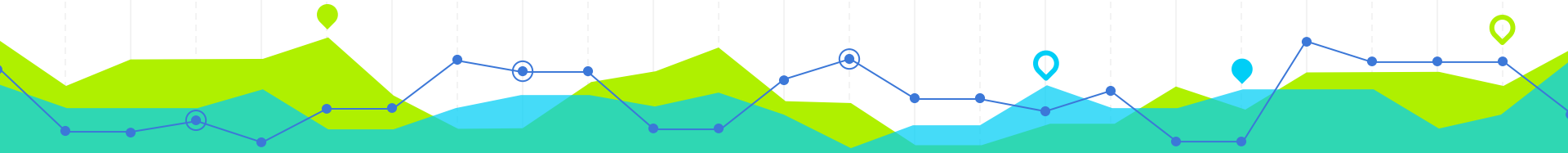
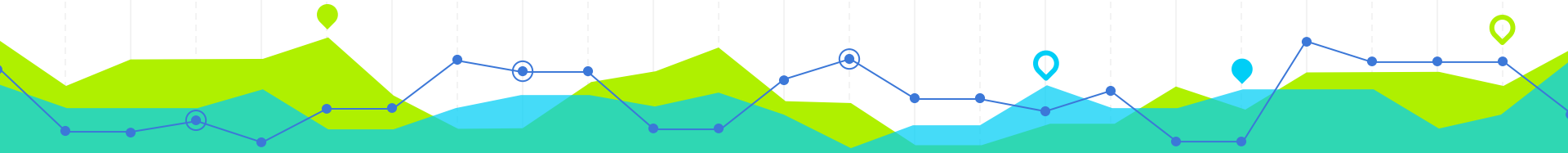


APLICAÇÃO DE TÉCNICAS DE MACHINE LEARNING PARA PREDIÇÃO DE TEMPO DE DESEMBARAÇO ADUANEIRO

Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data
PUC-MG



Contextualização

1

CONTEXTUALIZAÇÃO

Foco
Predição Tempo Despacho Importação

Base de Dados
146.397 Declarações de Importação do modal marítimo

Objetivo
Melhoria métricas e alocação de servidores

Ferramentas
Linguagem e bibliotecas Python

Data Science Workflow Canvas*

Start here. The sections below are ordered intentionally to make you state your goals first, followed by steps to achieve those goals. You're allowed to switch orders of these steps!

Title: APLICAÇÃO DE TÉCNICAS DE MACHINE LEARNING PARA PREDIÇÃO DE TEMPO DE DESEMBARAÇO ADUANEIRO

1 Problem Statement

What problem are you trying to solve?
What larger issues do the problem address?

A tarefa de Machine Learning é a predição do tempo de Despacho Aduaneiro nas Importações do modal marítimo. Trata-se, portanto, de uma tarefa de regressão.
A predição do tempo de desembaraço traz como benefícios a possibilidade de ajustes na alocação da força de trabalho das unidades aduaneiras e a melhoria das métricas de desempenho e produtividade. Já pelo lado do contribuinte permite uma racionalização da logística

2 Outcomes/Predictions

What prediction(s) are you trying to make?
Identify applicable predictor (X) and/or target (y) variables.

Variável de Predição: quantidade de horas brutas entre o registro e o desembaraço de Declarações de Importação.

Variáveis Preditoras: Dados abertos de Declarações de Importação.

Resultado: tempo de despacho aduaneiro de importação

3 Data Acquisition

Where are you sourcing your data from?
Is there enough data? Can you work with it?

Dados disponíveis no sítio da RFB https://receita.economia.gov.br/dados/resultados/aduana/estudos-e-analises/time-release-study-brasil já em formato de planilhas excel binárias.
De fácil obtenção e manipulação.

4 Modeling

What models are appropriate to use given your outcomes?

Dada a tarefa de regressão serão usados os modelos :

1. Arvore de Decisão de Regressão
2. Florestas Aleatórias de Regressão
3. CatBoost (Gradient Boosting)

5 Model Evaluation

How can you evaluate your model's performance?

Serão usadas as métricas padrão para regressão:

1. Coeficiente de Determinação (R2)
2. Erro Absoluto Médio (MAE)

6 Data Preparation

What do you need to do to your data in order to run your model and achieve your outcomes?

- Conversão do arquivo .xlsb para .xlsx
- Feature Engineering para obtenção de variáveis quantitativas e categóricas
- União dos datasets
- Encoding das variáveis categóricas para os modelos de Árvore de Decisão e Florestas Aleatórias.

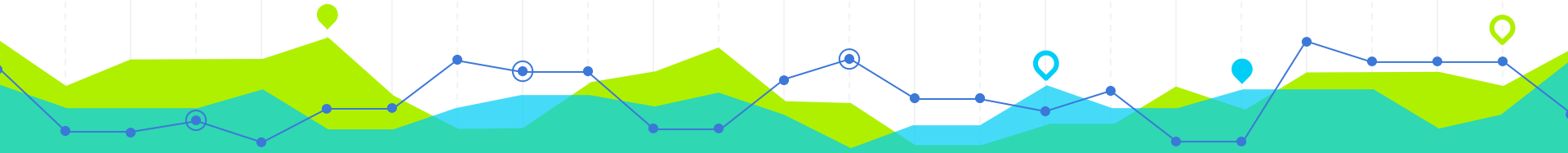
Activation

When you finish filling out the canvas above, now you can begin implementing your data science workflow in roughly this order.

1 Problem Statement → 2 Data Acquisition → 3 Data Prep → 4 Modeling → 5 Outcomes/Preds → 6 Model Eval

*** Note:** This canvas is intended to be used as a starting point for your data science projects. Data science workflows are typically nonlinear.

Conceptualized by Jasmine Vasandani using notes from General Assembly's Data Science Immersive. Format inspired by Business Model Canvas.



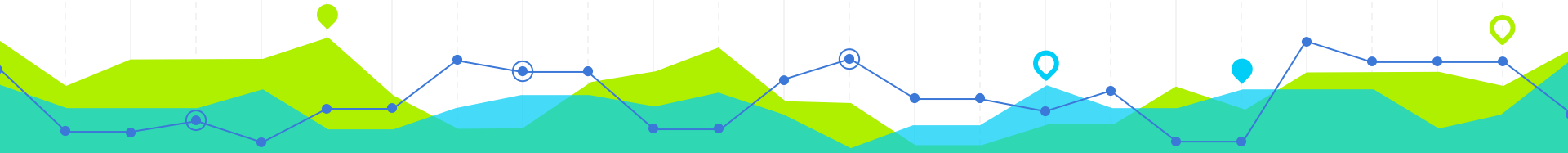
Coleta de Dados

2

COLETA DE DADOS

Conjunto de Planilhas Disponibilizadas no site da Receita Federal do Brasil:

- Dados DI
- Dados LI
- Dados Transito Aduaneiro



Tratamento dos Dados

3

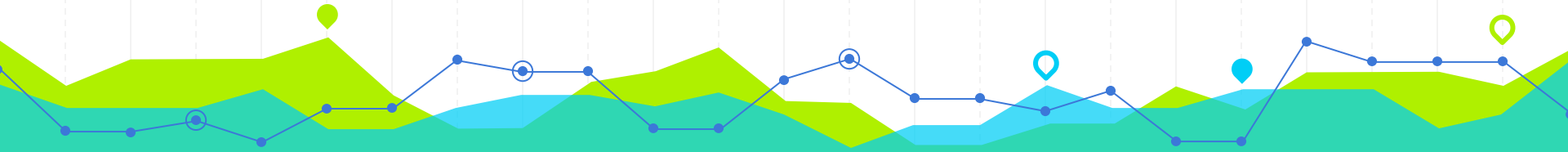
TRATAMENTO DE DADOS

- Processamento da variável rótulo
- Processamento de variáveis quantitativas
- Processamento de variáveis categóricas
- União dos Datasets
- Tratamento de Dados Ausentes

TRATAMENTO DE DADOS

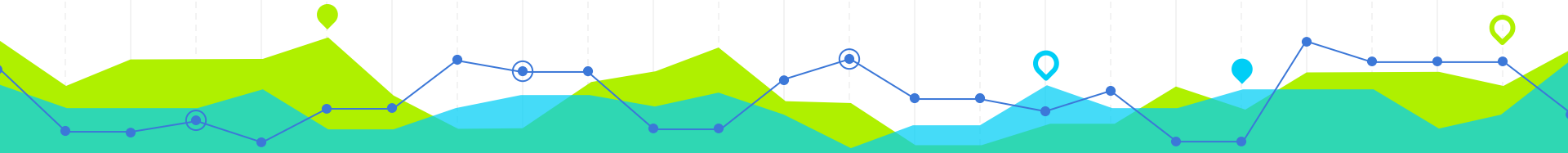
	ID DI	TIPO CE	OEA	MODALIDADE DESPACHO	TIPO DECLARACAO IMPORTACAO	CANAL	HORAS_EXIG	QTDE HORAS DESPACHO	QTDE HORAS PRESENCA	QTDE HORAS DISTRIBUICAO	...	QTDE LI MAPA	QTDE LI MCT	QTDE LI ANP	QT D
0	100010123323450	HBL	NAO	NORMAL	CONSUMO	VERDE	0.0	6.199444	12.200000	0.0	...	0.0	0.0	0.0	
1	10001863024225	HBL	NAO	NORMAL	CONSUMO	VERDE	0.0	21.353056	241.616667	0.0	...	0.0	0.0	0.0	
2	100031430658855	HBL	NAO	NORMAL	CONSUMO	VERDE	0.0	8.811111	40.150000	0.0	...	0.0	0.0	0.0	
3	100031604383988	HBL	NAO	NORMAL	CONSUMO	VERDE	0.0	5.201389	34.183333	0.0	...	0.0	0.0	0.0	
4	100036615433154	BL	NAO	NORMAL	CONSUMO	VERDE	0.0	0.000000	8.900000	0.0	...	0.0	0.0	0.0	

5 rows × 28 columns



146.937

Declarações de Importação do modal Marítimo



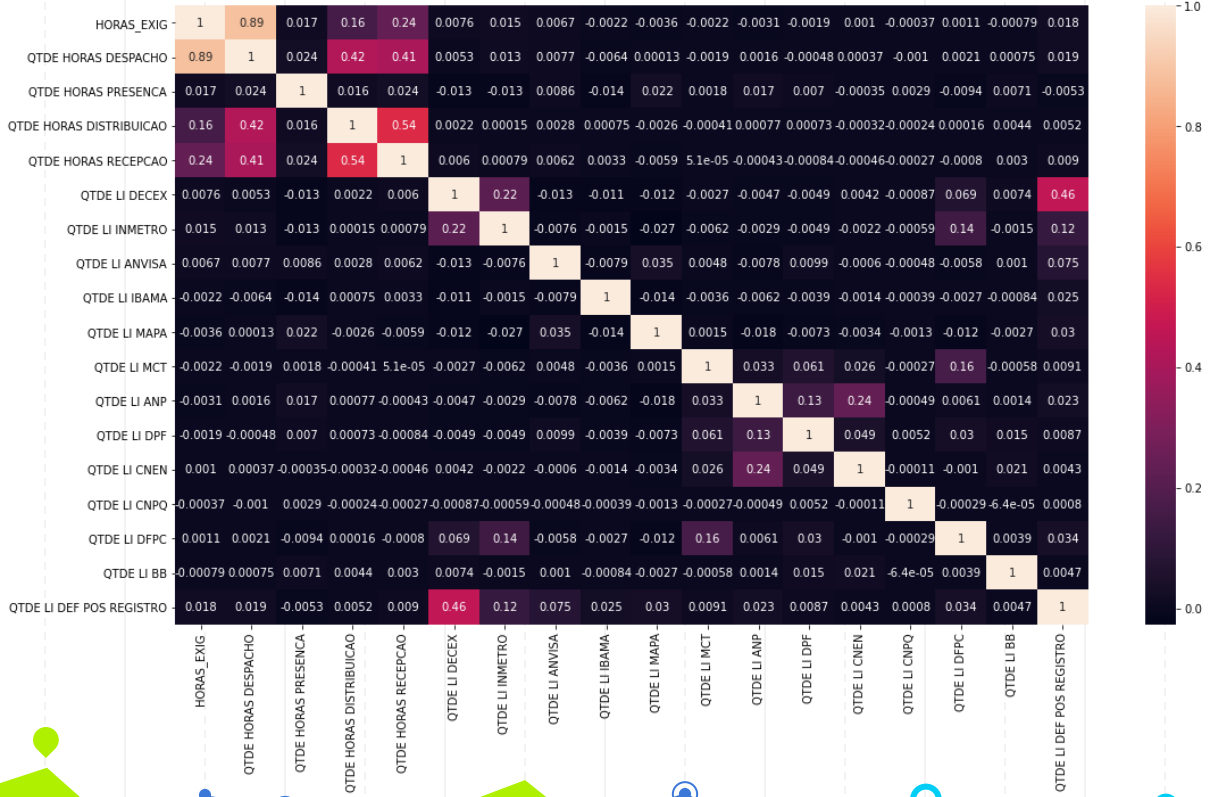
Análise e Exploração de Dados

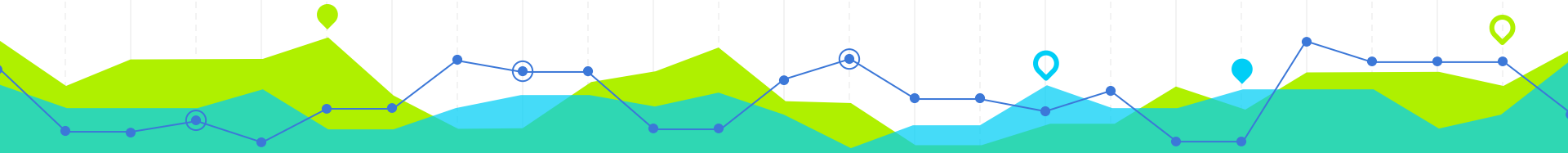
4

ANÁLISE E EXPLORAÇÃO DE DADOS

- Cardinalidade
- Correlação
- Seleção de Variáveis

HEATMAP CORRELAÇÕES





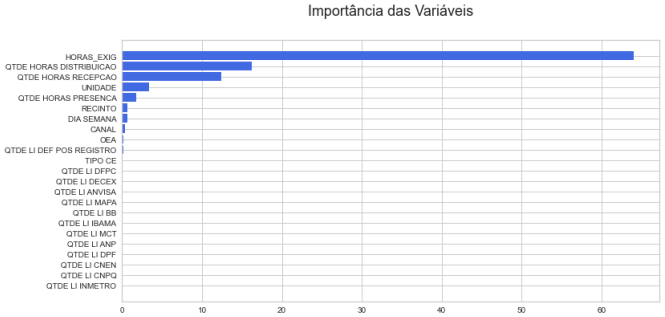
Modelos Machine Learning

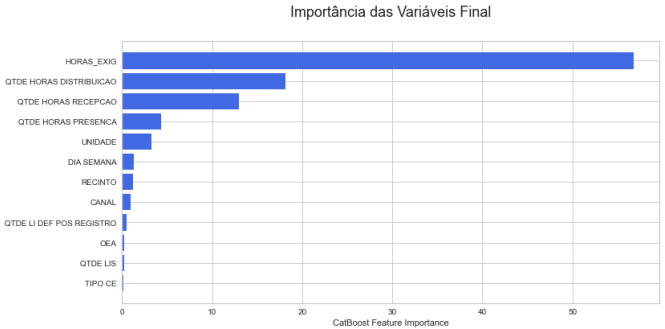
5

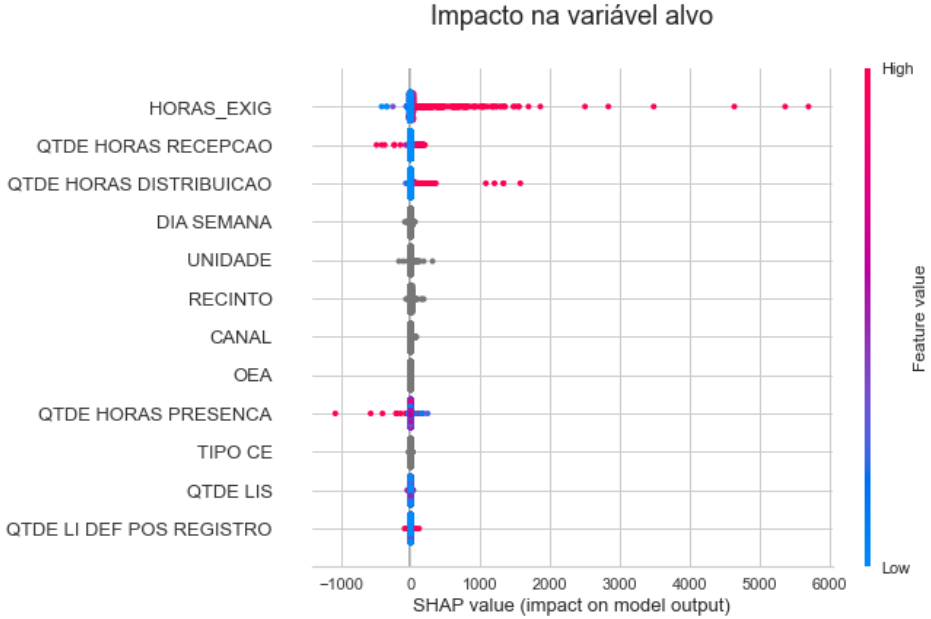
MODELOS MACHINE LEARNING

- Árvore de Decisão
- Floresta Aleatória
- Gradient Boosting (CatBoost)

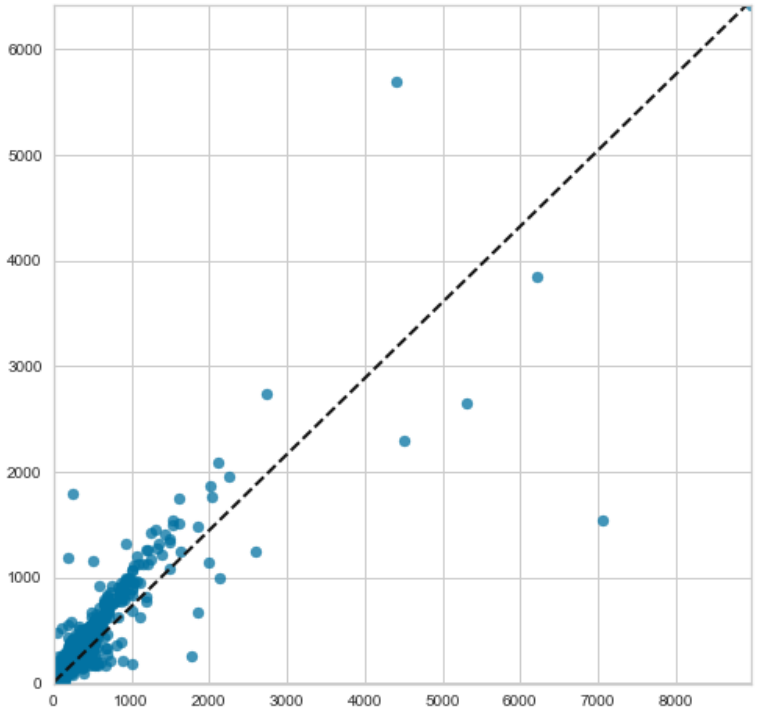
SELEÇÃO VARIÁVEIS COM BASE NA EXPLICAÇÃO DO MODELO







ERROS NA PREDIÇÃO



PRINCIPAIS RESULTADOS OBTIDOS

	R^2	MAE
Arvore de Decisão com busca em grade	0,6372	11,7339
Floresta Aleatória com busca em grade	0,7610	10,7627
CatBoost com busca em grade	0,8314	11,5921

CONCLUSÕES

- Resultado Satisfatório
- Seleção de variáveis não limitada
- Ganhos para a sociedade

OBRIGADO!

André Leonardo de La Corte

Auditor-Fiscal da Receita Federal do Brasil