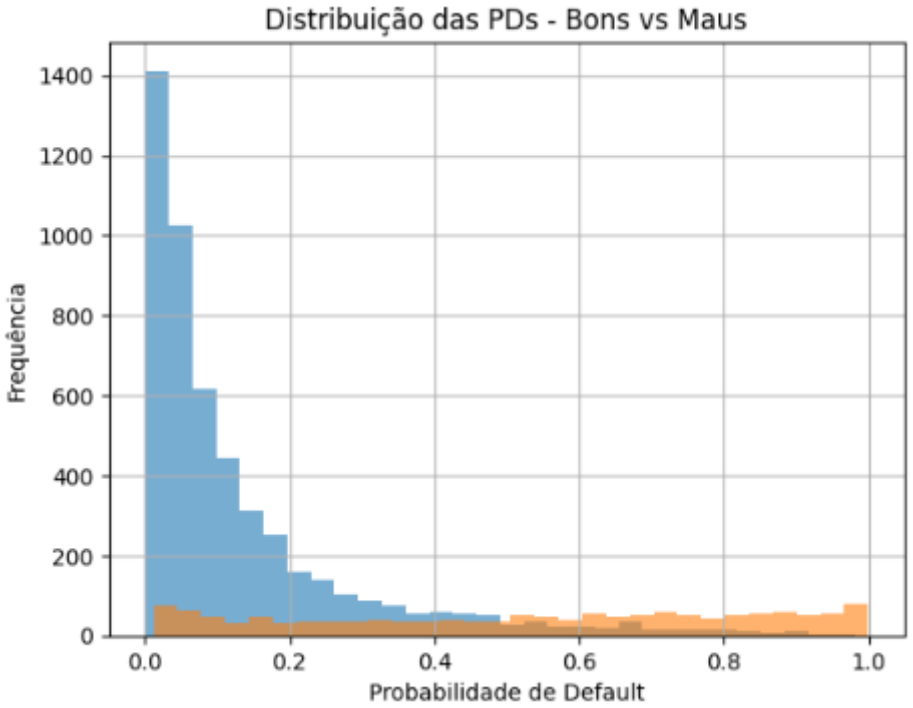
- Curva ROC (Curva Característica de Operação do Receptor): Gráfico que mostra a taxa de verdadeiros positivos contra a taxa de falsos positivos
- AUC (Área sob a Curva ROC): A área sob a curva que resume a capacidade do modelo de distinguir entre as classes, com valores variando de 0,5 (modelo aleatório) a 1,0 (modelo perfeito).
- Teste de Kolmogorov-Smirnov (KS): Gráfico que mostra a quantificação da distância absoluta entre duas funções de distribuição cumulativa, isto é, a distância entre as curvas acumuladas de verdadeiros e falsos marcados pelo modelo.
- Coeficientes de Regressão Logística: Gráfico que demonstra o impacto de cada variável na predição da variável resposta do modelo, onde valores positivos indicam maior impacto na marcação de clientes inadimplentes (PD = 1) e negativos na marcação de clientes adimplentes (PD = 0).

A seguir os gráficos dos resultados e a análise de cada uma das métricas do modelo:



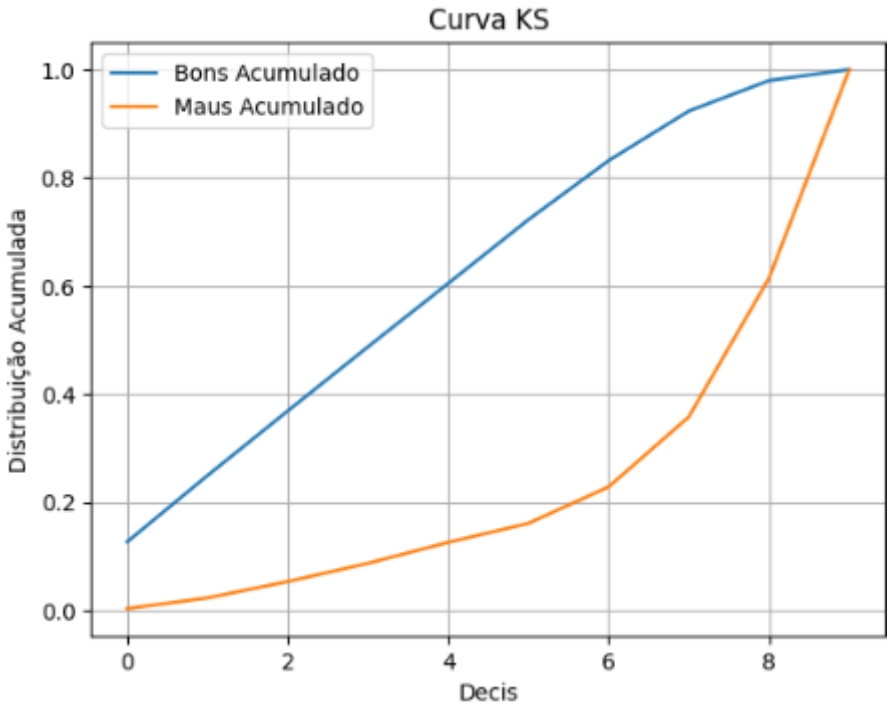
A Curva ROC do modelo apresenta formato fortemente convexo em direção ao canto superior esquerdo do gráfico, indicando elevada taxa de verdadeiros positivos mesmo para baixos níveis de falsos positivos. Esse comportamento confirma a alta capacidade discriminatória do modelo, em consonância com o valor de AUC obtido (0,87).

Observa-se que já nos primeiros percentis da distribuição de score o modelo é capaz de capturar parcela significativa dos inadimplentes, mantendo sob controle a classificação incorreta de bons clientes como maus pagadores.

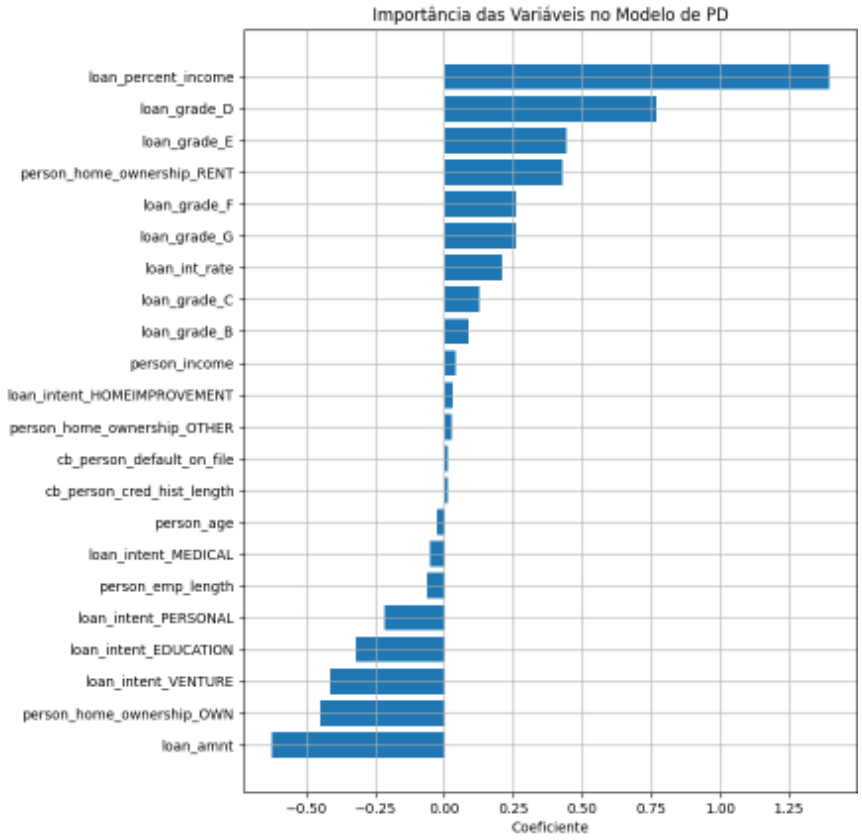


A análise da distribuição das probabilidades estimadas evidencia clara separação entre clientes adimplentes e inadimplentes. Observa-se que os bons pagadores concentram-se majoritariamente em faixas de baixa probabilidade de default, enquanto os inadimplentes apresentam maior incidência em faixas elevadas do score.

Existe uma região intermediária de sobreposição entre as distribuições, correspondente aos clientes de risco moderado, o que é esperado em modelos reais de crédito. De forma geral, o gráfico confirma visualmente o elevado poder discriminatório do modelo, em consonância com os valores de AUC e KS obtidos.



O modelo apresentou valor de KS igual a 0,603, indicando excelente capacidade de separação entre clientes adimplentes e inadimplentes. Esse resultado evidencia que, ao longo da ordenação do score, existe uma diferença superior a 60 pontos percentuais entre as distribuições acumuladas de bons e maus pagadores, o que caracteriza um modelo altamente discriminativo do ponto de vista de risco de crédito.



A análise dos coeficientes do modelo de regressão logística evidenciou que a variável com maior impacto positivo na probabilidade de inadimplência é a razão entre valor do empréstimo e renda do cliente (loan_percent_income), reforçando que altos níveis de comprometimento da renda elevam significativamente o risco de default.

Adicionalmente, observa-se que as categorias de pior classificação de crédito (loan_grade D, E, F e G) apresentam coeficientes positivos relevantes, indicando aumento progressivo do risco conforme a piora do rating do cliente. O fato de residir em imóvel alugado também se mostrou um importante fator de aumento de risco.

Por outro lado, variáveis como posse de imóvel próprio (home_ownership_OWN), maior valor de empréstimo, maior estabilidade no emprego, idade mais elevada e empréstimos com finalidade educacional ou de investimento apresentaram efeito redutor sobre a probabilidade de inadimplência.

De forma geral, os sinais e magnitudes dos coeficientes estão fortemente alinhados à teoria econômica e às práticas de mercado em gestão de risco de crédito.

Conclusão

Observando os resultados pode-se concluir que o modelo consegue predizer satisfatoriamente as ocorrências de inadimplência e adimplência de contratos de crédito utilizando-se com mais relevância das informações de percentual de renda comprometida, *rating* de crédito, tipo de propriedade da residência do tomador, idade e valor do contrato conforme visto nas métricas estatísticas.

As variáveis mais importantes para a definição da probabilidade de *default* segundo os coeficientes de regressão logística observados foram o percentual de comprometimento da renda, *ratings* de crédito inferiores, situação de residência do tipo aluguel e taxa de juros para previsão da inadimplência. Já para a previsão de adimplência, as variáveis mais importantes foram o valor do crédito, situação de residência do tipo própria e motivo do empréstimo do tipo educacional, pessoal e alavancagem.

Autoavaliação

Avaliando os resultados do modelo, concluo que pude desenvolver um MVP satisfatório para responder as perguntas postuladas no início do processo, agregando os conhecimentos adquiridos e trabalhados durante a sprint, como os modelos de dados, construção do pipeline, conceitos de banco de dados, governança e tratamento dos dados. Mantive a organização e a qualidade da documentação, facilitando o entendimento e replicação deste projeto.

Desenvolver este projeto me desafiou a construir desde a origem um modelo de aprendizado de máquina, observando os processos, metodologias e ferramentas, além de me proporcionar o contato com uma plataforma tão completa e poderosa como o Databricks, que está cada vez mais difundida na minha área de atuação como analista de dados no contexto de Risco de Crédito e Mercado. Essa oportunidade foi enriquecedora, pois agora tenho uma experiência e um modelo funcional que posso adicionar ao meu portfólio e representa um grande passo inicial na minha carreira como cientista de dados.