Predição do desempenho do alunos nas avaliações finais

Alicja Mazur

São Paulo, 1 de Dezembro 2020

Agenda

- O problema
- Dados disponíveis, transformação e processamento, correlação
- Dataset final
- Pipeline de Machine Learning
- Desempenho do modelo
- Desempenho do outros modelos testados
- Análise de predições do modelo
- Possíveis melhoras

O problema

Como identificar o aluno que pode falhar na prova final?

Predição de falha na prova final com modelo de classificação Machine Learning.

- Análise exploratória e processamento de dados disponíveis.
- Treinamento de vários algoritmos para comparação de desempenho.
- Minimização do falsos negativos: alunos que não passaram na prova que foram classificados como “passaram”.
- Análise e explicação de decisões do modelo final.

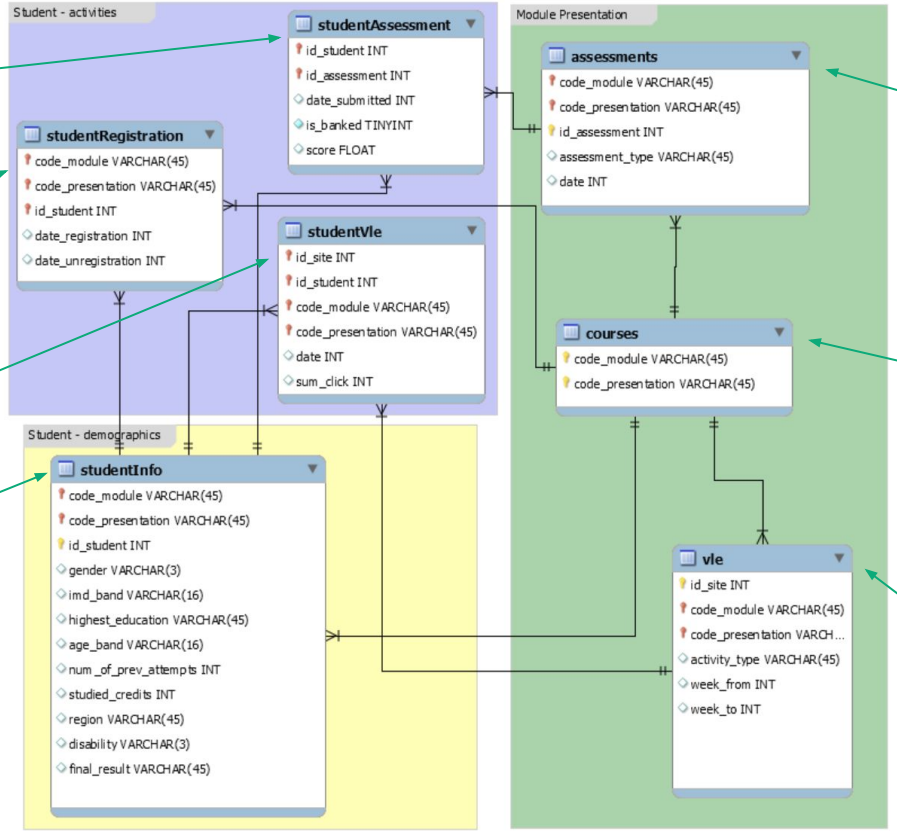
Bases de dados disponíveis

Desempenho nas provas

Matrícula

Atividade na plataforma online

Dados pessoais



Provas

Cursos

Materiais na plataforma online

Variáveis disponíveis

Matrícula:
- Código do curso
- Código do módulo de curso
- Número de identificação do aluno
- Data de matricula
- Data de cancelamento

Dados pessoais:
- Código do curso
- Código do módulo de curso
- Número de identificação do aluno
- Gênero,
- Categoria do Indice IMD
- Educação
- Idade
- Número de tentativas previas
- Número de créditos
- Região
- Deficiência
- Nota final

Desempenho nas provas:
- Número de identificação do aluno
- Número de identificação da prova
- Nota transferida?
- Nota

Provas:
- Código do curso
- Código do módulo
- Número de identificação da prova
- Tipo de avaliação
- Data

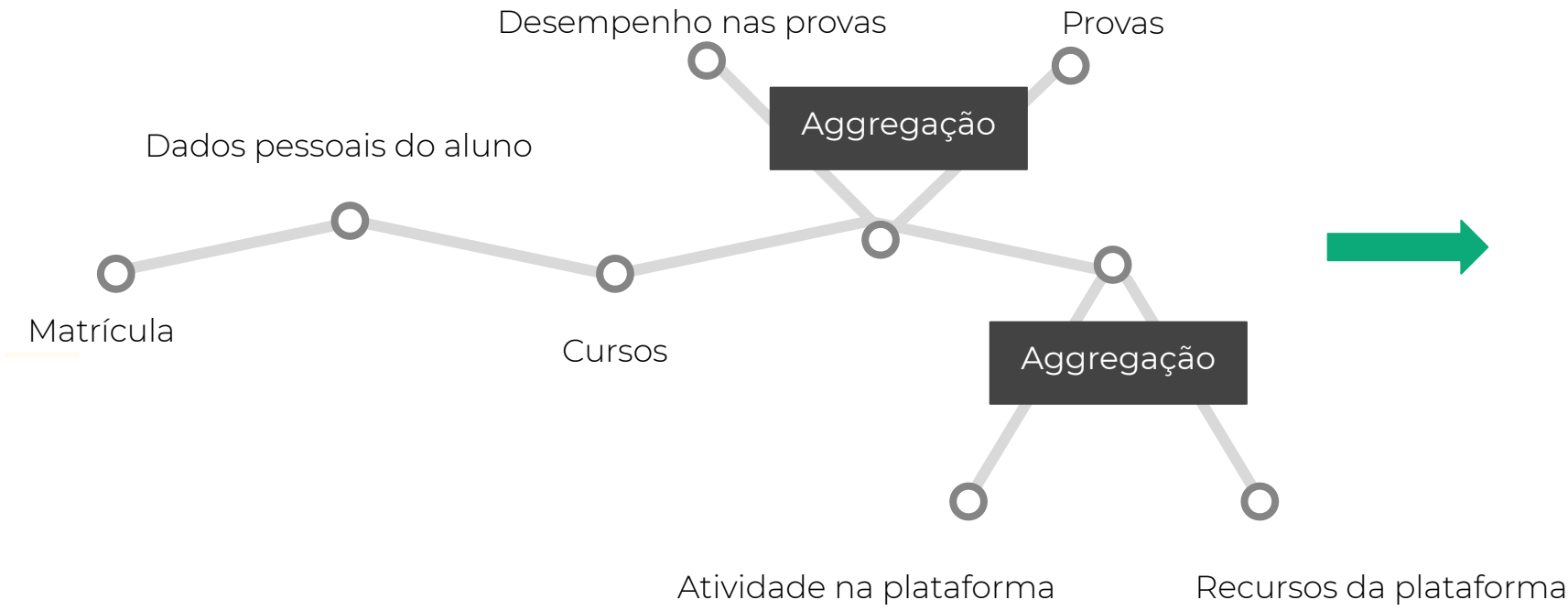
Cursos:
- Código do curso
- Código do módulo

Recursos da plataforma online:
- Número de identificação da página
- Código de curso
- Código de módulo
- Tipo de atividade
- Semana desde qual recurso vai ser usado
- Semana até qual recurso vai ser usado

Atividade do aluno na plataforma online:
- Número de identificação da página
- Número de identificação do aluno
- Código de curso
- Código de módulo
- Data
- Soma de clicks

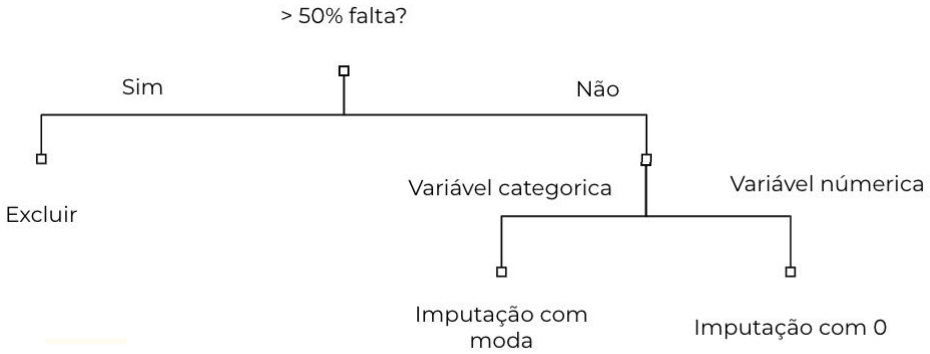
Transformação de dados - JOINs

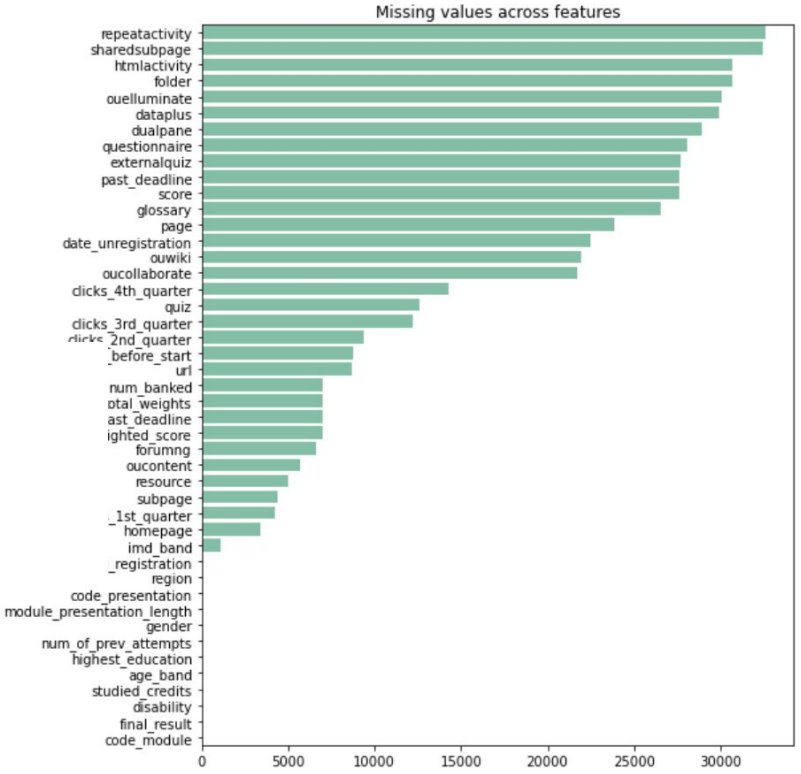
ID: **inscrição** do aluno



Processamento de dados

Grande parte dos features possuem valores faltantes.

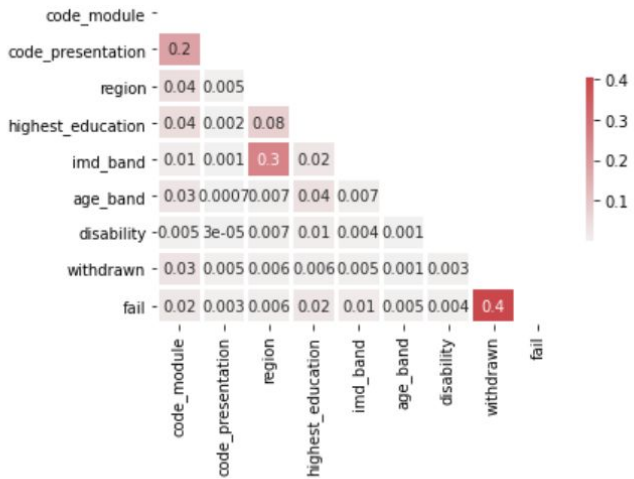


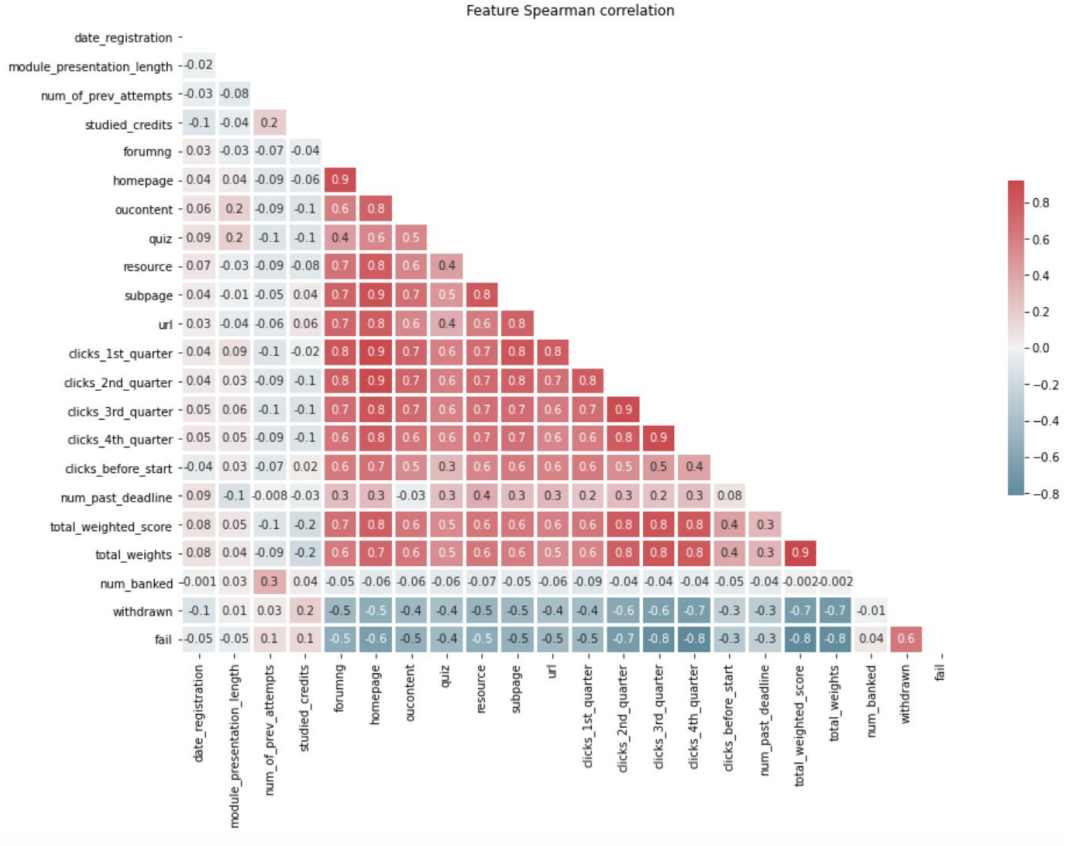


Valores numéricos faltantes se devem à ausência de registros das provas e atividades do aluno. Foi assumida a falta de atividade na plataforma online e ausência na prova de aluno.

Correlação entre as variáveis

Forte correlação entre a variável dependente e as variáveis independentes que apresentam cancelamento da matrícula, nota total nas provas, número de provas feitas, atividade na plataforma online





Dataset final

Informações sobre o curso

code_module	code_presentation	date_registration	module_presentation_length
AAA	2013J	-159.00	268
AAA	2013J	-53.00	268
AAA	2013J	-92.00	268
AAA	2013J	-52.00	268

Desempenho acumulado na provas

num_past_deadline	total_weighted_score	total_weights	num_banked	withdrawn
0.00	82.40	100.00	0.00	0
2.00	65.40	100.00	0.00	0
0.00	0.00	0.00	0.00	1
0.00	76.30	100.00	0.00	0

Dados pessoais do aluno

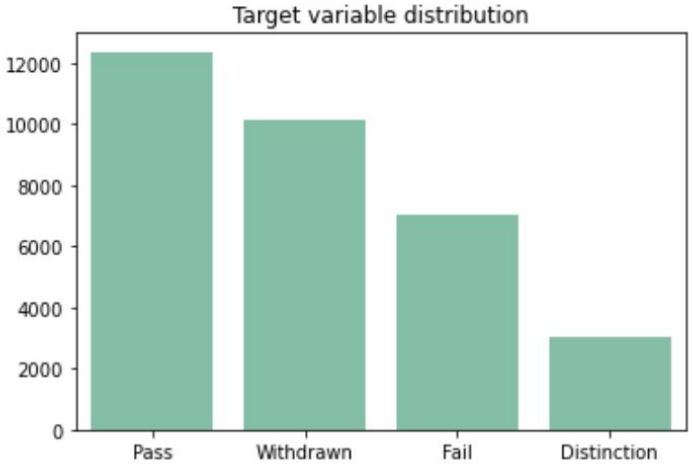
highest_education	imd_band	age_band	num_of_prev_attempts	studied_credits	disability
HE Qualification	90-100%	55<=	0	240	N
HE Qualification	20-30%	35-55	0	60	N
A Level or Equivalent	30-40%	35-55	0	60	Y
A Level or Equivalent	50-60%	35-55	0	60	N

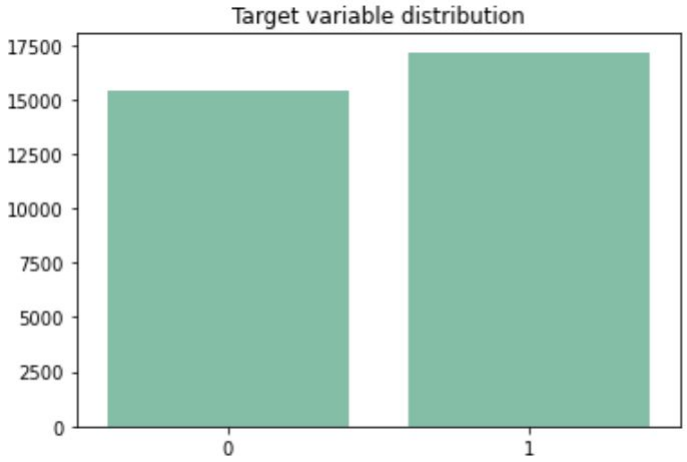
Atividade acumulada na plataforma

forumng	homepage	oucontent	quiz	resource	subpage	url	clicks_1st_quarter	clicks_2nd_quarter	clicks_3rd_quarter	clicks_4th_quarter
193.00	138.00	553.00	0.00	13.00	32.00	5.00	447.00	165.00	38.00	186.00
417.00	324.00	537.00	0.00	12.00	87.00	48.00	508.00	225.00	360.00	127.00
126.00	59.00	66.00	0.00	4.00	22.00	4.00	179.00	0.00	0.00	0.00
634.00	432.00	836.00	0.00	19.00	144.00	90.00	674.00	539.00	457.00	319.00

Classificação binária

Transformação da variável dependente: problema de **classificação binária**.





Pipeline de Machine Learning

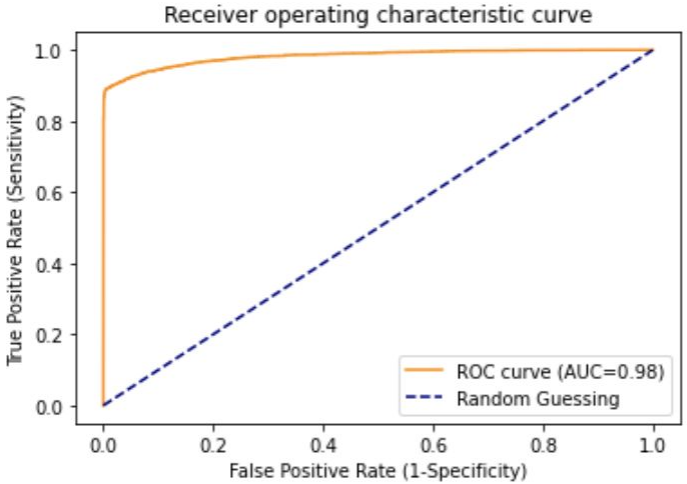
Pre-processamento variáveis numéricas	Pre-processamento variáveis categoricas	Seleção de features
RobustScaler(), adequado para dados esparso com *outliers*.	OneHotEncoder() **27 → 52 variáveis**	SelectFromModel() usando estimador XGBClassifier() em todos os experimentos (alguns algoritmos não suportam esse método). **52 → 26 variáveis**

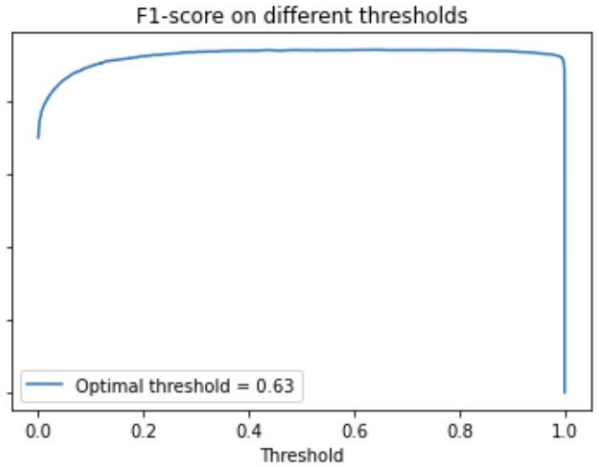
Pipeline de Machine Learning

Treinamento	Cross-validação	Análise de métricas
Modelo final: **XGBClassifier** alto performance, adequado para os dados esparsos, com *outliers*. **+Experimentos** com: Regressão Logistica Support Vector Machine K-Nearest Neighbours Random Forest	5 folds	Escolha de *threshold* baseada nos **F1-scores**: bom equilibrio entre os falsos negativos e falsos positivos. A escolha baseada no Recall gera o maior número (2000%) de falsos positivos.

Desempenho do modelo no conjunto de teste

fit_time	score_time	val_accuracy	val_roc_auc	val_precision	val_recall	val_f1score	test_accuracy	test_recall	test_precision	test_auc	threshold	test_f1score
4.18	0.21	0.94	0.98	0.97	0.91	0.94	0.94	0.90	0.98	0.98	0.63	0.94
4.16	0.18	0.94	0.98	0.96	0.92	0.94	0.94	0.90	0.98	0.98	0.63	0.94
3.76	0.21	0.94	0.98	0.96	0.91	0.94	0.94	0.90	0.98	0.98	0.63	0.94





Modelo **robusto** com Recall satisfatório.

Matrix de confusão

True Negative: 3026

Alunos que não passaram na prova e foram classificados como “não passaram”.

False Positive: 51

Alunos que não passaram na prova e foram classificados como “passaram”.

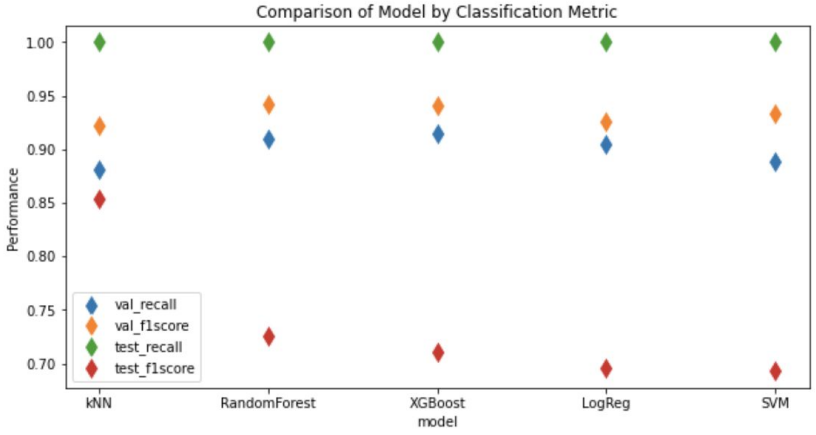
False Negative: 340

Alunos que passaram na prova e foram classificados como “passaram”.

True Positive: 3093

Alunos que passaram na prova e foram classificados como “passaram’.

Comparação de desempenho de outros modelos



Os algoritmos baseados em **árvores de decisão** acabaram sendo os de **melhor performance.**

val_accuracy	val_roc_auc	val_precision	val_recall	val_f1score	model	test_accuracy	test_recall	test_precision	test_auc	threshold	test_f1score
0.94	0.98	0.97	0.91	0.94	XGBoost	0.94	0.90	0.98	0.98	0.59	0.94
0.94	0.98	0.97	0.91	0.94	RandomForest	0.94	0.91	0.98	0.98	0.51	0.94
0.93	0.97	0.98	0.89	0.93	SVM	0.93	0.90	0.96	0.97	0.45	0.93
0.92	0.98	0.95	0.91	0.93	LogReg	0.93	0.90	0.95	0.98	0.52	0.93
0.92	0.96	0.97	0.88	0.92	kNN	0.92	0.92	0.91	0.96	0.40	0.92

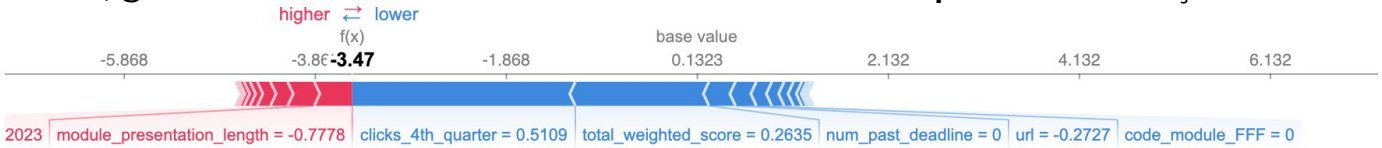
Análise de predições com Shapley values

2 amostras aleatórias

code_module	code_presentation	date_registration	module_presentation_length	highest_education	imd_band	age_band	num_of_prev_attempts	studied_credits	disability
GGG	2014B	-17.00	241	A Level or Equivalent	0-10%	0-35	0	30	N
EEE	2014B	-45.00	241	Lower Than A Level	40-50%	35-55	0	60	N

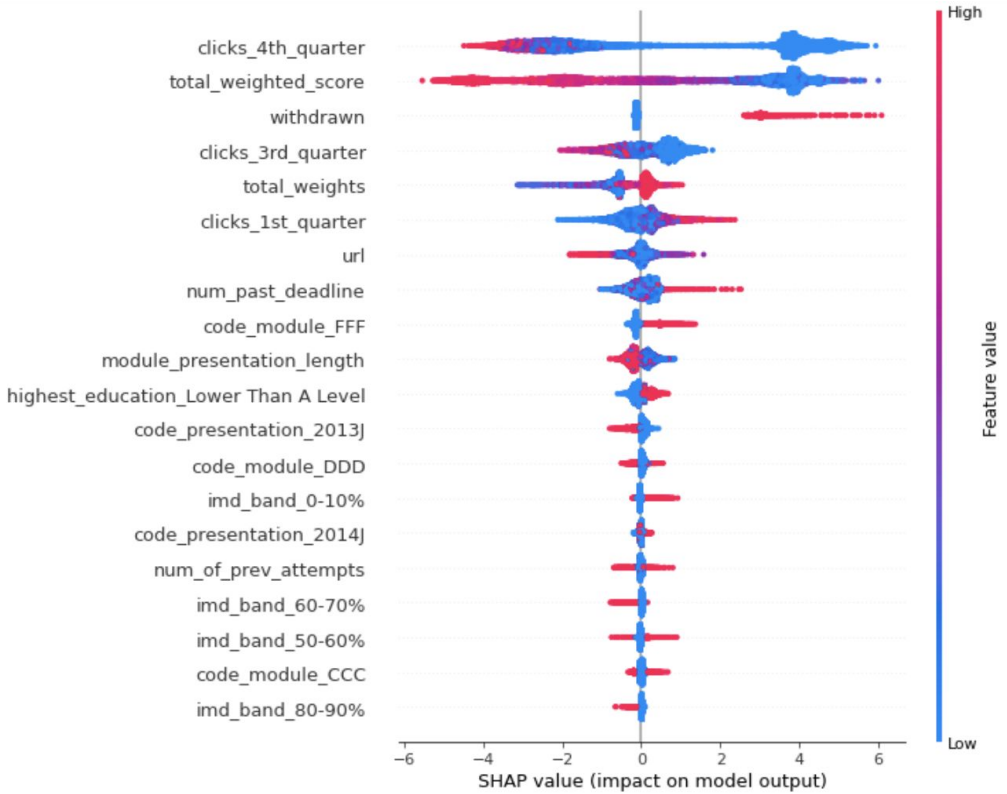
forumng	homepage	oucontent	quiz	resource	subpage	url	clicks_1st_quarter	clicks_2nd_quarter	clicks_3rd_quarter	clicks_4th_quarter	num_past_deadline	total_weighted_score	total_weights	num_banked	withdrawn
366.00	204.00	263.00	100.00	25.00	43.00	0.00	537.00	171.00	84.00	154.00	0.00	65.77	100.00	0.00	0
149.00	196.00	823.00	111.00	6.00	23.00	44.00	810.00	197.00	421.00	0.00	4.00	69.16	100.00	0.00	1

True negative (“passou”). **A favor de “passou”:** alta soma de notas, grande atividade no quarta parte do curso, grande número de URL clicks. **A favor de “não passou”:** duração do curso.



True positive (“falhou”). **A favor de “passou”** : alta soma de notas, grande atividade no terceira parte do curso, grande número de clicks nas urls. **A favor de “não passou”:** baixa atividade na quarta parte do curso, cancelamento da matrícula, alta atividade no primeira parte do curso.

Análise de predições com Shapley values



⬆⬆ abs(SHAP VALUE)
=
Maior impacto na predição do modelo

Possíveis melhorias

- Diminuir o numéro de variáveis categóricas através de *binning*.
- Análise de erro de falsos negativos e transformação posterior das variáveis preditivas.
- Tuning de hiperparâmetros com *random search* ou otimização bayesiana para aumentar a precisão do modelo.
- Considerar o custo verdadeiro do falso negativos e falso positivo e escolher o *threshold* baseado na métrica adequada.