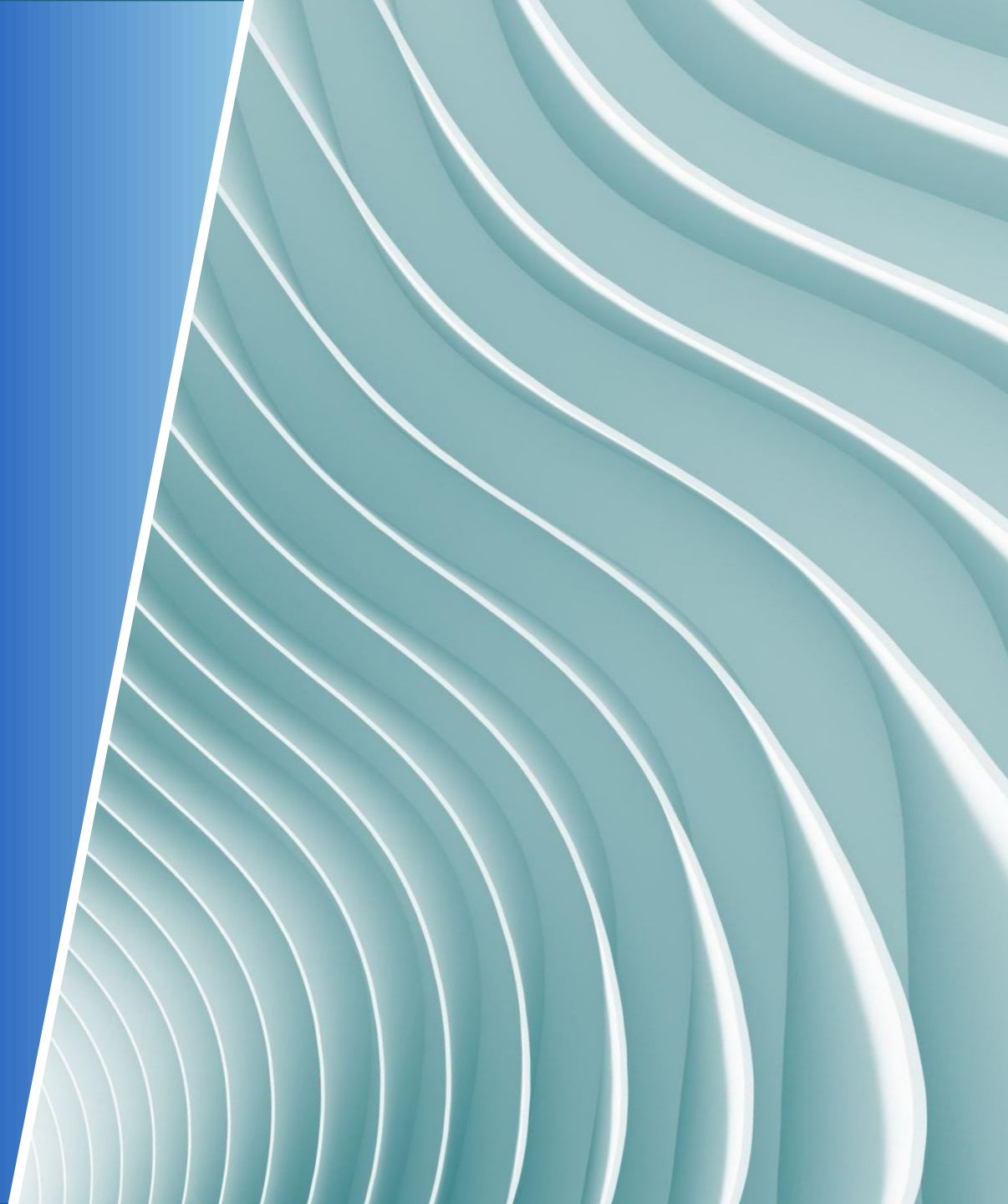
Vitalis Bank

Modelo de Concessão de Crédito | CC2506

Marina Cavalca



Vitalis Bank — um banco criado com o objetivo de proporcionar saúde financeira, combinando tecnologia, sensibilidade de mercado e compromisso com a eficiência.

Como primeira iniciativa, apresentamos nosso **modelo de crédito próprio** — desenvolvido com foco em inovação, responsabilidade e inteligência de dados.

Esse modelo reflete não apenas nossa visão estratégica, mas também o compromisso com decisões mais assertivas, inclusivas e **orientadas por dados**.

Público e Carteira

CC2506

Volume de Público

184.350

Inadimplência

7.74%

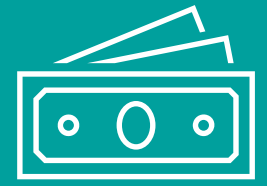
Empréstimo Rotativo (Revolving Loans)

- Volume de público:

17.443 (~10%)

- Inadimplência:

5.21%



Empréstimo Pessoal (Cash Loans)

- Volume de público:

166.907 (~90%)

- Inadimplência:

8.01%

Cliente Médio

CC2506

Renda (Anual)

131 Mil

Valor do Empréstimo

607 Mil

Idade (Anos)

44

Ocupação

18% - Serviços Gerais

Nível Escolar

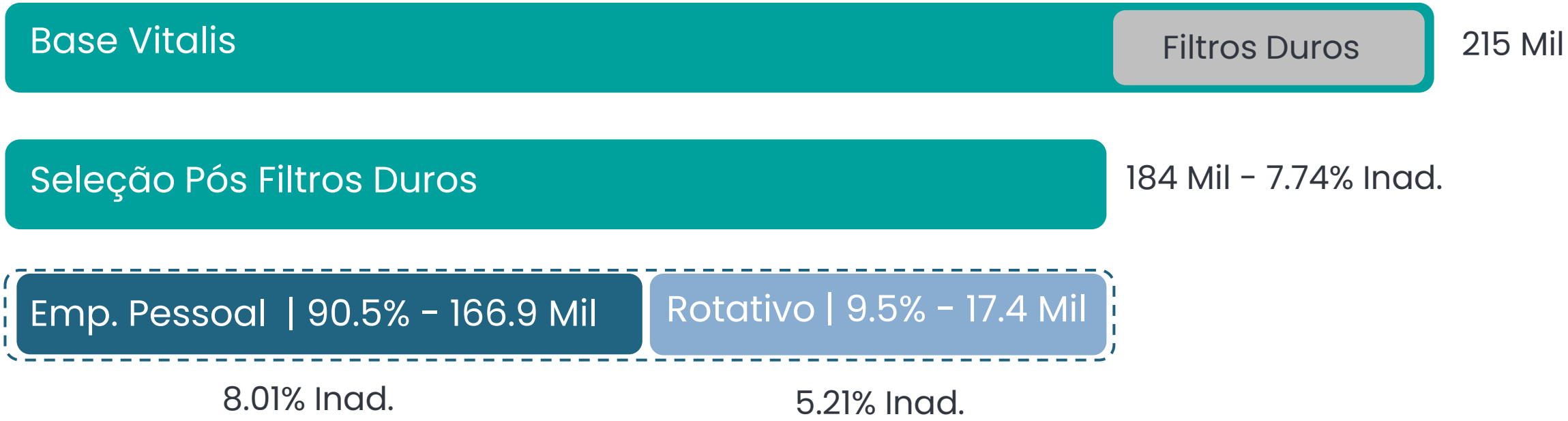
70% Ensino Médio

Membros da Família

50% - 2 membros

Definição do Público - Funil

CC2506



Modelagem

Base de dados

- Seleção das bases com os dados de interesse:
  - Dados externos (desconhecidos) + Dados do Público (internos e conhecidos (3 scores externos))
- Feature Engineering
- Seleção de Variáveis

Modelo de Machine Learning (ML)

- XGBoost e LightGBM
- Modelo base
- Seleção da Variáveis com valores de Importância > 0
- Otimização utilizando Optuna: Menor diferença entre AUC e KS treino e teste (Gráfico de Pareto)

Modelagem Base de Dados - Book 1.0

CC2506

Bureau – 2 (vars) Dados de outras instituições financeiras

Bureau Balance – 15 (vars) Informações mensais sobre créditos anteriores do cliente em outras instituições

Dados externos (desconhecidos)

Application – 170 (vars) Base de público para treino do modelo (Contendo 3 variáveis de score e 167 variáveis)

Dados do público (conhecidos)

187 variáveis Total de variáveis somando os três books

Modelagem Base de Dados - Criação do Book 2.0

CC2506

Bureau – 2 (vars)

Bureau Balance – 15 (vars)

Application – 170 (vars)

Book 1.0 com 187 variáveis

Feature Engineering

Book 2.0 – 962 variáveis

Modelagem Base de Dados - Seleção de Variáveis

Book 2.0 – 962 variáveis

Retira vars alta % missing, aplica One hot Encoder (vars cat) e inputa missing como -999

Retira vars constantes e quase constantes, calcula IV e KS para retirar vars por correlação (Pearson e Spearman)

Variáveis Selecionadas - 41

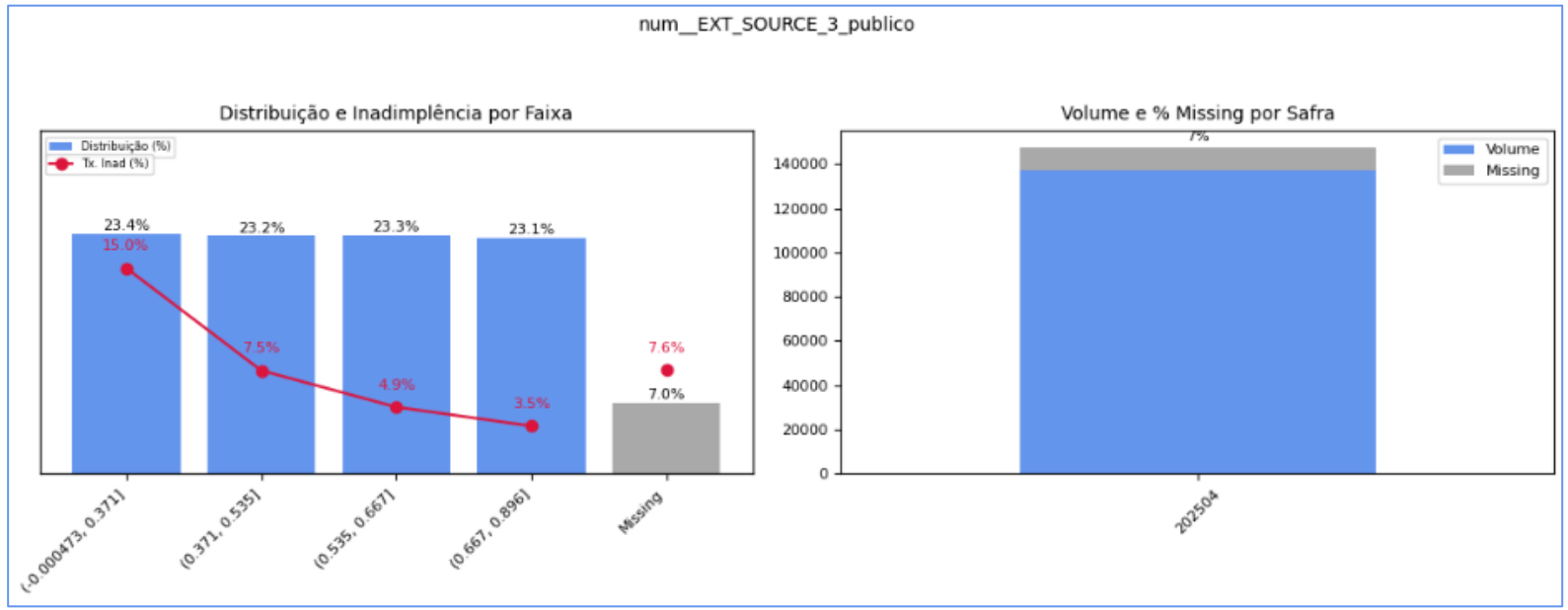
Modelagem Base de Dados – 20 Variáveis com maior IV

CC2506

	Variável	IV	KS
1	num__EXT_SOURCE_3_publico	32,6%	25,2%
2	num__EXT_SOURCE_2_publico	28,0%	21,2%
3	num__EXT_SOURCE_1_publico	10,3%	11,8%
4	num__PAYMENT_RATE_publico	10,2%	9,4%
5	num__AMT_GOODS_PRICE_publico	10,1%	9,2%
6	num__DAYS_EMPLOYED_publico	9,8%	8,5%
7	num__DAYS_BIRTH_publico	8,0%	11,5%
8	num__VL_TOT_VL_TOT_AMT_CREDIT_SUM_DEBT_ULTIMOS_12_MESES_externo	7,8%	13,3%
9	num__CREDIT_TO_GOODS_RATIO_publico	7,6%	12,2%
10	num__INCOME_TO_EMPLOYED_RATIO_publico	7,5%	12,0%
11	num__VL_MED_VL_TOT_AMT_CREDIT_SUM_DEBT_ULTIMOS_24_MESES_externo	6,8%	11,9%
12	num__QT_MAX_QT_MAX_DAYS_CREDIT_UPDATE_ULTIMOS_24_MESES_externo	6,4%	11,1%
13	num__QT_MAX_QT_MAX_DAYS_CREDIT_UPDATE_ULTIMOS_6_MESES_externo	5,9%	11,1%
14	num__REGION_RATING_CLIENT_W_CITY_publico	5,0%	6,2%
15	num__QT_MAX_QT_MAX_DAYS_CREDIT_ENDDATE_ULTIMOS_36_MESES_externo	4,7%	8,6%
16	num__VL_TOT_VL_TOT_AMT_CREDIT_SUM_LIMIT_ULTIMOS_24_MESES_externo	4,6%	6,9%
17	num__VL_TOT_VL_TOT_AMT_CREDIT_MAX_OVERDUE_ULTIMOS_24_MESES_externo	4,6%	6,8%
18	cat__NAME_EDUCATION_TYPE_publico_Higher_education	4,3%	8,5%
19	num__DAYS_LAST_PHONE_CHANGE_publico	4,2%	9,1%
20	cat__NAME_INCOME_TYPE_publico_Working	4,1%	10,1%

Modelagem Base de Dados – Variáveis com maior IV

CC2506



Variável EXT_SOURCE_3:

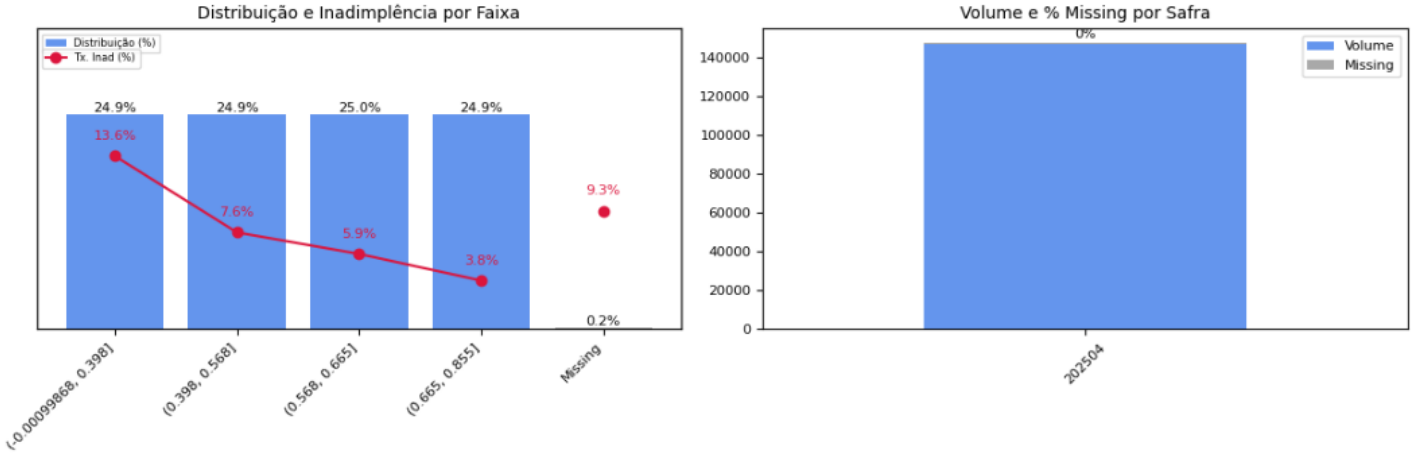
- A Bad Rate DIMINUI de acordo com que o Score AUMENTA.
- Os valores missing tem Bad Rate de 7.6% e uma volumetria em torno de 7%. O valor médio desse score é de 0.510 com uma Bad de 7.5%

EXT_SOURCE_3: Pontuação normalizada da fonte de dados externa 3. Esse escore externo é uma combinação de várias variáveis fortes, já com entendimento de negócio, que já tem sua estabilidade testada e aprovada.

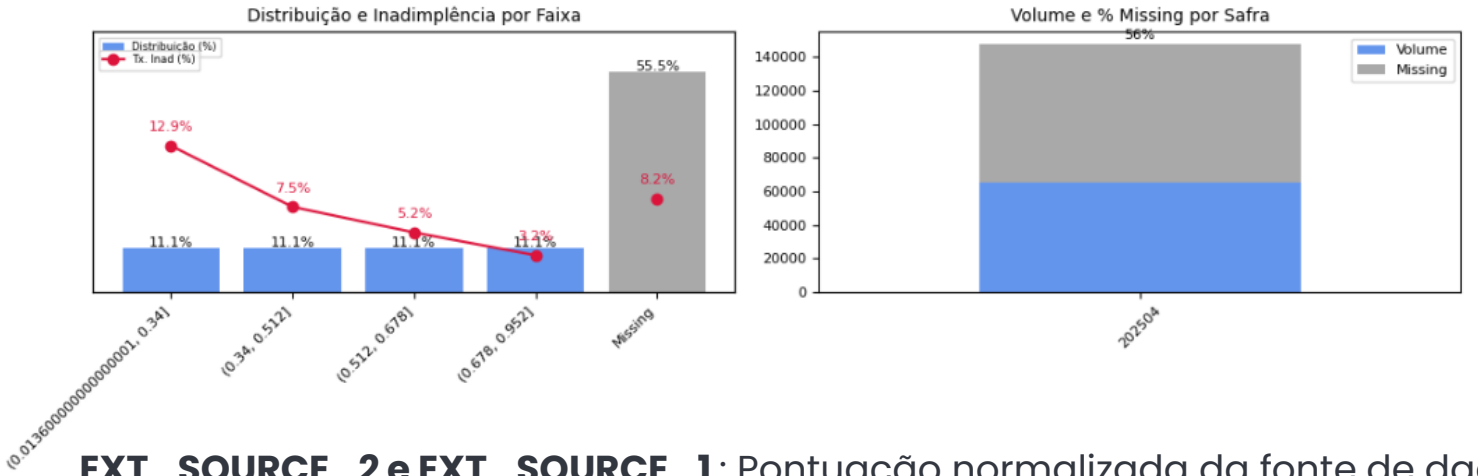
Modelagem Base de Dados – Variáveis com maior IV

CC2506

num__EXT_SOURCE_2_publico



num__EXT_SOURCE_1_publico



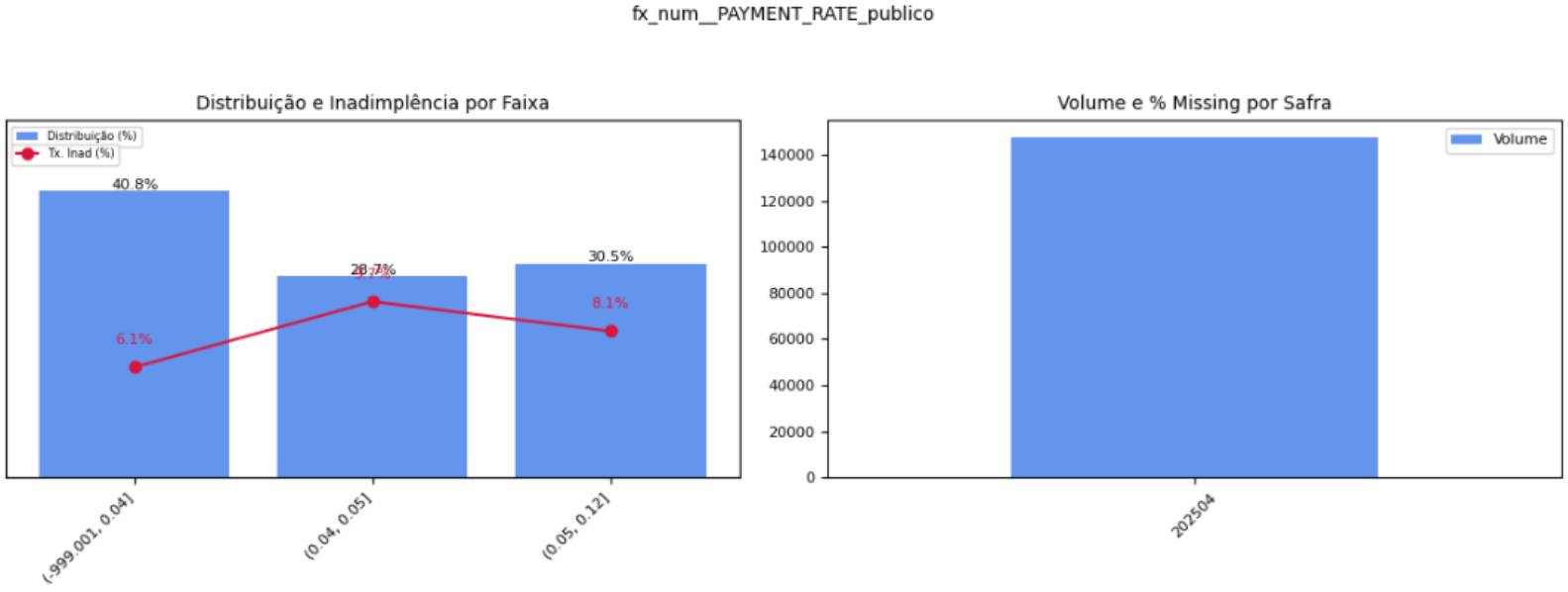
Variáveis EXT_SOURCE_2 e EXT_SOURCE_1:

- A Bad Rate DIMINUI de acordo com que o Score AUMENTA para os dois casos.
- Os valores missing de EXT_SOURCE_2 tem Bad Rate de 9.3% e uma volumetria em torno de 0.2%. O valor médio desse score é 0.517 e a Bad de 7.6%.
- Os valores missing de EXT_SOURCE_1 tem Bad Rate de 8.2% e uma volumetria em torno de 55.5%. O valor médio desse score é 0.555 e a Bad de 7.5%.

EXT_SOURCE_2 e EXT_SOURCE_1: Pontuação normalizada da fonte de dados externa 2 e 1. Esses escores externos são uma combinação de várias variáveis fortes, já com entendimento de negócio, que já tem sua estabilidade testada e aprovada.

Modelagem Base de Dados – Variáveis com maior IV

CC2506



Variável PAYMENT RATE (Taxa de Pagamento):

- A Bad Rate AUMENTA de 6.1% para 9.7% para uma taxa de Pagamento de zero até 5%, e DIMINUI para 8.1% de 0.51 até 12%.
- A variável PAYMENT_RATE não apresenta monotonicidade com a inadimplência porque pode estar interagindo fortemente com outra variável
- Os valores missing são quase zero

PAYMENT_RATE: Taxa de Pagamento, representa a proporção entre o valor do pagamento feito pelo cliente (parcela paga ou valor mínimo) e o valor total da dívida ou prestação devida.

Modelagem Base de Dados – Variáveis com maior IV

CC2506

VOLUMETRIA

	NAME_CONTRACT_TYPE_publico		
fx_PAYMENT_RATE_publico	Cash loans	Revolving loans	Total
(-999.001, 0.04]	60.224	0	60.224
(0.04, 0.05]	28.274	13.985	42.259
(0.05, 0.12]	44.977	20	44.997
Total	133.475	14.005	147.480

TARGET

	NAME_CONTRACT_TYPE_publico		
fx_PAYMENT_RATE_publico	Cash loans	Revolving loans	Total
(-999.001, 0.04]	6,1%	0,0%	6,1%
(0.04, 0.05]	11,9%	5,3%	9,7%
(0.05, 0.12]	8,1%	0,0%	8,1%
Total	8,0%	5,3%	7,7%

*__Variável NAME_CONTRACT_TYPE:__

Cash Loans (Emp. Pessoal)

Revolving Loans (Emp. Rotativo)

Interação da PAYMENT RATE com a Name Contract Type (Nome do Tipo de Contrato):

- A faixa intermediária (0.04, 0.05] tem: A maior concentração de contratos Revolving, e a maior inadimplência para esse tipo de contrato.
- A faixa seguinte (> 0.05) tem menos contratos Revolving, e por isso a inadimplência média cai, mesmo com o aumento da PAYMENT_RATE.
- Para empréstimos pessoais (Cash Loans), ela reflete diretamente a capacidade e disciplina de pagamento.
- Para crédito rotativo (Revolving Loans), ela reflete comportamento de uso do crédito, podendo indicar risco futuro mesmo quando o pagamento está em dia.
- Por isso, a PAYMENT_RATE interage com o CONTRACT_TYPE, afetando seu significado e sua relação com a inadimplência.

Modelagem Base de Dados – Variáveis com maior IV

CC2506

Variável AMT GOODS PRICE(Valor do Bem Financiado):

- A distribuição de clientes está bem equilibrada entre as faixas de valor financiado (cerca de 25% por faixa).
- A inadimplência não é monotônica:
- Sobe para 10,6% na faixa entre R$ 238.500 e R$ 450.000.
- Cai para 5,5% na faixa acima de R$ 688.500.
- Isso sugere que valores intermediários têm maior risco do que valores muito altos.

AMT_GOODS_PRICE: Para empréstimos ao consumo, é o preço dos bens para os quais o empréstimo é concedido.

Modelagem Base de Dados – Variáveis com maior IV

CC2506

VOLUMETRIA

	cat_Bens_Consumo*				
fx_AMT_GOODS_PRICE_publico	**0.0**	**1.0**	**2.0**	**3.0**	**Total**
(-999.001, 238500.0]	7.483	7.479	19.354	3.075	37.391
(238500.0, 450000.0]	7.839	8.943	16.463	3.958	37.203
(450000.0, 688500.0]	7.592	8.471	16.940	3.875	36.878
(688500.0, 4050000.0]	6.531	10.508	13.787	5.182	36.008
Total	29.445	35.401	66.544	16.090	147.480

TARGET

	cat_Bens_Consumo*				
fx_AMT_GOODS_PRICE_publico	**0**	**1**	**2**	**3**	**Total**
(-999.001, 238500.0]	8,7%	6,7%	7,5%	6,9%	7,5%
(238500.0, 450000.0]	12,0%	9,7%	10,7%	9,2%	10,6%
(450000.0, 688500.0]	7,8%	7,0%	7,4%	6,7%	7,3%
(688500.0, 4050000.0]	6,1%	5,2%	5,8%	4,7%	5,5%
Total	8,8%	7,1%	7,9%	6,7%	7,7%

*Bens de Consumo:
0: Não comprou imóvel nem veículo
1: Comprou imóvel e veículo
2: Comprou imóvel, não veículo
3: Não comprou imóvel, comprou veículo

Interação da AMT_GOODS_PRICE com a Bens_Consumo:

A inadimplência tende a diminuir conforme o valor do bem adquirido aumenta, exceto na faixa entre R$ 238.500 e R$ 450.000, que concentra o maior risco da amostra.

▲ **Faixa com maior inadimplência (R$ 238.500 a R$ 450.000):**

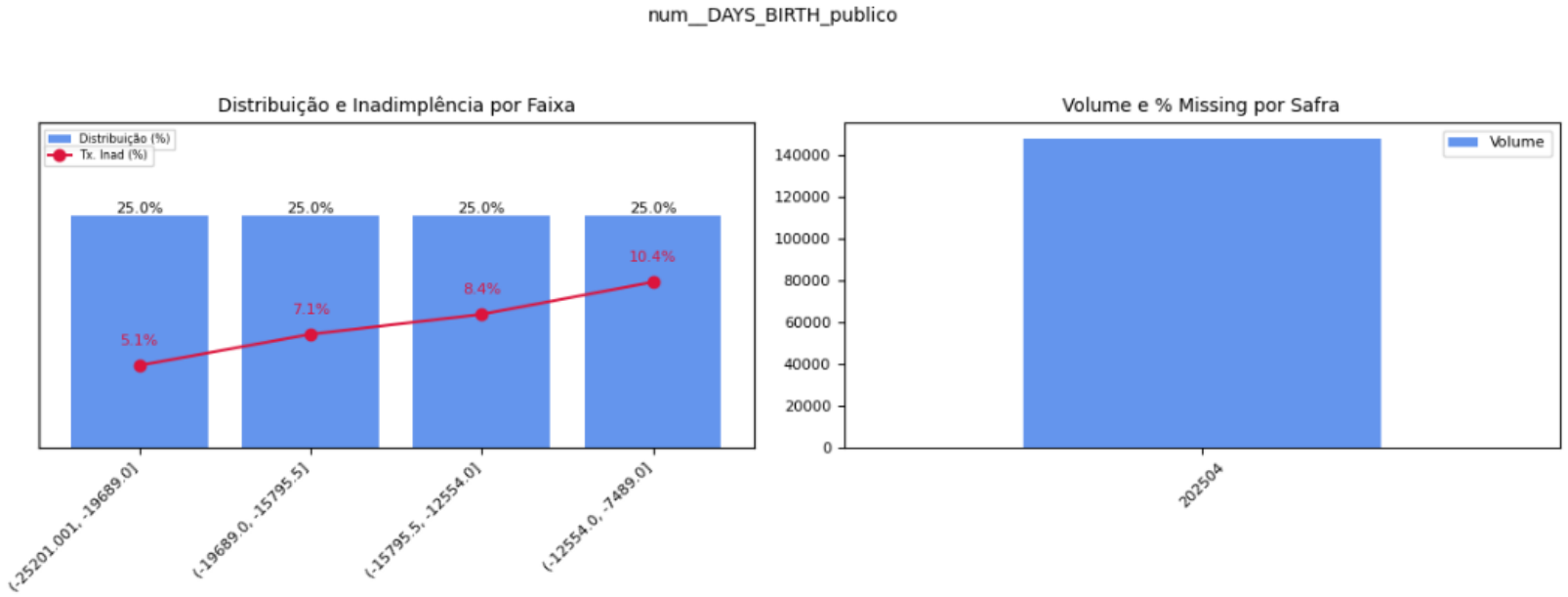
- Clientes que não compraram bens duráveis apresentam maior risco — possivelmente por destinarem o crédito a fins não colateralizados ou de maior volatilidade financeira.

▼ **Faixa com menor inadimplência (R$ 688.500 a R$ 4.050.000):**

- Valores mais altos de bens estão associados a perfis de menor risco, especialmente quando o crédito está vinculado à aquisição de ativos reais como imóvel ou veículo.

Modelagem Base de Dados – Variáveis com maior IV

CC2506



Variável DAYS_BIRTH:

- A Bad Rate AUMENTA de acordo com que a idade do cliente DIMINUI.
- A Bad Rate dobra de 5.1% para clientes na média dos 60 anos, para uma Bad de 10.4% para clientes na média dos 27 anos.

DAYS_BIRTH: Valor negativo que indica a quantidade de dias desde o nascimento.

Modelos

Foram utilizados dois tipos de modelos:

XGBoost (Extreme Gradient Boosting):

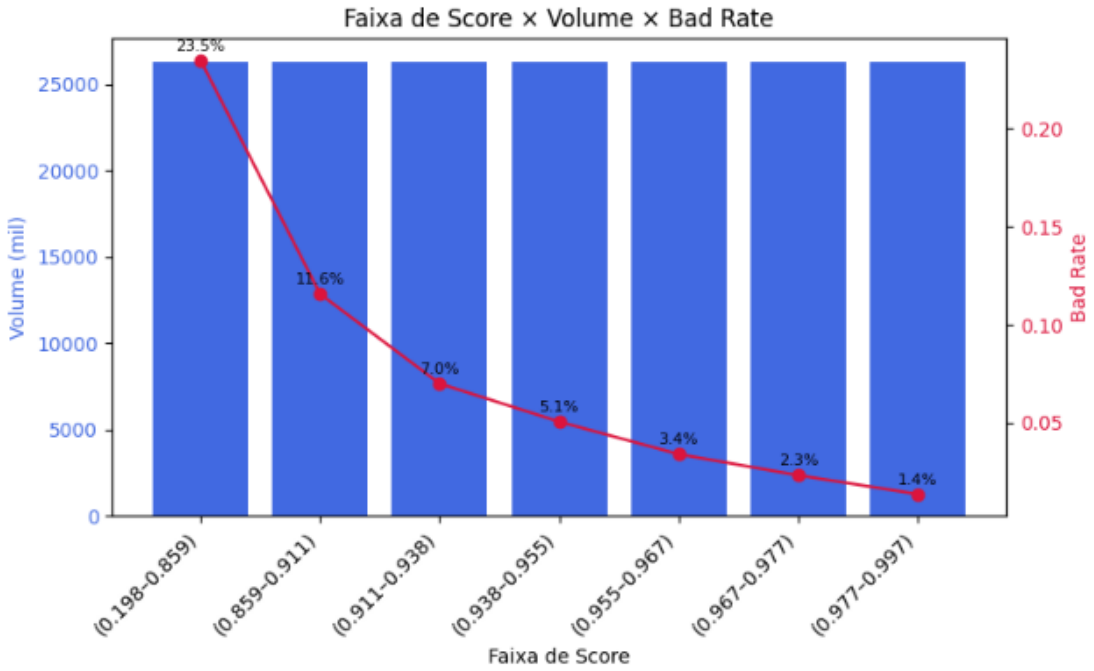
- Algoritmo de boosting que constrói árvores de decisão de forma sequencial, corrigindo os erros das anteriores.
- Robusto com capacidade de lidar com dados desbalanceados e oferecer alto desempenho

LightGBM (Light Gradient Boosting Machine)

- Variação otimizada do boosting, desenvolvida para ser extremamente rápida e eficiente, especialmente com grandes volumes de dados.
- Utiliza técnicas como crescimento por folha e histogramas discretos, o que permite treinar modelos com menor consumo de memória e tempo.

Modelo XGBoost - Performance

CC2506



Faixa de Score vs. Volume vs. Inadimplência

- **Monotonicidade**: inadimplência decresce de forma ordenada entre grupos.
- **Score com bom alcance**: score mínimo de 0,199 até 0,997, ampla cobertura.
- **Segmentação balanceada**: cada faixa de score cobre ~14% do público → boa granularidade.

Desempenho do Modelo

Métrica	Treino	Teste	Geral
AUC	76,30%	75,54%	76,15%
Gini	52,60%	51,09%	52,30%
KS	39,66%	38,72%	39,44%

O modelo apresenta **alto poder de discriminação**, com **performance consistente entre treino e teste**, indicando boa capacidade preditiva e baixo risco de overfitting:

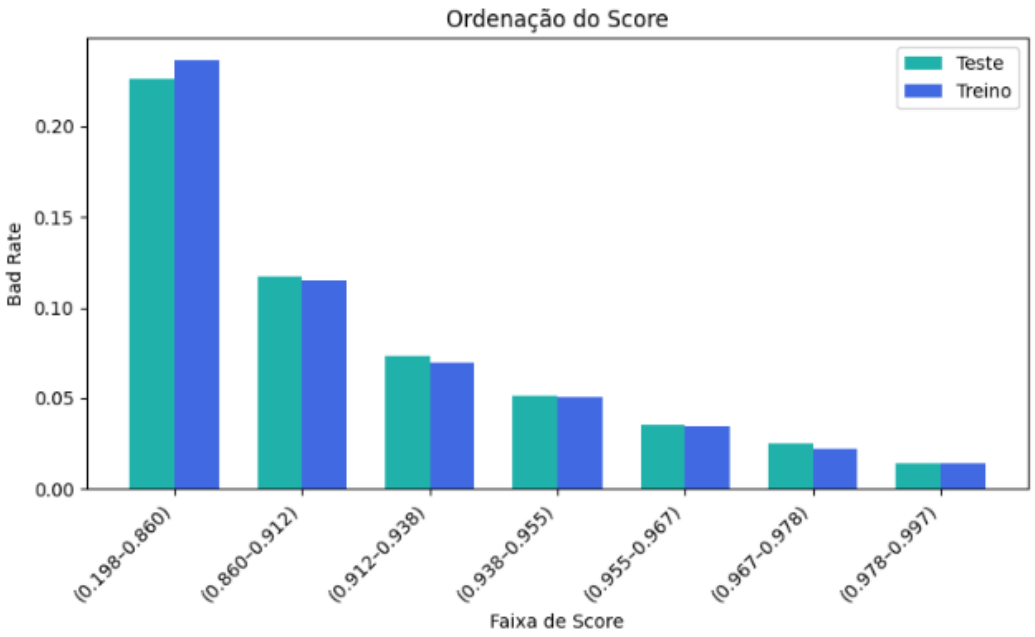
•**AUC de 75,54% no teste**, sinalizando boa separação entre bons e maus pagadores.

•**KS de 38,72% no teste**, acima do patamar de referência (>30%), demonstrando forte diferenciação entre os grupos.

•**Estabilidade entre treino e teste** (diferença de <1 p.p. nas métricas), reforçando a **robustez e generalização** do modelo.

Modelo XGBoost - Performance

CC2506



Ordenação de Score:

Os scores estão ordenados de modo que quanto maior o score, menor o risco, ou seja, a Inadimplência diminui com o aumento do score. Os valores de Bad Rate entre os dados de Treino e Teste não apresentam grandes variações.

Scorecard

gh	faixa_score	min_score	max_score	bad_rate	volume	vol_acum	% vol
1	(0,977–0,997)	0,977445	0,996867	1,37%	26.336	26.336	14%
2	(0,967–0,977)	0,966848	0,977444	2,33%	26.336	52.672	29%
3	(0,955–0,967)	0,954694	0,966848	3,41%	26.335	79.007	43%
4	(0,938–0,955)	0,937588	0,954694	5,06%	26.336	105.343	57%
5	(0,911–0,938)	0,911349	0,937587	7,00%	26.335	131.678	71%
6	(0,859–0,911)	0,859137	0,911347	11,57%	26.336	158.014	86%
7	(0,198–0,859)	0,19921	0,859137	23,47%	26.336	184.350	100%

Apetite financeiro:

- O corte de aprovação pode ser feito com base no apetite de risco (por ex., até GH4 → inadimplência máxima de 5%, cobre 57% do público.
- Aprovar até o GH 5, com 71% do público com uma Inadimplência de 7% (atual Inadimplência do público de modelagem)

* Novos cortes podem ser feitos de acordo com a política

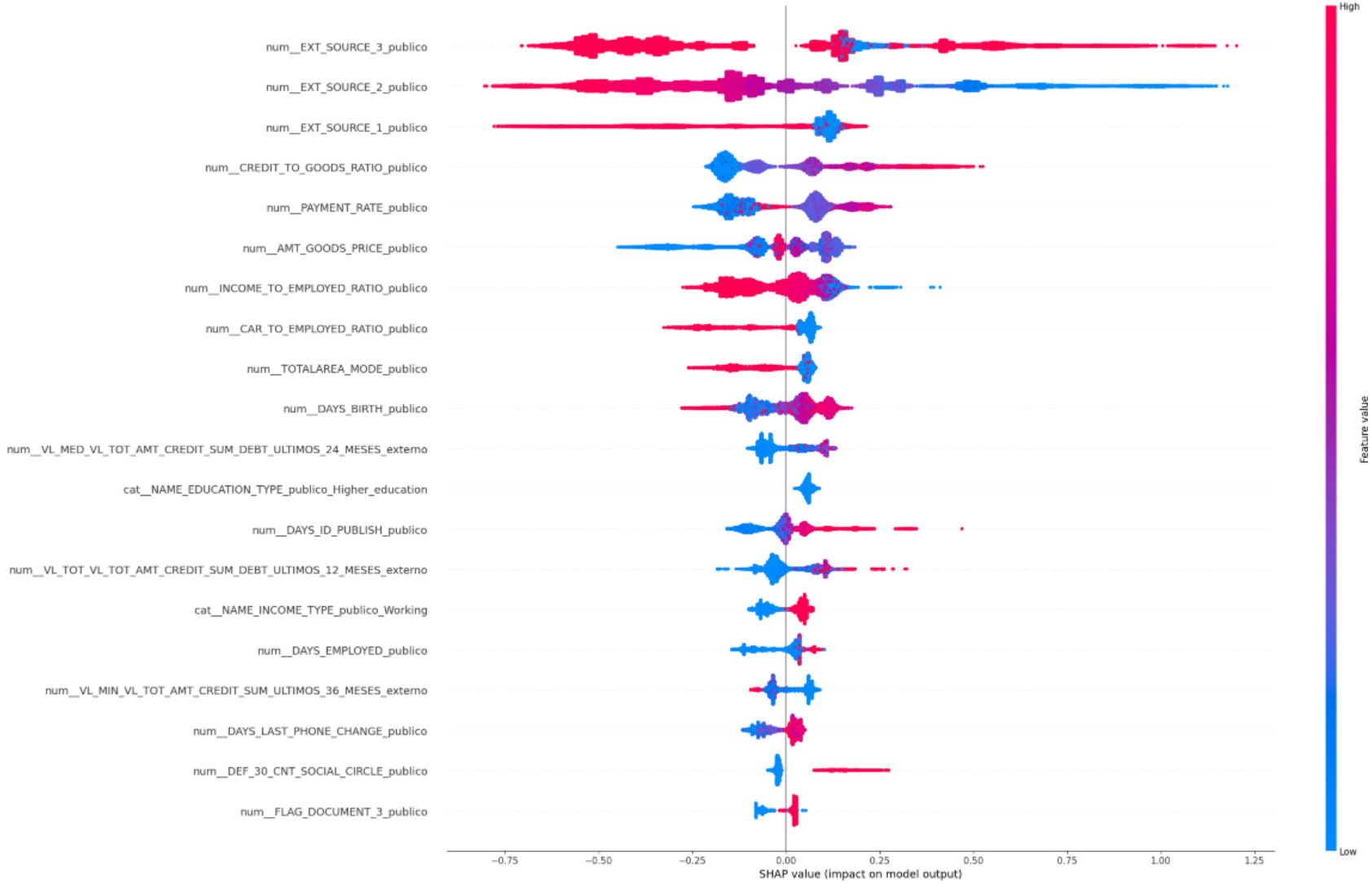
Modelo XGBoost – 31 Variáveis Selecionadas

CC2506

	Variável	Importância	IV	KS
1	num__EXT_SOURCE_3_publico	12,2%	32,8%	25,3%
2	num__EXT_SOURCE_2_publico	10,4%	27,7%	21,2%
3	num__VL_TOT_VL_TOT_AMT_CREDIT_MAX_OVERDUE_ULTIMOS_12_MESES_externo	6,5%	3,8%	4,2%
4	num__VL_TOT_VL_TOT_AMT_CREDIT_SUM_DEBT_ULTIMOS_12_MESES_externo	6,0%	7,5%	13,2%
5	num__CREDIT_TO_GOODS_RATIO_publico	5,6%	7,7%	12,2%
6	num__EXT_SOURCE_1_publico	4,6%	10,5%	11,7%
7	num__CAR_TO_EMPLOYED_RATIO_publico	4,2%	3,5%	6,8%
8	cat__NAME_EDUCATION_TYPE_publico_Higher_education	3,7%	3,3%	6,8%
9	num__INCOME_TO_EMPLOYED_RATIO_publico	3,3%	7,6%	12,2%
10	num__DAYS_BIRTH_publico	3,3%	8,2%	11,7%
11	num__FLAG_DOCUMENT_3_publico	3,2%	2,9%	7,4%
12	num__PAYMENT_RATE_publico	2,7%	10,5%	9,6%
13	num__TOTALAREA_MODE_publico	2,6%	3,4%	8,3%
14	cat__NAME_INCOME_TYPE_publico_Working	2,5%	4,1%	10,0%
15	num__DEF_30_CNT_SOCIAL_CIRCLE_publico	2,5%	1,5%	3,8%
16	num__QT_MAX_QT_MAX_DAYS_CREDIT_UPDATE_ULTIMOS_6_MESES_externo	2,2%	5,8%	11,2%
17	num__AMT_GOODS_PRICE_publico	1,9%	10,0%	9,2%
18	num__QT_MAX_QT_MAX_DAYS_CREDIT_UPDATE_ULTIMOS_24_MESES_externo	1,9%	5,8%	10,9%
19	cat__OCCUPATION_TYPE_publico_Laborers	1,9%	2,3%	6,1%
20	num__VL_MED_VL_TOT_AMT_CREDIT_SUM_DEBT_ULTIMOS_24_MESES_externo	1,8%	6,5%	11,7%
21	num__DAYS_ID_PUBLISH_publico	1,7%	3,7%	8,1%
22	num__DAYS_LAST_PHONE_CHANGE_publico	1,7%	3,8%	8,8%
23	num__REGION_RATING_CLIENT_W_CITY_publico	1,7%	4,8%	6,2%
24	num__QT_MAX_QT_MAX_CREDIT_DAY_OVERDUE_ULTIMOS_36_MESES_externo	1,7%	1,6%	1,4%
25	num__VL_TOT_VL_TOT_AMT_CREDIT_SUM_LIMIT_ULTIMOS_24_MESES_externo	1,6%	4,7%	6,8%
26	num__QT_MAX_QT_MAX_DAYS_CREDIT_ENDDATE_ULTIMOS_36_MESES_externo	1,6%	4,7%	8,6%
27	num__VL_TOT_VL_TOT_AMT_CREDIT_MAX_OVERDUE_ULTIMOS_24_MESES_externo	1,6%	4,8%	6,9%
28	num__DAYS_EMPLOYED_publico	1,5%	9,8%	8,3%
29	num__REGION_POPULATION_RELATIVE_publico	1,4%	3,3%	5,8%
30	num__INCOME_TO_BIRTH_RATIO_publico	1,2%	2,1%	5,5%
31	num__VL_MIN_VL_TOT_AMT_CREDIT_SUM_ULTIMOS_36_MESES_externo	1,1%	3,6%	5,3%

Modelo XGBoost – Shap Value

CC2506



Explicação do comportamento das variáveis:

num_EXT_SOURCE_3_publico (e EXT_SOURCE_2, EXT_SOURCE_1):

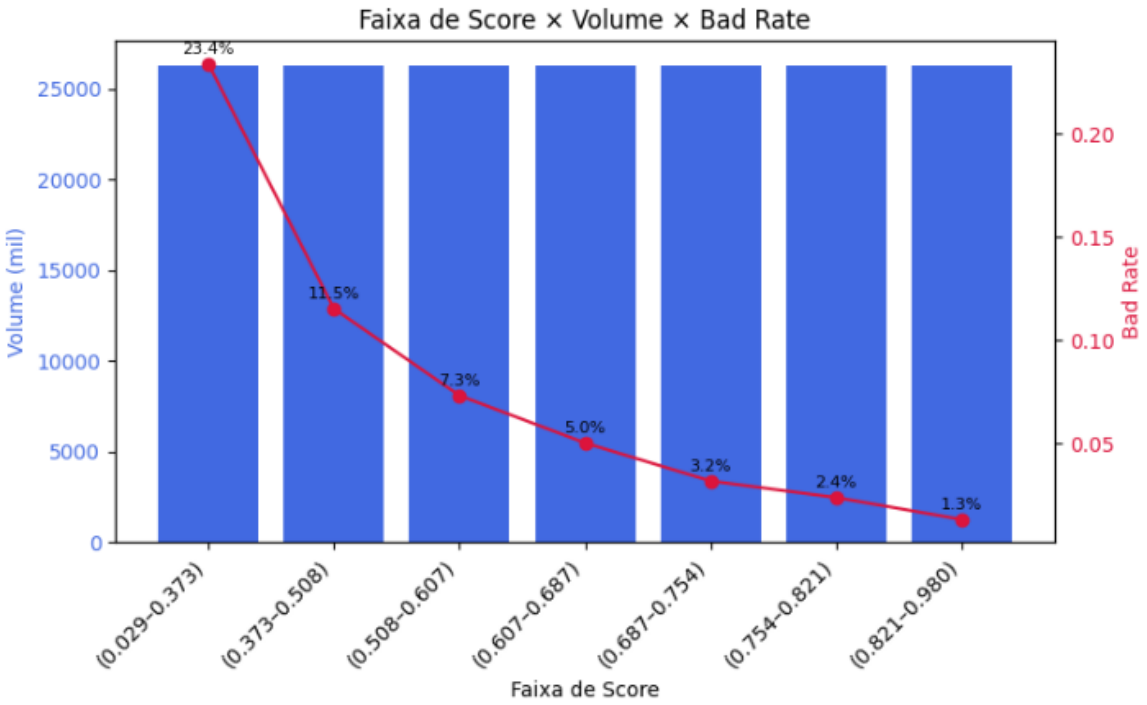
- Escores externos: quanto maior, menor o risco. Altos valores deslocam os SHAPs para a direita (↑score → menor risco). Baixos valores deslocam os SHAPs para a esquerda (↓score → maior risco).

num_VL_TOT_VL_TOT_AMT_CREDIT_SUM_DEBT_ULTIMOS_12_MESES_externo:

- Quando o valor da dívida é alto (vermelho) → SHAP negativo → menor score (maior risco).Valor baixo (azul) → impacto positivo no score.

num_CREDIT_TO_GOODS_RATIO_publico: Alta razão crédito/bens → risco maior (vermelho → SHAP negativo).Cliente com crédito menor proporcional ao bem tende a ter menor risco.

Modelo LightGBM - Performance



Faixa de Score vs. Volume vs. Inadimplência

- **Monotonicidade**: inadimplência decresce de forma ordenada entre grupos
- **Score com bom alcance**: score mínimo de 0,029 até 0,980, ampla cobertura.
- **Segmentação balanceada**: cada faixa de score cobre ~14% do público → boa granularidade.

Desempenho do Modelo

Métrica	Treino	Teste	Geral
AUC	76,36%	75,81%	76,25%
Gini	52,73%	51,63%	52,50%
KS	39,60%	38,58%	39,36%

O modelo apresenta **alto poder de discriminação**, com **performance consistente entre treino e teste**, indicando boa capacidade preditiva e baixo risco de overfitting:

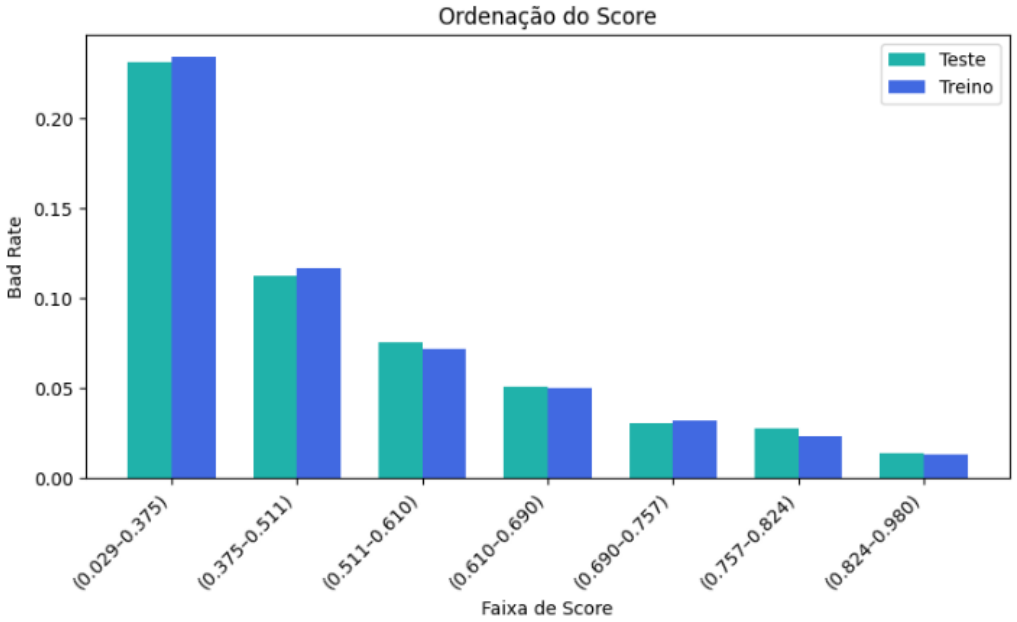
•**AUC de 75,81% no teste**, sinalizando boa separação entre bons e maus pagadores.

•**KS de 38,58% no teste**, acima do patamar de referência (>30%), demonstrando forte diferenciação entre os grupos.

•**Estabilidade entre treino e teste** (diferença de <1 p.p. nas métricas), reforçando a **robustez e generalização** do modelo.

Modelo LightGBM - Performance

CC2506



Ordenação de Score:

Os scores estão ordenados de modo que quanto maior o score, menor o risco, ou seja, a Inadimplência diminui com o aumento do score. Os valores de Bad Rate entre os dados de Treino e Teste não apresentam grandes variações.

Scorecard

gh	faixa_str	min_score	max_score	bad_rate	volume	vol_acum	% vol
1	(0,821–0,980)	0,821127	0,980132	1,33%	26.336	26.336	14%
2	(0,754–0,821)	0,754303	0,821124	2,39%	26.336	52.672	29%
3	(0,687–0,754)	0,686763	0,754298	3,19%	26.335	79.007	43%
4	(0,607–0,687)	0,607429	0,686763	5,03%	26.336	105.343	57%
5	(0,508–0,607)	0,508471	0,607429	7,33%	26.335	131.678	71%
6	(0,373–0,508)	0,373019	0,50847	11,54%	26.336	158.014	86%
7	(0,029–0,373)	0,029571	0,373017	23,40%	26.336	184.350	100%

Apetite financeiro:

- O corte de aprovação pode ser feito com base no apetite de risco (por ex., até GH4 → inadimplência máxima de 5%, cobre 57% do público.
- Aprovar até o GH 5, com 71% do público com uma Inadimplência de 7,3 % (atual Inadimplência do público de modelagem)

** Novos cortes podem ser feitos de acordo com a política

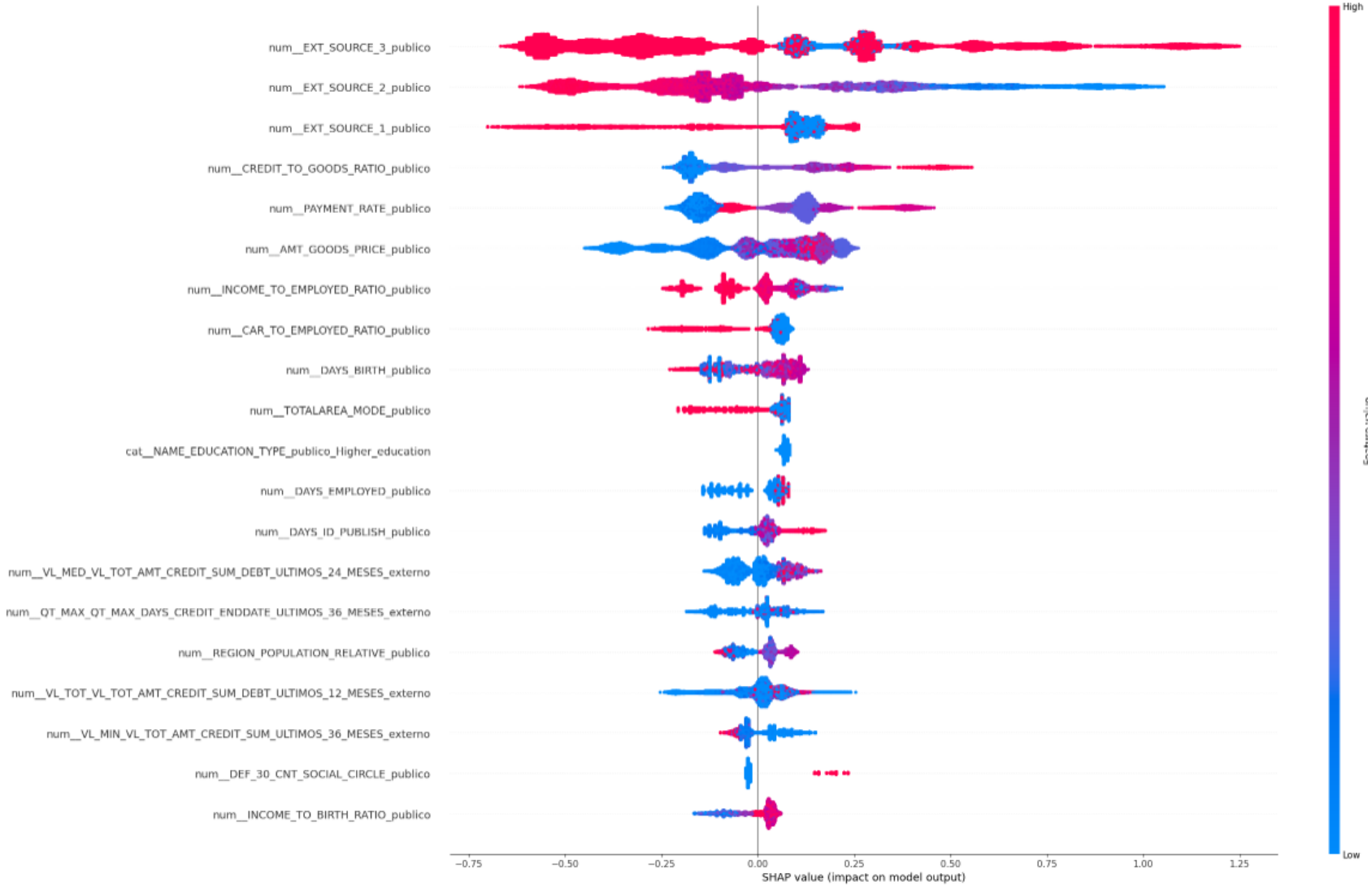
Modelo LightGBM – 26 Variáveis Selecionadas

CC2506

	Variável	% Importância	IV	KS
1	num__EXT_SOURCE_3_publico	12,5%	32,8%	25,3%
2	num__EXT_SOURCE_2_publico	10,7%	27,7%	21,2%
3	num__AMT_GOODS_PRICE_publico	8,5%	10,0%	9,2%
4	num__PAYMENT_RATE_publico	6,3%	10,5%	9,6%
5	num__VL_TOT_VL_TOT_AMT_CREDIT_SUM_DEBT_ULTIMOS_12_MESES_externo	6,3%	7,5%	13,2%
6	num__QT_MAX_QT_MAX_DAYS_CREDIT_ENDDATE_ULTIMOS_36_MESES_externo	4,7%	4,7%	8,6%
7	num__EXT_SOURCE_1_publico	4,7%	10,5%	11,7%
8	num__CREDIT_TO_GOODS_RATIO_publico	4,1%	7,7%	12,2%
9	num__REGION_POPULATION_RELATIVE_publico	3,4%	3,3%	5,8%
10	num__DAYS_EMPLOYED_publico	3,4%	9,8%	8,3%
11	num__DAYS_LAST_PHONE_CHANGE_publico	3,4%	3,8%	8,8%
12	num__DAYS_BIRTH_publico	3,1%	8,2%	11,7%
13	num__INCOME_TO_EMPLOYED_RATIO_publico	3,1%	7,6%	12,2%
14	num__VL_MED_VL_TOT_AMT_CREDIT_SUM_DEBT_ULTIMOS_24_MESES_externo	2,8%	6,5%	11,7%
15	num__TOTALAREA_MODE_publico	2,8%	3,4%	8,3%
16	num__VL_MIN_VL_TOT_AMT_CREDIT_SUM_ULTIMOS_36_MESES_externo	2,2%	3,6%	5,3%
17	num__VL_TOT_VL_TOT_AMT_CREDIT_SUM_LIMIT_ULTIMOS_24_MESES_externo	2,2%	4,7%	6,8%
18	num__INCOME_TO_BIRTH_RATIO_publico	1,9%	2,1%	5,5%
19	num__DAYS_ID_PUBLISH_publico	1,9%	3,7%	8,1%
20	num__CAR_TO_EMPLOYED_RATIO_publico	1,9%	3,5%	6,8%
21	num__QT_MIN_QT_MAX_DAYS_CREDIT_UPDATE_ULTIMOS_24_MESES_externo	1,9%	2,7%	6,1%
22	num__REGION_RATING_CLIENT_W_CITY_publico	1,9%	4,8%	6,2%
23	cat__NAME_EDUCATION_TYPE_publico_Higher_education	1,9%	3,3%	6,8%
24	num__QT_MAX_QT_MAX_DAYS_CREDIT_UPDATE_ULTIMOS_6_MESES_externo	1,6%	5,8%	11,2%
25	num__DAYS_REGISTRATION_publico	1,6%	2,6%	6,2%
26	num__DEF_30_CNT_SOCIAL_CIRCLE_publico	1,3%	1,5%	3,8%

Modelo LightGBM – Shap Values

CC2506



Explicação do comportamento das variáveis:

num_DAYS_BIRTH_publico:
Clientes mais jovens (azul) têm SHAP negativo → risco maior. Clientes mais velhos (vermelho) têm SHAP positivo → risco menor.

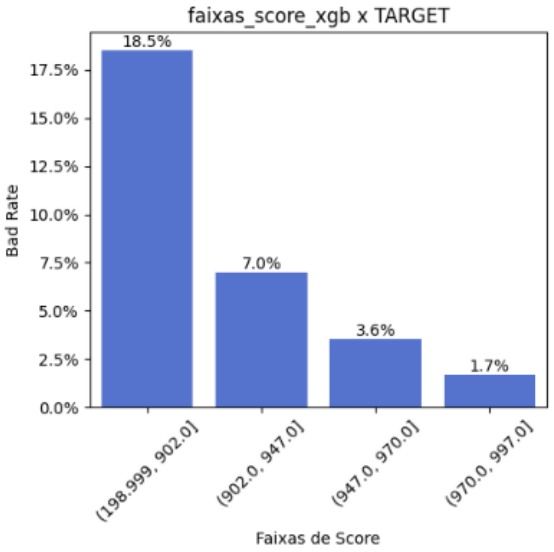
num_PAYMENT_RATE_publico:
Alta taxa de pagamento (vermelho) → desloca para a direita → menor risco. Baixa taxa → aumenta risco.

cat_NAME_EDUCATION_TYPE_publico_Higher_education e cat_NAME_INCOME_TYPE_publico_Working:
Pessoas com ensino superior ou emprego formal (flag = 1) tendem a ter SHAPs positivos → menor risco.

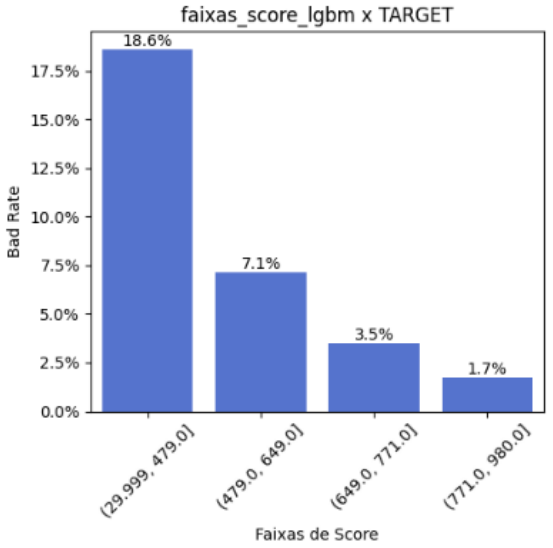
cat_OCCUPATION_TYPE_publico_Laborers:
Clientes que trabalham como operários (flag = 1) puxam SHAP negativo → maior risco.

Modelos XGBoost e LightGBM - Comparação

CC2506



Métrica	Treino	Teste	Geral
AUC	76,30%	75,54%	76,15%
Gini	52,60%	51,09%	52,30%
KS	39,66%	38,72%	39,44%



Métrica	Treino	Teste	Geral
AUC	76,36%	75,81%	76,25%
Gini	52,73%	51,63%	52,50%
KS	39,60%	38,58%	39,36%

Resumo de Performance

Desempenho dos Modelos

Modelo	# Variáveis	AUC Teste	Gini Teste	KS Teste
LGBM	26	74,84%	49,68%	37,88%
XGB	31	74,25%	48,50%	36,68%

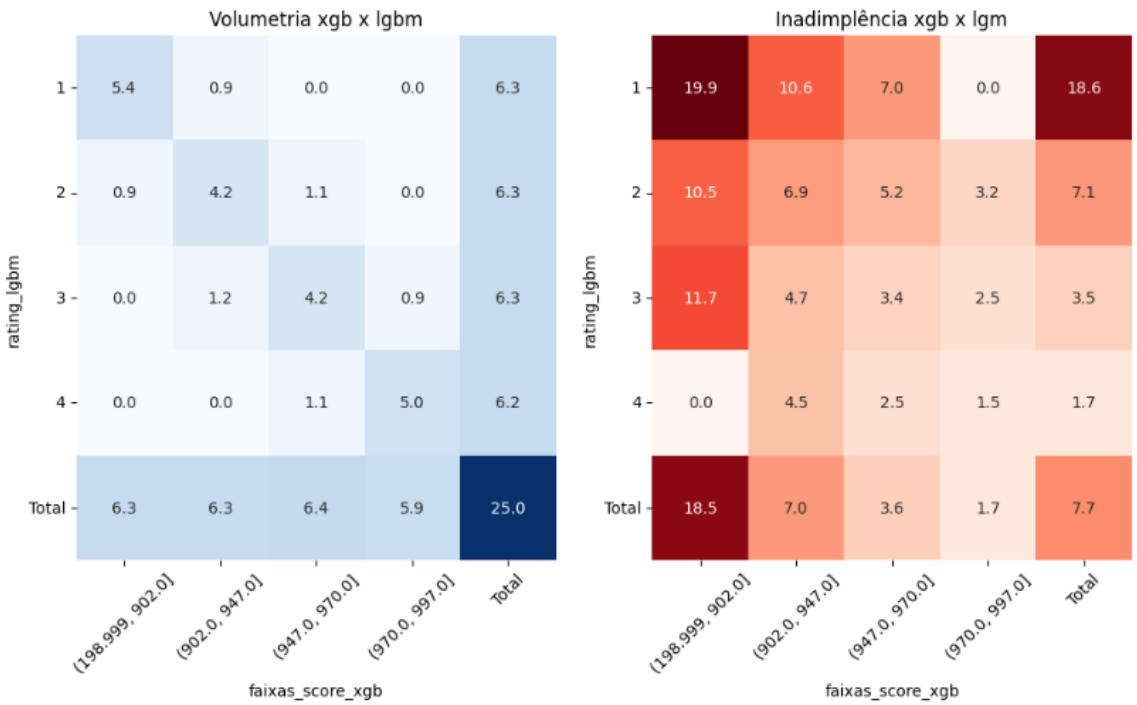
Bad Rate por Faixa

Faixa de Score	Faixa de Score LGBM	Faixa de Score XGB	LightGBM	XGBoost
Faixa 1 (pior risco)	(29.999, 479.0]	(198.999, 902.0]	18.6%	18.5%
Faixa 2	(479.0, 649.0]	(902.0, 947.0]	7.1%	7.0%
Faixa 3	(649.0, 771.0]	(947.0, 970.0]	3.5%	3.6%
Faixa 4 (melhor risco)	(771.0, 980.0]	(970.0, 997.0]	1.7%	1.7%

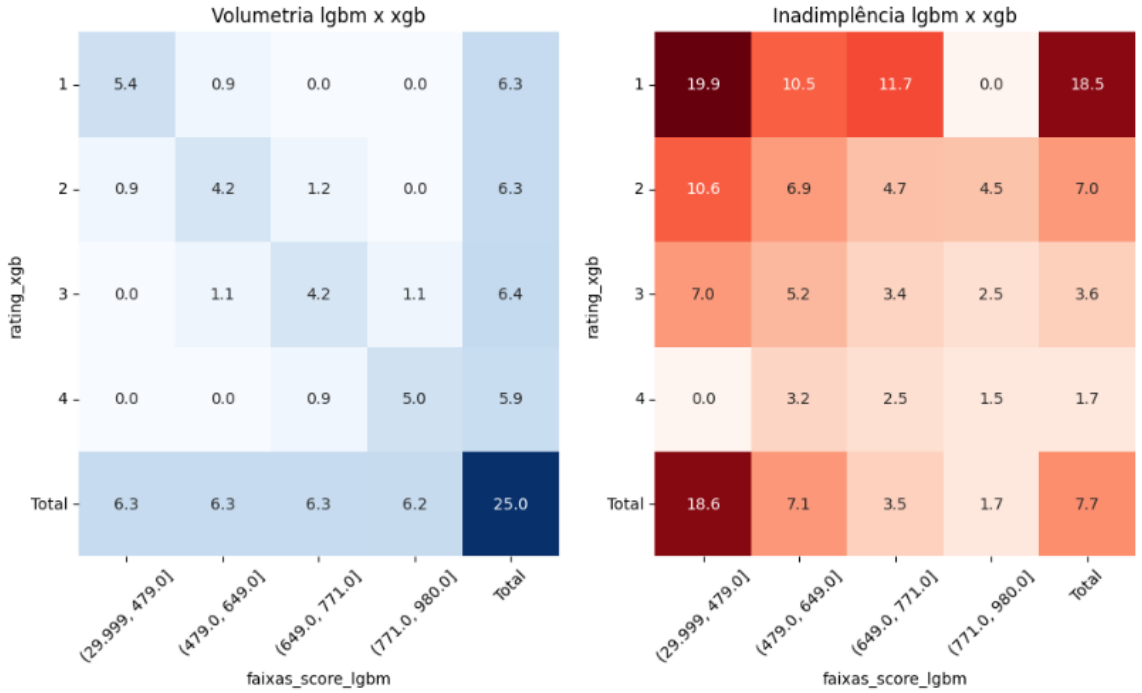
Matriz de Migração

CC2506

XGBoost x LightGBM

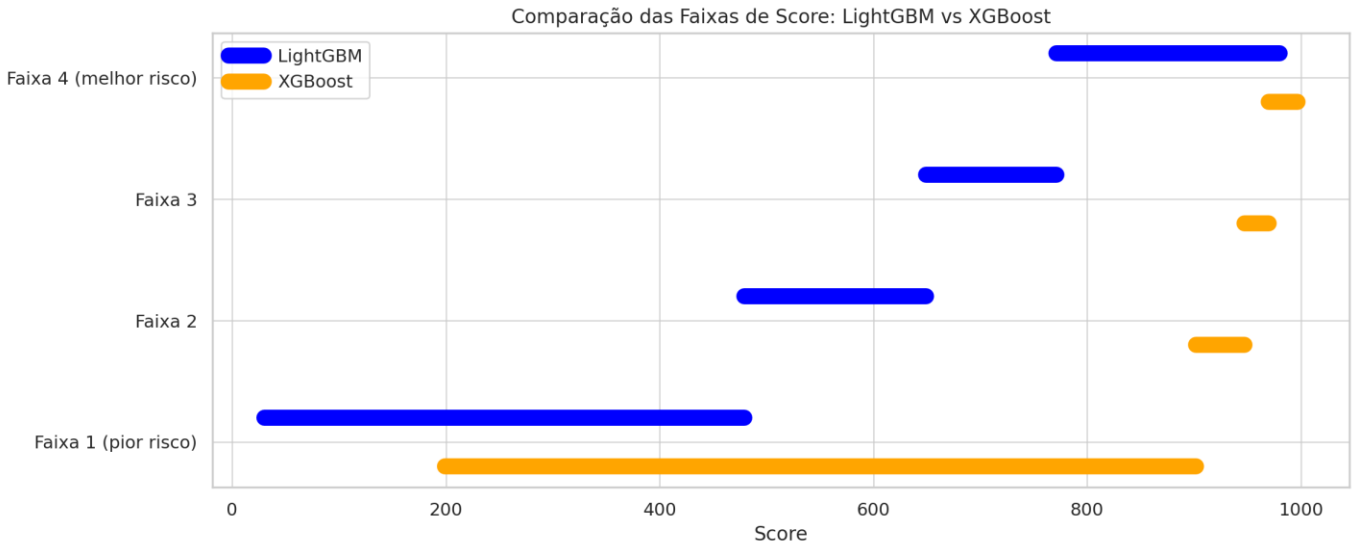


LightGBM x XGBoost



Ambos os modelos conseguem ordenar bem o risco de crédito: as taxas de inadimplência caem progressivamente da Faixa 1 (score menor) para a Faixa 4 (score maior). Mesmo com intervalos de score diferentes, os resultados finais são muito parecidos, indicando que ambos os modelos estão aprendendo padrões semelhantes de risco.

Comparação das faixas de Score



Faixas de Score

- LightGBM tem uma escala mais ampla e bem distribuída, o que facilita segmentações, regras de corte, políticas de crédito e explicações regulatórias.
- XGBoost, apesar de ter menos amplitude, entrega performance similar — mas pode exigir ajustes para calibrar melhor os cortes de score.
- É visível que o LightGBM utiliza uma faixa de score mais ampla e distribuída, enquanto o XGBoost tem faixas mais comprimidas e começa em valores mais altos.
- O LightGBM pode oferecer melhor separação entre os perfis de risco, mesmo com número menor de variáveis.

Análise de valor incremental

Uma abordagem comum em risco de crédito — especialmente quando se trabalha com variáveis externas ou custosas, como scores — consiste em seguir uma estratégia incremental, conforme os passos abaixo:

1. Nova seleção de variáveis internas: realizar uma nova seleção sem considerar os scores externos, garantindo que o modelo base reflita apenas as informações de variáveis explicativas.

2.Treinamento do modelo base: treinar um modelo apenas com as variáveis explicativas selecionadas.

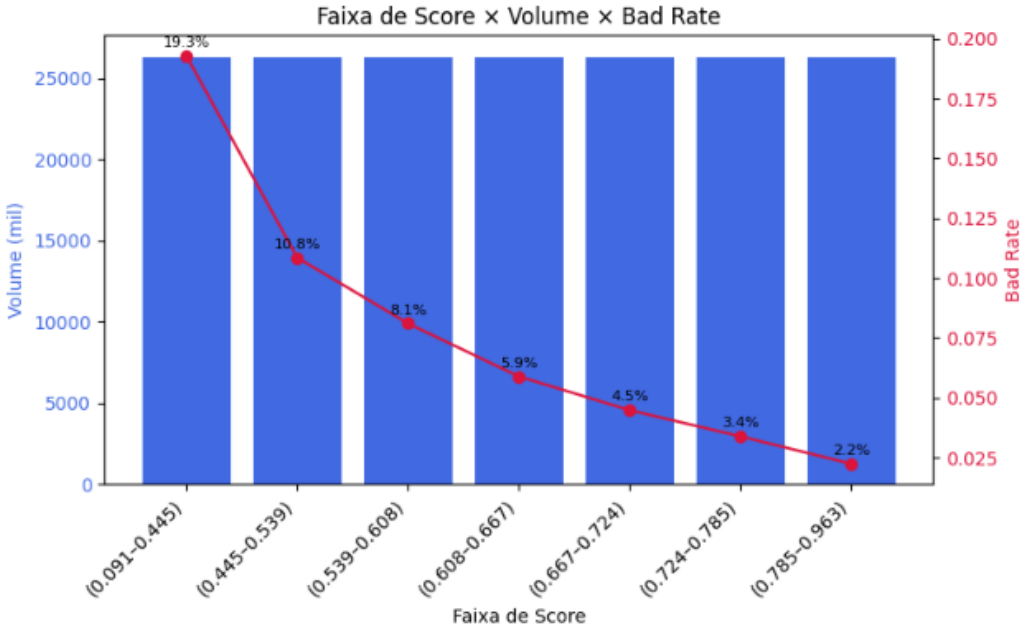
3. Análise incremental dos scores: adicionar individualmente cada score externo ou suas combinações ao modelo base, avaliando o ganho de performance em cada caso.

4. Avaliação por camadas de score (Blend): testar a inclusão dos scores em diferentes camadas — por exemplo, adicionando um score por vez, em ordens distintas — para entender o valor marginal de cada camada adicional.

Essa abordagem ajuda a quantificar o valor agregado de cada score e a justificar, de forma técnica e estratégica, o uso de variáveis custosas no modelo de crédito.

Modelo sem scores LightGBM - Performance

CC2506



Faixa de Score vs. Volume vs. Inadimplência

- Apesar de manter a **monotonicidade da inadimplência entre faixas**, o modelo tem **níveis mais altos de risco em cada faixa**, comparado aos modelos anteriores com scores.
- **Score com bom alcance**: score mínimo de 0,091 até 0,963, ampla cobertura.
- **Segmentação balanceada**: cada faixa de score cobre ~14% do público → boa granularidade.

Desempenho do Modelo

Métrica	Treino	Teste	Geral
AUC	70,72%	69,85%	70,55%
Gini	41,44%	39,71%	41,09%
KS	30,58%	30,64%	30,50%

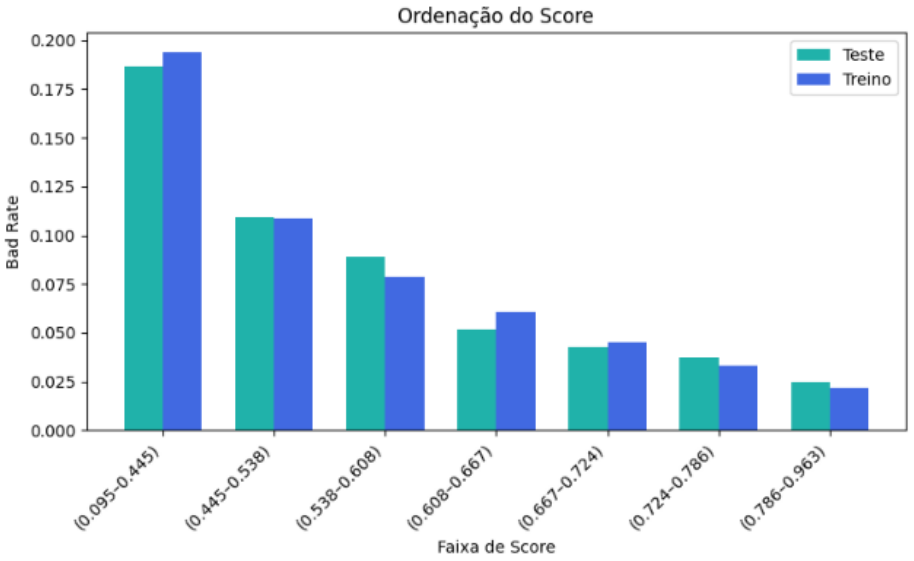
O modelo continua estável entre treino e teste (diferenças pequenas), mas a qualidade preditiva geral é inferior aos modelos anteriores. Comparando somente com o modelo Lightgbm, vamos chama-lo de Lightgbm Core temos:

- **Queda de ~6 pontos percentuais em AUC** no teste (de 75,81% para 69,85%)

- **O KS caiu de 38,72% para 30,64%**, um recuo de mais de ~8 p.p., sugerindo menor capacidade de separar bons e maus pagadores.

Modelo sem scores LightGBM - Performance

CC2506



Ordenação de Score:

Os scores estão ordenados de modo que quanto maior o score, menor o risco, ou seja, a Inadimplência diminui com o aumento do score. Os valores de Bad Rate entre os dados de Treino e Teste não apresentam grandes variações.

Scorecard

gh	faixa_score	min_score	max_score	bad_rate	volume	vol_acum	% vol
1	(0,785–0,963)	0,78543	0,963105	2,24%	26.336	26.336	14%
2	(0,724–0,785)	0,723688	0,78543	3,38%	26.336	52.672	29%
3	(0,667–0,724)	0,667339	0,723686	4,48%	26.335	79.007	43%
4	(0,608–0,667)	0,608059	0,667337	5,89%	26.336	105.343	57%
5	(0,539–0,608)	0,538701	0,608053	8,10%	26.335	131.678	71%
6	(0,445–0,539)	0,444656	0,538699	10,85%	26.336	158.014	86%
7	(0,091–0,445)	0,091888	0,444648	19,28%	26.336	184.350	100%

Apetite financeiro:

- Para atingir o mesmo patamar de inadimplência (até 5%), o modelo sem scores externos só pode aprovar até GH3 (43% da base), enquanto o modelo anterior permitia aprovar até GH4 (57% da base).
- Isso representa 14 p.p. a menos de volume aprovado com o mesmo nível de risco.

Modelo sem scores LightGBM – 21 Variáveis Selecionadas

CC2506

	Variável	Importância	IV	KS
1	num__DAYS_REGISTRATION_publico	7,8%	2,7%	6,1%
2	num__AMT_GOODS_PRICE_publico	7,7%	10,1%	9,2%
3	num__VL_TOT_VL_TOT_AMT_CREDIT_SUM_DEBT_ULTIMOS_12_MESES_externo	7,1%	7,8%	13,3%
4	num__DAYS_BIRTH_publico	6,5%	8,0%	11,5%
5	num__VL_MED_VL_TOT_AMT_CREDIT_SUM_DEBT_ULTIMOS_24_MESES_externo	6,5%	6,8%	11,9%
6	num__PAYMENT_RATE_publico	6,4%	10,2%	9,4%
7	num__DAYS_EMPLOYED_publico	6,3%	9,8%	8,5%
8	num__INCOME_TO_EMPLOYED_RATIO_publico	6,3%	7,5%	12,0%
9	num__REGION_POPULATION_RELATIVE_publico	5,7%	3,6%	5,8%
10	num__CREDIT_TO_GOODS_RATIO_publico	4,7%	7,6%	12,2%
11	num__DAYS_LAST_PHONE_CHANGE_publico	4,7%	4,2%	9,1%
12	num__QT_MIN_QT_MAX_DAYS_CREDIT_UPDATE_ULTIMOS_24_MESES_externo	4,4%	2,8%	6,3%
13	num__DAYS_ID_PUBLISH_publico	3,8%	3,6%	8,3%
14	num__VL_TOT_VL_TOT_AMT_CREDIT_SUM_LIMIT_ULTIMOS_24_MESES_externo	3,6%	4,6%	6,9%
15	num__INCOME_TO_BIRTH_RATIO_publico	3,6%	2,3%	5,4%
16	num__VL_MIN_VL_TOT_AMT_CREDIT_SUM_ULTIMOS_36_MESES_externo	3,6%	3,6%	5,2%
17	num__TOTALAREA_MODE_publico	2,6%	3,8%	8,7%
18	num__CAR_TO_EMPLOYED_RATIO_publico	2,3%	3,6%	7,0%
19	num__VL_TOT_VL_TOT_AMT_CREDIT_MAX_OVERDUE_ULTIMOS_24_MESES_externo	2,3%	4,6%	6,8%
20	num__QT_MAX_QT_MAX_DAYS_CREDIT_UPDATE_ULTIMOS_6_MESES_externo	2,0%	5,9%	11,1%
21	num__REGION_RATING_CLIENT_W_CITY_publico	2,0%	5,0%	6,2%

Análise Incremental – Scores como variáveis explicativas

CC2506

	Scores Usados	AUC Treino (%)	AUC Teste (%)	Gini Treino (%)	Gini Teste (%)	KS Treino (%)	KS Teste (%)
1	**EXT_SOURCE_1 + EXT_SOURCE_2 + EXT_SOURCE_3**	75,67	75,05	51,35	50,11	38,78	37,90
2	**EXT_SOURCE_2 + EXT_SOURCE_3**	75,34	74,72	50,68	49,44	38,33	37,59
3	**EXT_SOURCE_1 + EXT_SOURCE_3**	74,34	73,67	48,68	47,33	36,71	35,12
4	**EXT_SOURCE_1 + EXT_SOURCE_2**	73,86	73,04	47,73	46,09	35,38	35,20
5	**EXT_SOURCE_3**	73,37	72,80	46,73	45,60	35,01	34,39
6	**EXT_SOURCE_2**	73,20	72,40	46,40	44,79	34,47	33,70
7	**EXT_SOURCE_1**	72,02	71,24	44,03	42,49	32,84	32,26
8	**Sem scores**	70,72	69,85	41,44	39,71	30,58	30,64

Valor Incremental dos Scores como variáveis internas no modelo

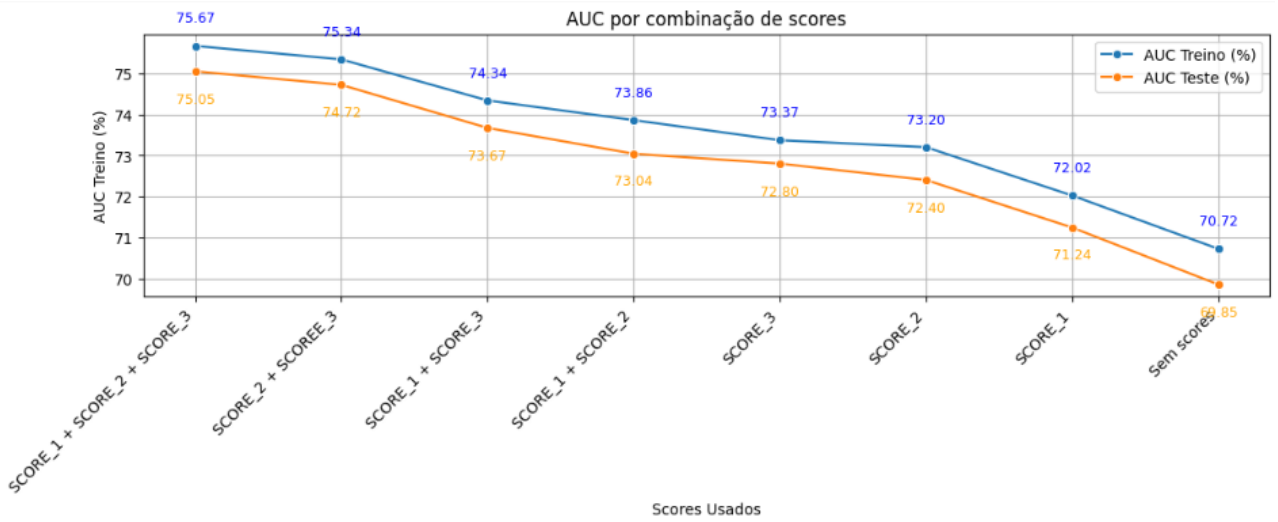
- AUC Teste sobe de 69,85% para 75,05%, Gini sobe de 39,71% para 50,11% e KS sobe de 30,64% para 37,90% com os 3 scores.

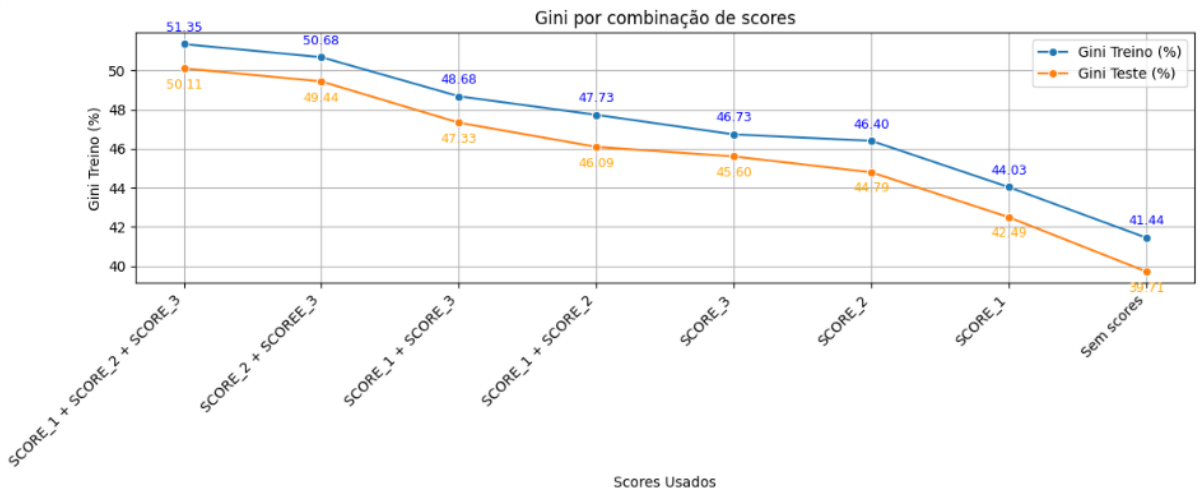
Valor individual de cada score

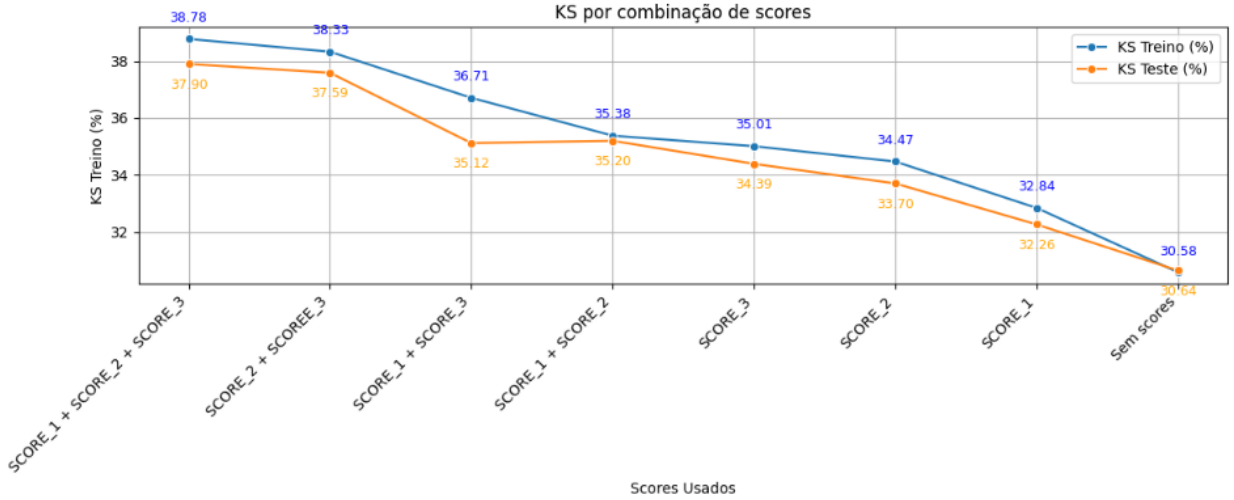
- Todos os scores, individualmente, contribuem, mas em diferentes intensidades:
- EXT_SOURCE_3 é o mais forte sozinho: AUC Teste 72,80%
- EXT_SOURCE_2 vem logo depois: AUC Teste 72,40%
- EXT_SOURCE_1 tem menor impacto isolado: AUC Teste 71,24%.

Análise Incremental – Scores como variáveis explicativas

CC2506







- Os scores externos EXT_SOURCE_1, EXT_SOURCE_2 e EXT_SOURCE_3 agregam valor real ao modelo.
- A combinação ideal envolve os três, mas EXT_SOURCE_2 e EXT_SOURCE_3 já oferecem praticamente todo o ganho quando usados juntos.

Blend de scores por camada

CC2506

Scores_usados	AUC Treino (%)	AUC Teste (%)	Gini Treino (%)	Gini Teste (%)	KS Treino (%)	KS Teste (%)
prob0 + 3 camadas (EXT_SOURCE_1 + EXT_SOURCE_2 + EXT_SOURCE_3)	75,59	74,84	51,18	49,68	38,24	37,88
prob0 + 2 camadas (EXT_SOURCE_2 + EXT_SOURCE_3)	75,01	74,20	50,03	48,41	37,34	36,66
prob0 + 2 camadas (EXT_SOURCE_1 + EXT_SOURCE_3)	74,03	73,24	48,05	46,49	35,73	34,97
prob0 + 2 camadas (EXT_SOURCE_1 + EXT_SOURCE_2)	73,69	72,59	47,38	45,17	35,23	34,46
prob0 + 1 camada (EXT_SOURCE_3)	73,15	72,32	46,3	44,63	34,33	33,29
prob0 + 1 camada (EXT_SOURCE_2)	72,99	71,87	45,97	43,73	34,11	33,19
prob0 + 1 camada (EXT_SOURCE_1)	71,93	70,79	43,86	41,58	32,42	31,51
Somente prob0	70,84	69,75	41,68	39,5	30,63	30,37

Valor Incremental dos Scores

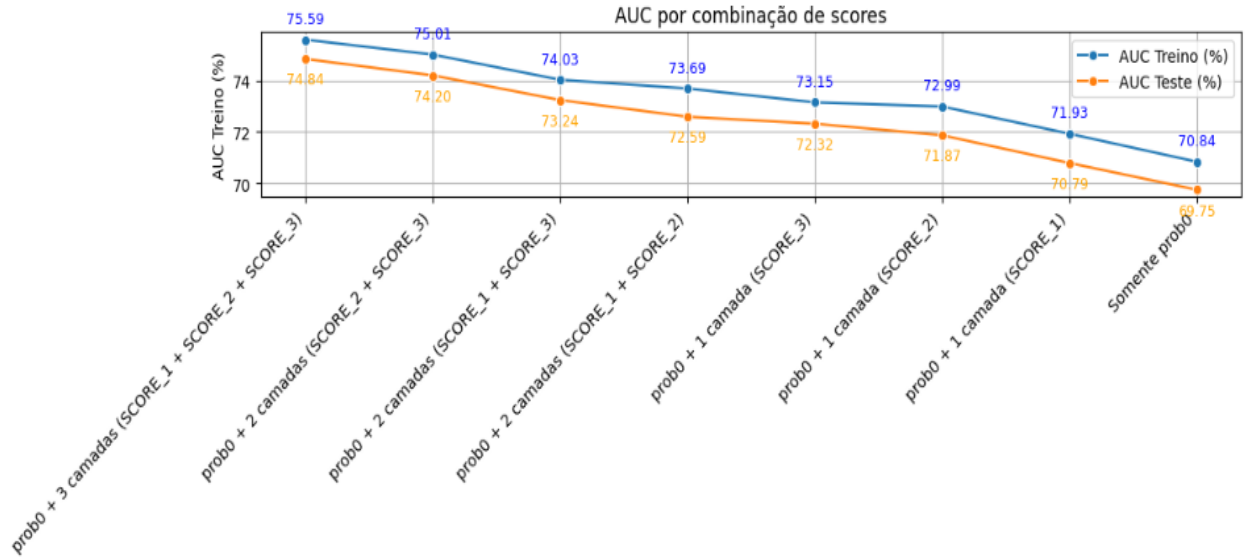
- AUC Teste sobe de 69,75% para 74,84%, Gini sobe de 39,50% para 49,68% e KS sobe de 30,37% para 37,88% com os 3 scores

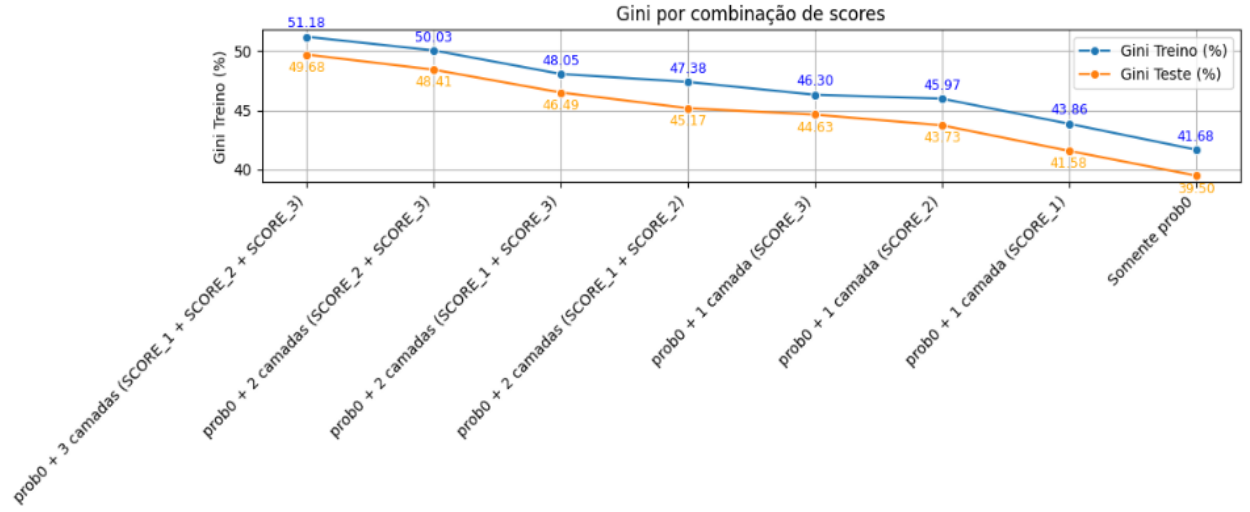
Valor individual de cada score

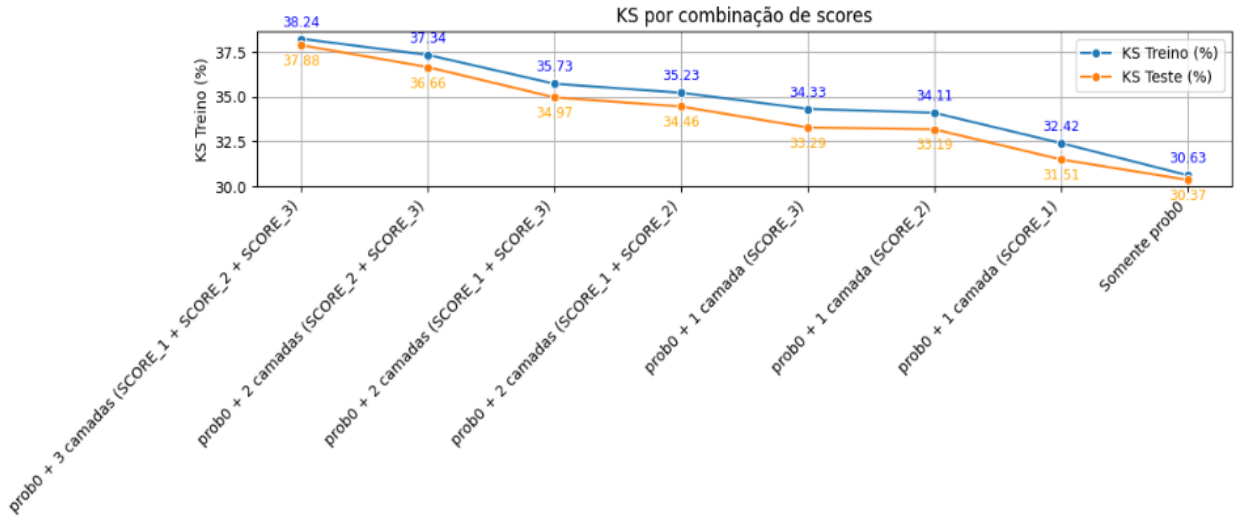
- Todos os scores, individualmente, contribuem, mas em diferentes intensidades:
- EXT_SOURCE_3 é o mais forte sozinho: AUC Teste 72,32%, EXT_SOURCE_2 vem logo depois: AUC Teste 71;87%EXT_SOURCE_1 tem menor impacto isolado: AUC Teste 70,79%.

Blend de scores por camada

CC2506



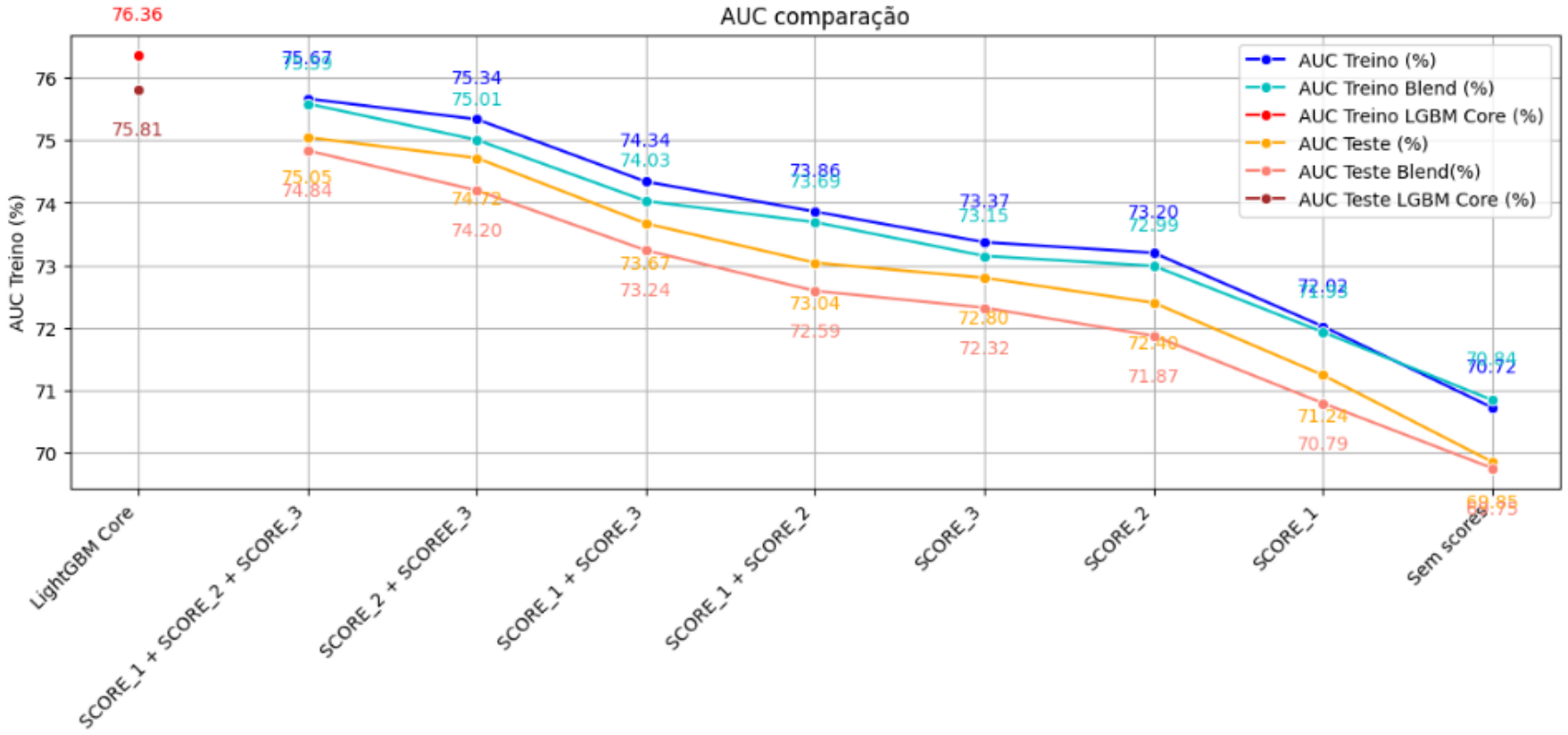




- Os scores externos EXT_SOURCE_1, EXT_SOURCE_2 e EXT_SOURCE_3 agregam valor real ao modelo.
- A combinação ideal envolve os três, mas EXT_SOURCE_2 e EXT_SOURCE_3 já oferecem praticamente todo o ganho quando usados juntos.
- **Como agora os scores externos não interagem mais internamente vemos uma perda maior em performance no teste**

Comparativo entre os modelos LightGBM

CC2506



Comparativo

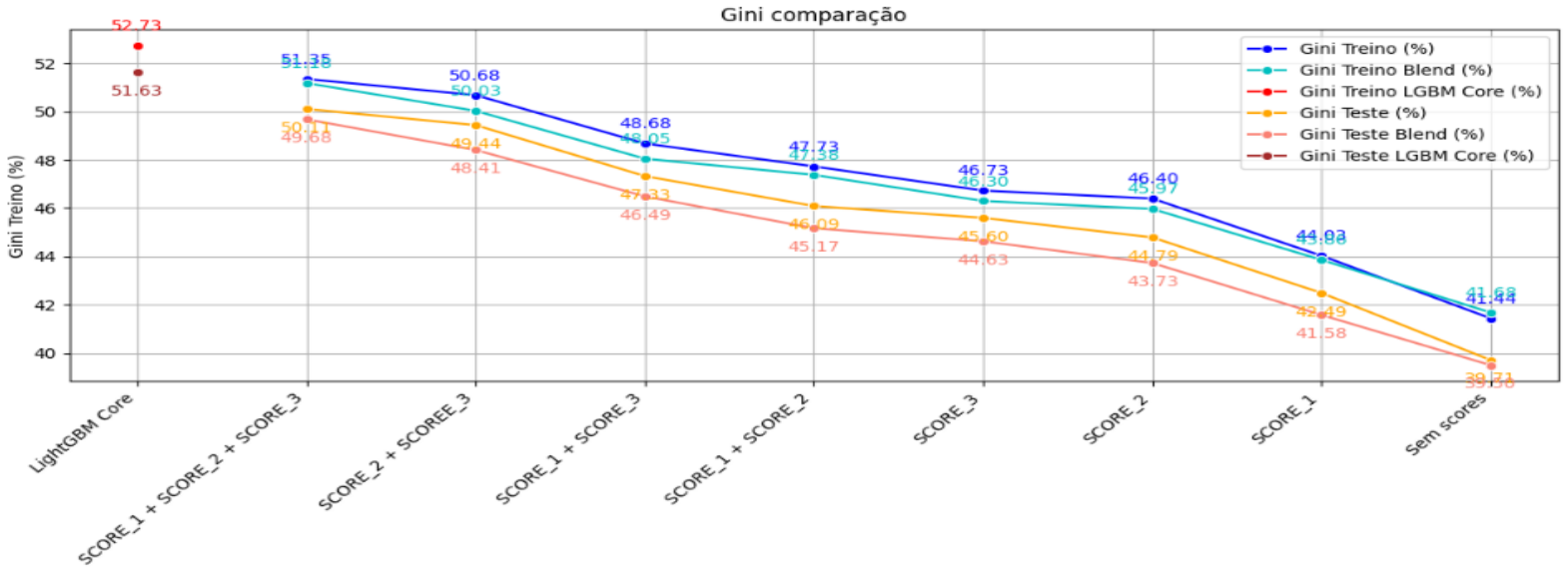
O LGBM Core tem o maior valor de AUC teste 75,81, seguido pelo que adiciona os 3 scores externos diretamente com AUC teste 75,05 e o blend com 3 camadas com AUC teste de 74,84 (com -0,21 p.p. Em relação ao modelo anterior.

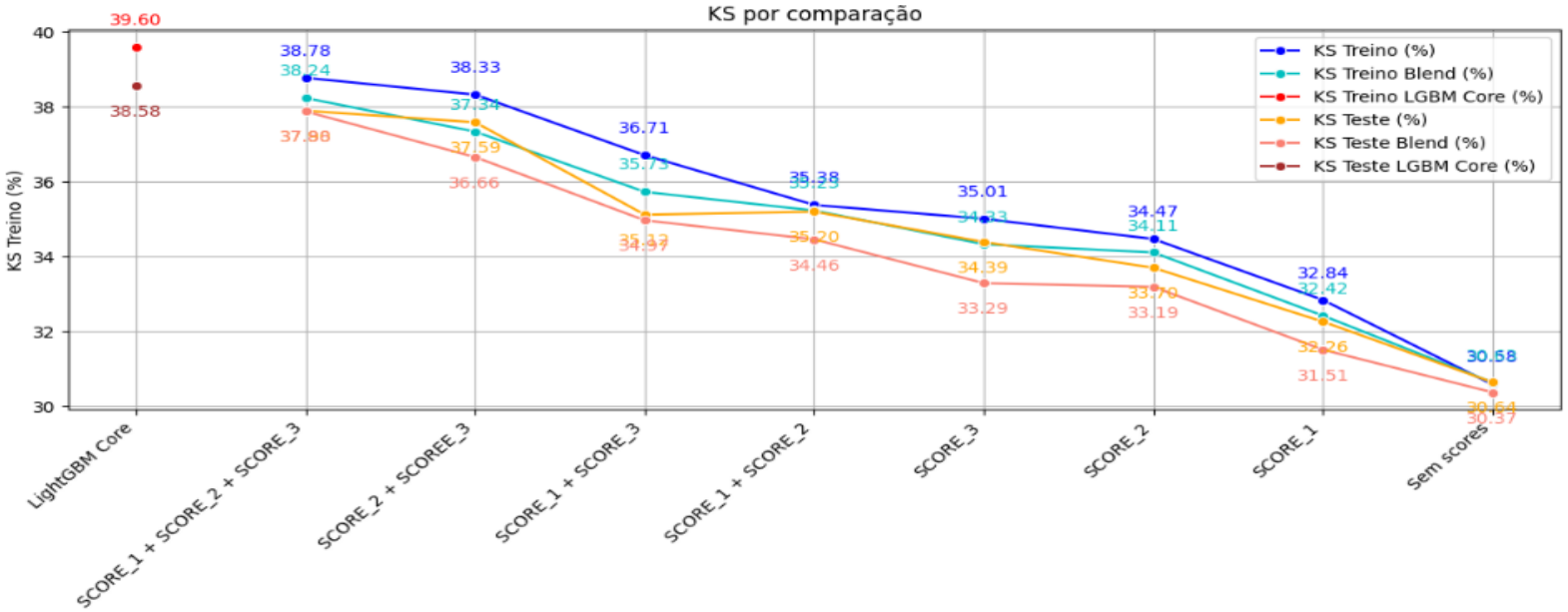
A leve queda na AUC do modelo em camadas ocorre porque parte do poder preditivo das variáveis internas é "comprimido" em um único score, e o modelo que vem depois não consegue aprender as interações ricas entre os scores externos e as variáveis internas originais

*LGBM Core: modelo LightGBM com 26 variáveis, tendo os 3 scores como variáveis internas do modelo

Comparativo entre os modelos LightGBM

CC2506





Comparativo

- O modelo base somente com variáveis internas apresenta uma AUC = 70,72%, Gini 41,44 e KS = 30,58%.
- Adicionar os 3 scores externos diretamente melhora o AUC Teste em 5,2 p.p., Gini em 10,4 p.p. e KS em 7,3 p.p. Esses ganhos indicam que os scores externos são altamente informativos e complementares
- Modelo Blend (Camadas de Score): Desempenho muito semelhante ao modelo anterior, com uma pequena queda (ex: KS de 37,90 → 37,88).
- LGBM Core (variáveis + scores juntos): Melhor desempenho geral. O ganho em relação ao modelo Base é significativo (+6 p.p. em AUC Teste), e mesmo comparado ao modelo com os scores adicionados simples ou via blend, ele ainda tem vantagem

Ganho Incremental

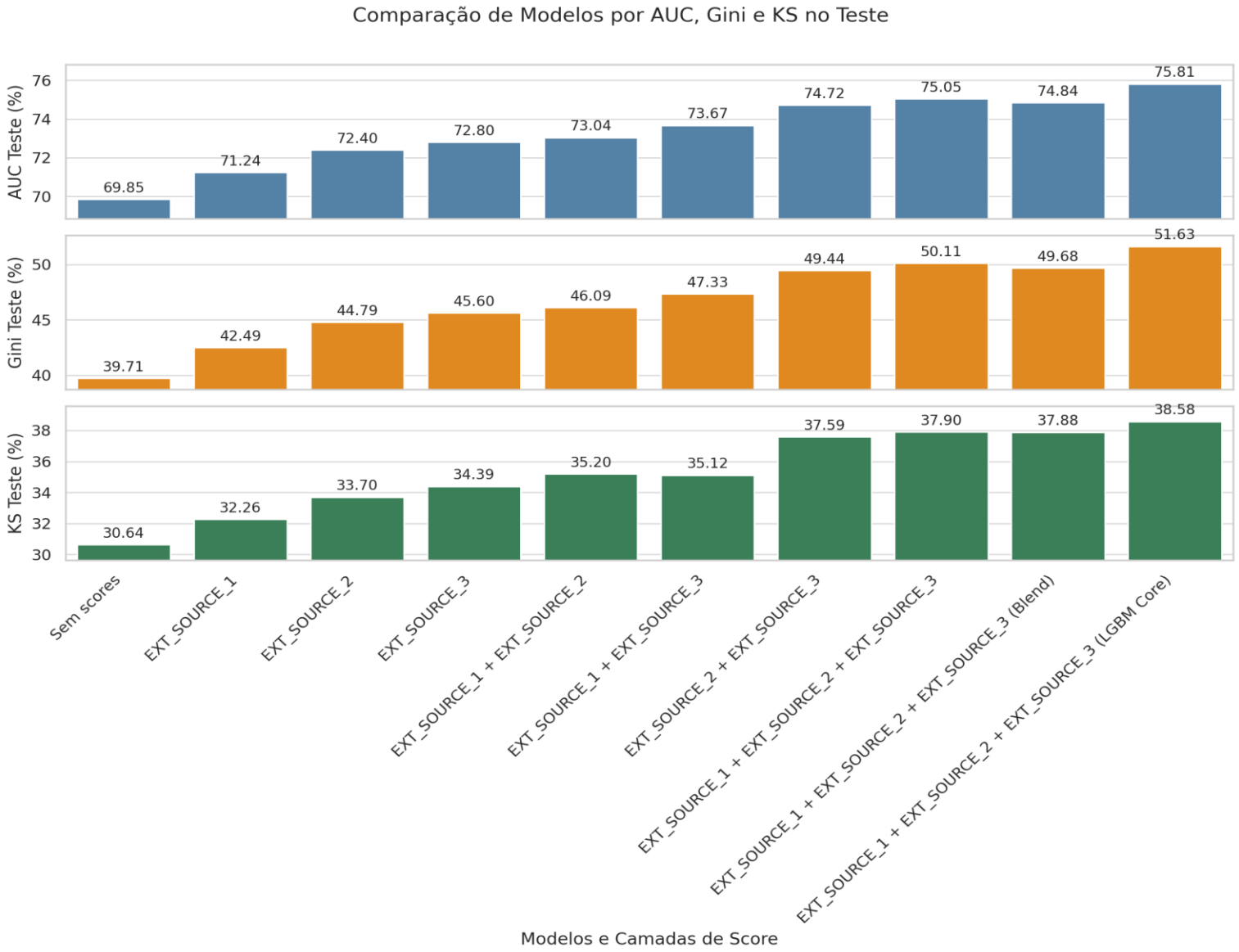


Gráfico resumo mostrando de forma clara o ganho incremental ao incluir scores externos no modelo LightGBM de previsão de inadimplência.

Resumo do Ganho Incremental

CC2506

1. Modelo base (Sem scores)

AUC: 69,85% | Gini: 39,71% | KS: 30,64%

- Esse é o modelo apenas com as 21 variáveis internas. Serve como referência para avaliar o valor agregado pelos scores externos.

2. Análise incremental com inclusão de scores

- À medida que scores externos são adicionados, as métricas melhoram continuamente.
- O maior ganho ocorre com a inclusão de EXT_SOURCE_3, seguido por EXT_SOURCE_2 e depois EXT_SOURCE_1.
- EXT_SOURCE_1 sozinho eleva a AUC para 71,24% (+1,39pp).
- EXT_SOURCE_2 + EXT_SOURCE_3 sobem para 74,72% (+4,87pp).
- Todos os 3 scores juntos chegam a 75,05% (+5,2pp), com Gini de 50,11% e KS de 37,90%.

3. Modelos com camadas (Blend)

- 21 variáveis internas foram usados para gerar um score base (prob0), que depois foi combinado em camadas com os scores externos.
- Apesar de ligeiramente inferior ao modelo com os três scores diretos, o modelo Blend ainda entrega:
  AUC: 74,84% | Gini: 49,68% | KS: 37,88%
- Isso mostra que o modelo em camadas consegue capturar boa parte do sinal, mesmo com transformação em score intermediário.

4. Modelo LGBM Core

- Esse modelo incorpora as 26 variáveis (23 internas + 3 scores) diretamente, considerando interações complexas.
- Ele entrega os melhores resultados absolutos:
  AUC: 75,81% | Gini: 51,63% | KS: 38,58%

Conclusões e Recomendações

- **Melhor modelo**: LGBM Core com 26 variáveis diretas. Ele extrai o máximo valor dos 3 scores e suas interações com as variáveis internas.
- **Modelo mais interpretável/modular**: O Blend é interessante se você quer separar responsabilidades por camada ou usar os scores em fluxos diferentes (por exemplo, operacional vs. analítico). Aplicar camadas com scores externos em decisões de maior risco ou valor. Também possibilita atualizar cada camada separadamente sem retrabalhar o modelo todo.
- **Ganho real:** O uso de scores externos melhora a AUC em quase 6 pontos percentuais e o KS em quase 8pp, o que é um ganho significativo em modelagem de risco de crédito.
- **Interações:** O LGBM Core possivelmente está aprendendo interações não lineares entre os scores e variáveis internas — por isso seu desempenho superior.

Marina Cavalca, Cientista de Dados do Vitalis Bank

Resolução completa do projeto em:

https://github.com/MaCavalca/Modelo_Credito_Vitalis_Bank

Vitalis Bank

“No Vitalis Bank, acreditamos que Vital para os nossos clientes é proporcionar saúde financeira para uma vida com mais propósito, liberdade e paz.”

**Projeto fictício para fins de portifólio*