ISAG EXECUTIVE ACADEMY

Apresentação Projeto Final

MODELO PREVISÃO PAGAMENTOS EM ATRASO

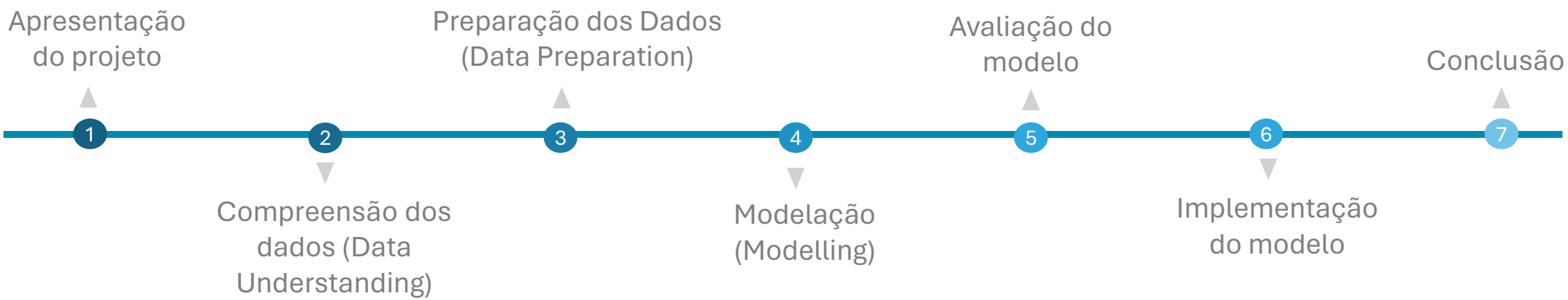
Porto, 3 de maio de 2024

Realizado por: Sara Oliveira da Silva
Orientado por: Rita Faria e Ivo Nogueira

Pós Graduação Data Science e Business Inteligence

Agenda

Metodologia CRISP-DM



Apresentação do Projeto

Motivação

O atraso no pagamento de faturas é um desafio comum e transversal a muitas empresas

Pode causar problemas de liquidez e impactar o desempenho financeiros das organizações.

Capacidade de identificar os clientes mais propensos a atrasarem o pagamento das faturas

Apresentação do Projeto

Objetivo

SOLUÇÃO

Desenvolver um modelo para prever os pagamentos em atraso.

Capacidade de identificar os clientes mais propensos a atrasarem o pagamento das faturas

Apresentação do Projeto

Objetivo | Resumo

O objetivo principal é identificar clientes que são propensos a não efetuarem ou a atrasarem o pagamento das faturas dos produtos/serviços que usufruíram.

Compreensão dos Dados

Identificação das fontes de dados

- Dataset extraído do Kaggle
- Ficheiro em formato CSV (Comma-separated values)

Compreensão dos Dados

Descrição dos dados

invoiceNumber	countryCode	customerID	PaperlessDate	InvoiceDate	DueDate	InvoiceAmount	Disputed	SettledDate	PaperlessBill	DaysToSettle	DaysLate
611365	391	0379-NEVHP	4/6/2013	1/2/2013	2/1/2013	55.94	No	1/15/2013	Paper	13	0
7900770	406	8976-AMJEO	3/3/2012	1/26/2013	2/25/2013	61.74	Yes	3/3/2013	Electronic	36	6
9231909	391	2820-XGXSB	1/26/2012	7/3/2013	8/2/2013	65.88	No	7/8/2013	Electronic	5	0
9888306	406	9322-YCTQO	4/6/2012	2/10/2013	3/12/2013	105.92	No	3/17/2013	Electronic	35	5
15752855	818	6627-ELFBK	11/26/2012	10/25/2012	11/24/2012	72.27	Yes	11/28/2012	Paper	34	4

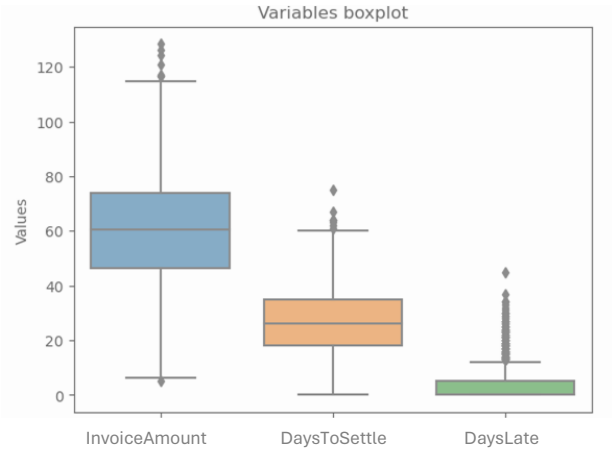
Dataset composto por

- 12 variáveis
- 2466 data points

Nome da Variável	Significado
countryCode	Código do país do cliente.
customerID	Identificação Única do Cliente.
PaperlessDate	Data em que cliente consentiu fatura digital.
invoiceNumber	Número da fatura.
InvoiceDate	Data da fatura.
DueDate	Data Vencimento da fatura.
InvoiceAmount	Valor total da fatura.
Disputed	Indica se fatura foi contestada pelo cliente. Valores Possíveis: (‘Yes’ (fatura contestada) ou ‘No’ (fatura não contestada)
SettledDate	Data em que fatura foi paga ou contestação resolvida.
PaperlessBill	Indica se cliente recebe fatura digital. Valores Possíveis: (‘Paper’ (fatura em papel) ou ‘Electronic’ (fatura eletrónica).
DaysToSettle	Número de dias que cliente demorou para efetuar pagamento.
DaysLate	Número de dias que fatura ficou em atraso em relação à data de vencimento.

Compreensão dos Dados

EDA | Missing Values e Outliers



Missing Values

0

Outliers

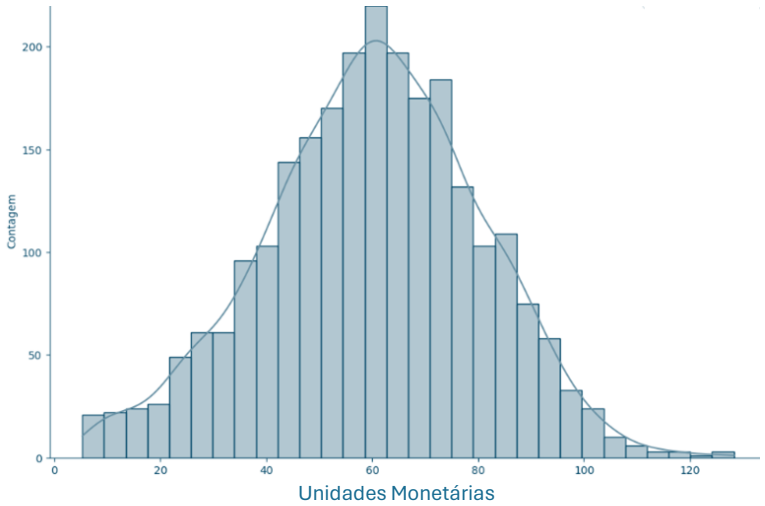
259

Devido ao contexto, não se fez tratamento dos outliers uma vez que é pertinente considerar que existem diversos cenários que influenciam o comportamento dos clientes

Compreensão dos Dados

Estatísticas Descritivas
Variáveis Numéricas

‘InvoiceAmount’



Média (μ): 59,90
Desvio Padrão (σ^2): 20,44
Mínimo (min): 5,26
Máximo (max): 128,28

Simetria: -0,1234
Curtose: -0,0952

Compreensão dos Dados

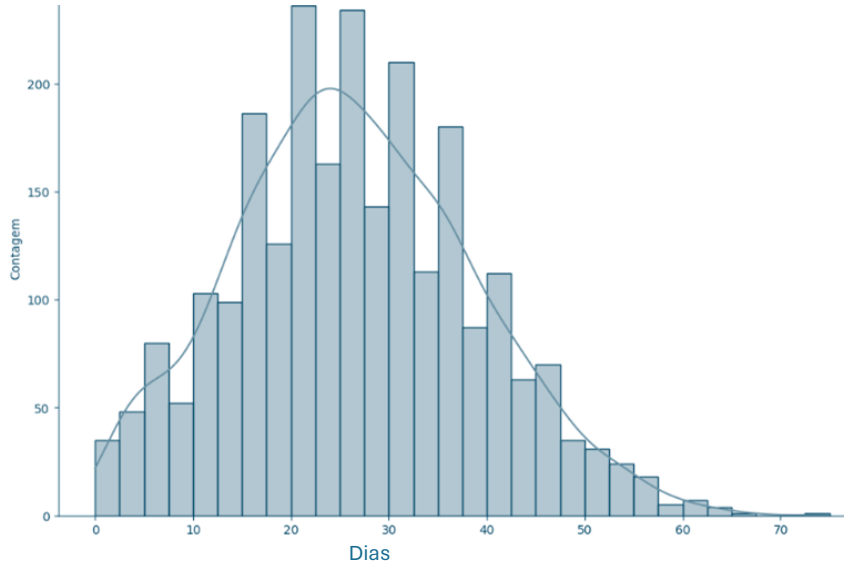
Estatísticas Descritivas
Variáveis Numéricas

‘InvoiceAmount’

Média (μ): 59,90
Desvio Padrão (σ^2): 20,44
Mínimo (min): 5,26
Máximo (max): 128,28

Simetria: -0,1234
Curtose: -0,0952

‘DaysToSettle’



Média (μ): 26
Desvio Padrão (σ^2): 12
Mínimo (min): 0
Máximo (max): 26

Simetria: 0,2415
Curtose: -0,2007

Compreensão dos Dados

Estatísticas Descritivas
Variáveis Numéricas

‘InvoiceAmount’

Média (μ): 59,90
Desvio Padrão (σ^2): 20,44
Mínimo (min): 5,26
Máximo (max): 128,28

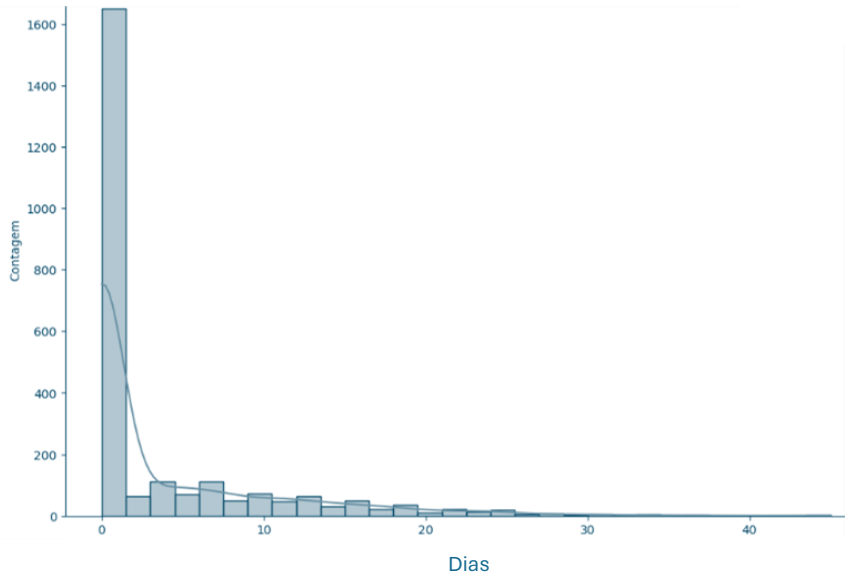
Simetria: -0,1234
Curtose: -0,0952

‘DaysToSettle’

Média (μ): 26
Desvio Padrão (σ^2): 12
Mínimo (min): 0
Máximo (max): 26

Simetria: 0,2415
Curtose: -0,2007

‘DaysLate’



Média (μ): 3
Desvio Padrão (σ^2): 6
Mínimo (min): 0
Máximo (max): 45

Simetria: 2,1589
Curtose: 4,6978

Compreensão dos Dados

Estatísticas Descritivas
Variáveis Categóricas

‘Disputed’

Fatura não contestada (0): 1905 registos

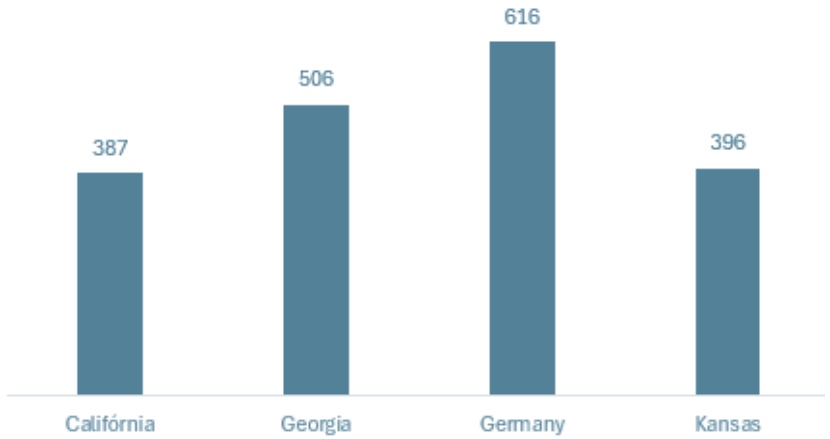
Fatura contestada (1): 561 registos

‘PaperlessBill’

Fatura em papel (0): 1263 registos

Fatura eletrónica (1): 1203 registos

‘Country_{}’



Preparação dos Dados

Transformação dummies | Normalização

Transformação Dummies

['CountryCode'] → ['Country_Germany', 'Country_California', 'Country_Georgia', 'Country_Kansas']

columns = ['PaperlessDate', 'InvoiceDate', 'DueDate', 'SettledDate']

→ For col in columns:

[col]_year,
[col]_month,
[col]_day,
[col]_weekDay

Normalização

	InvoiceAmount	Disputed	DaysToSettle	Country_California, US	Country_Georgia, US	Country_Germany	Country_Kansas, US	PaperlessBill_dummy
0	-0.193614	-0.542668	-1.090203	-0.431448	-0.508098	1.732988	-0.437384	0.975958
1	0.090259	1.842748	0.774799	-0.431448	-0.508098	-0.577038	-0.437384	-1.024634
2	0.292885	-0.542668	-1.738899	-0.431448	-0.508098	1.732988	-0.437384	-1.024634
3	2.252586	-0.542668	0.693712	-0.431448	-0.508098	-0.577038	-0.437384	-1.024634
4	0.605635	1.842748	0.612625	2.317778	-0.508098	-0.577038	-0.437384	0.975958
...	...	...	...	...	...	...	...	...
2461	0.964881	-0.542668	1.423495	-0.431448	-0.508098	1.732988	-0.437384	-1.024634
2462	-1.059425	-0.542668	-0.198246	-0.431448	-0.508098	1.732988	-0.437384	0.975958
2463	0.376579	-0.542668	-0.117159	-0.431448	1.968125	-0.577038	-0.437384	0.975958
2464	-0.329677	-0.542668	-0.441507	-0.431448	1.968125	-0.577038	-0.437384	0.975958
2465	0.428948	-0.542668	-1.009116	-0.431448	-0.508098	-0.577038	-0.437384	-1.024634

Modelling

Metodologia

Target = [‘DaysLate’]

Divisão do dataset em conjuntos de treino (80%) e teste (20%)

Testados vários algoritmos de regressão

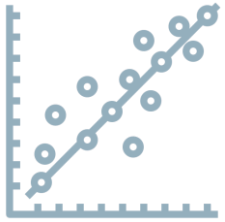
Métricas de avaliação utilizadas:

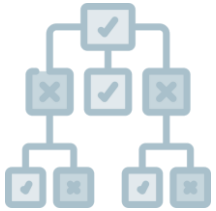
- RMSE
- R^2

Otimização Hiperparâmetros

Modelling

Tentativas





Modelo

Avaliação

Regressão Linear

LinearRegression()

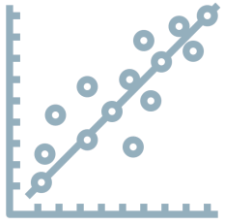
RMSE	3,3435
R^2	0,7056

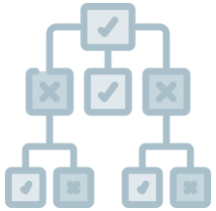
Modelling Avaliação

1 2 3 4 5 6 7

Modelling

Tentativas





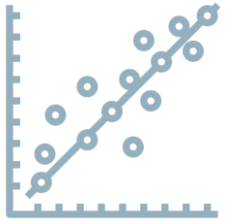
	Regressão Linear	Regressão Ridge
Modelo	LinearRegression()	Ridge (alpha = 0,5)
Avaliação	RMSE 3,3435 R^2 0,7056	RMSE 3,335 R^2 0,707

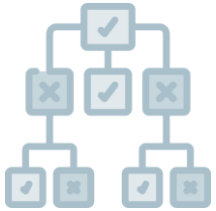
Modelling Avaliação

1 2 3 4 5 6 7

Modelling

Tentativas



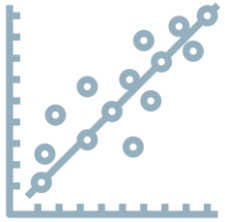


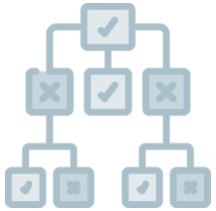
	Regressão Linear	Regressão Ridge	Regressão Lasso
Modelo	LinearRegression()	Ridge (alpha = 0,5)	Lasso (alpha = 0,5)
Avaliação	RMSE 3,3435 R^2 0,7056	RMSE 3,335 R^2 0,707	RMSE 3,310 R^2 0,711

Modelling Avaliação

Modelling

Tentativas





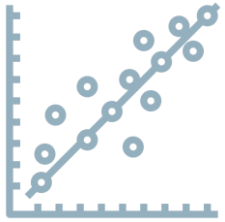
	Regressão Linear	Regressão Ridge	Regressão Lasso	Árvore Decisão
Modelo	LinearRegression()	Ridge (alpha = 0,5)	Lasso (alpha = 0,5)	DecisionTreeRegressor (max_depth=5, random_state=42)
Avaliação	RMSE 3,3435 R^2 0,7056	RMSE 3,335 R^2 0,707	RMSE 3,310 R^2 0,711	RMSE 0,597 R^2 0,991

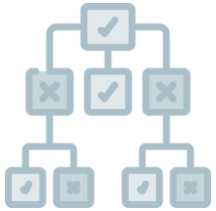
Modelling Avaliação

1 2 3 4 5 6 7

Modelling

Tentativas





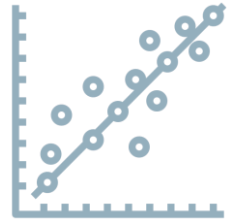
	Regressão Linear	Regressão Ridge	Regressão Lasso	Árvore Decisão	Random Forest
Modelo	LinearRegression()	Ridge (alpha = 0,5)	Lasso (alpha = 0,5)	DecisionTreeRegressor (max_depth=5, random_state=42)	RandomForestRegressor(n_estimators=100, max_depth=5, random_state=42)
Avaliação	RMSE 3,3435 R^2 0,7056	RMSE 3,335 R^2 0,707	RMSE 3,310 R^2 0,711	RMSE 0,597 R^2 0,991	RMSE 0,644 R^2 0,99

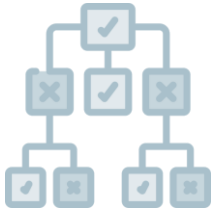
Modelling Avaliação

1 2 3 4 5 6 7

Modelling

Modelo Escolhido





Modelo

Gradient Boosting

GradientBoostingRegressor
(n_estimators=1000,
learning_rate=0.1,
max_depth=5,
random_state=42)

Avaliação

RMSE	0,524
R^2	0,994

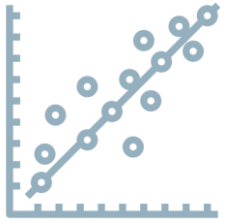
- Boa precisão do modelo em prever os valores reais
- Explica a maior parte da variância

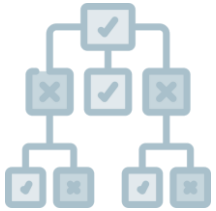
Modelling Avaliação

1 2 3 4 5 6 7

Modelling

Otimização





Modelo

Gradient Boosting

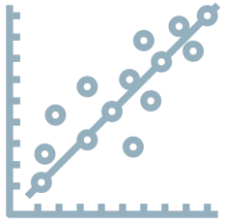
GradientBoostingRegressor
(n_estimators=1000,
learning_rate=0.1,
max_depth=5,
random_state=42)

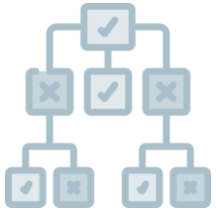
Avaliação

RMSE	0,524
R^2	0,994

Modelling

Otimização





Modelo

Avaliação

Gradient Boosting

GradientBoostingRegressor
(n_estimators=1000,
learning_rate=0.1,
max_depth=5,
random_state=42)

RMSE	0,524
R^2	0,994

RandomizedSearchCV()

Melhores Hiperparâmetros

hyperparameters = {

'learning rate' = 0,4177,

'max_depth' = 3,

'max_features' = None,

'min_samples_leaf' = 3,

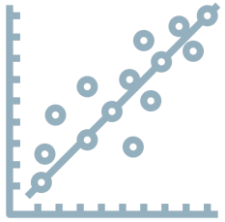
'min_samples_split' = 4,

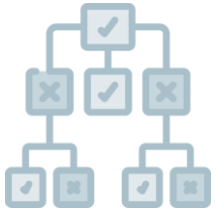
'n_estimators' = 388,

'subsample' = 0,803 }

Modelling

Otimização





Modelo

Avaliação

Gradient Boosting

GradientBoostingRegressor
(n_estimators=1000,
learning_rate=0.1,
max_depth=5,
random_state=42)

RMSE	0,524
R^2	0,994

RandomizedSearchCV()

Melhores Hiperparâmetros

hyperparameters = {

'learning rate' = 0,4177,

'max_depth' = 3,

'max_features' = None,

'min_samples_leaf' = 3,

'min_samples_split' = 4,

'n_estimators' = 388,

'subsample' = 0,803 }

OTIMIZAÇÃO

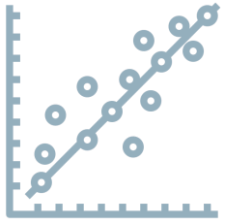
Gradient Boosting

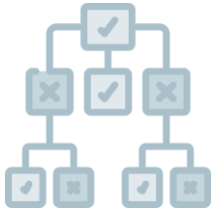
GradientBoostingRegressor
(**hyperparameters)

RMSE	0,057
R^2	0,998

Modelling

Otimização





Modelo

Avaliação

Gradient Boosting

GradientBoostingRegressor
(n_estimators=1000,
learning_rate=0.1,
max_depth=5,
random_state=42)

RMSE	0,524
R^2	0,994

RandomizedSearchCV()

Melhores Hiperparâmetros

hyperparameters = {

'learning rate' = 0,4177,

'max_depth' = 3,

'max_features' = None,

'min_samples_leaf' = 3,

'min_samples_split' = 4,

'n_estimators' = 388,

'subsample' = 0,803 }

OTIMIZAÇÃO

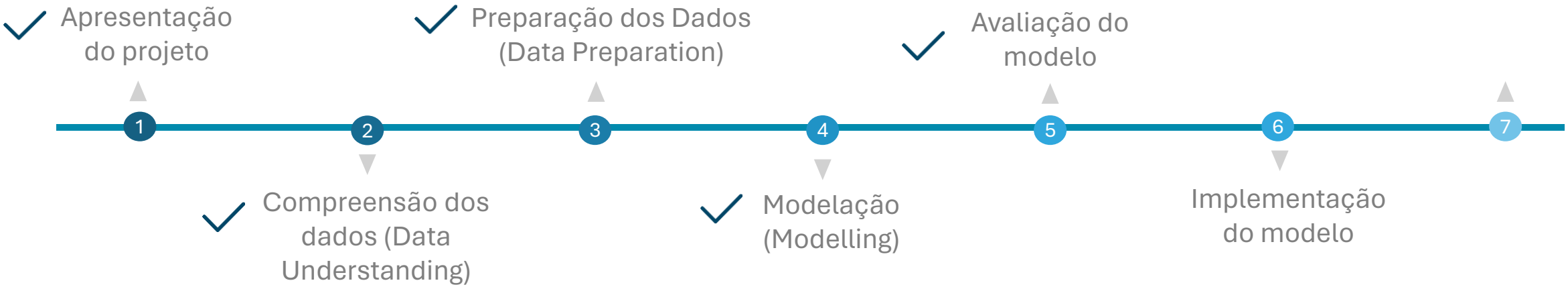
Gradient Boosting

GradientBoostingRegressor
(**hyperparameters)

RMSE	0,057
R^2	0,998

Implementação do Modelo

Enquadramento



Implementação do Modelo

Sugestões

Construção fluxos de trabalho em tempo real para preparação e limpeza dos dados

Implementação do Modelo

Sugestões

Construção fluxos de trabalho em tempo real para preparação e limpeza dos dados

Integração do modelo com os sistemas existentes

Implementação do Modelo

Sugestões

Construção fluxos de trabalho em tempo real para preparação e limpeza dos dados

Integração do modelo com os sistemas existentes

Monitorização contínua para acompanhar desempenho do modelo

Implementação do Modelo

Sugestões

Construção fluxos de trabalho em tempo real para preparação e limpeza dos dados

Integração do modelo com os sistemas existentes

Monitorização contínua para acompanhar desempenho do modelo

Retreinar periodicamente o modelo com dados novos

Implementação do Modelo

Sugestões

Construção fluxos de trabalho em tempo real para preparação e limpeza dos dados

Integração do modelo com os sistemas existentes

Monitorização contínua para acompanhar desempenho do modelo

Retreinar periodicamente o modelo com dados novos

Manter a documentação atualizada

Conclusão

Go Sara! Sem suporte! 😊

OBRIGADA!