Feature Engineering

Prof. Leandro Romualdo da Silva

Big Data Borat
@BigDataBorat

Follow

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

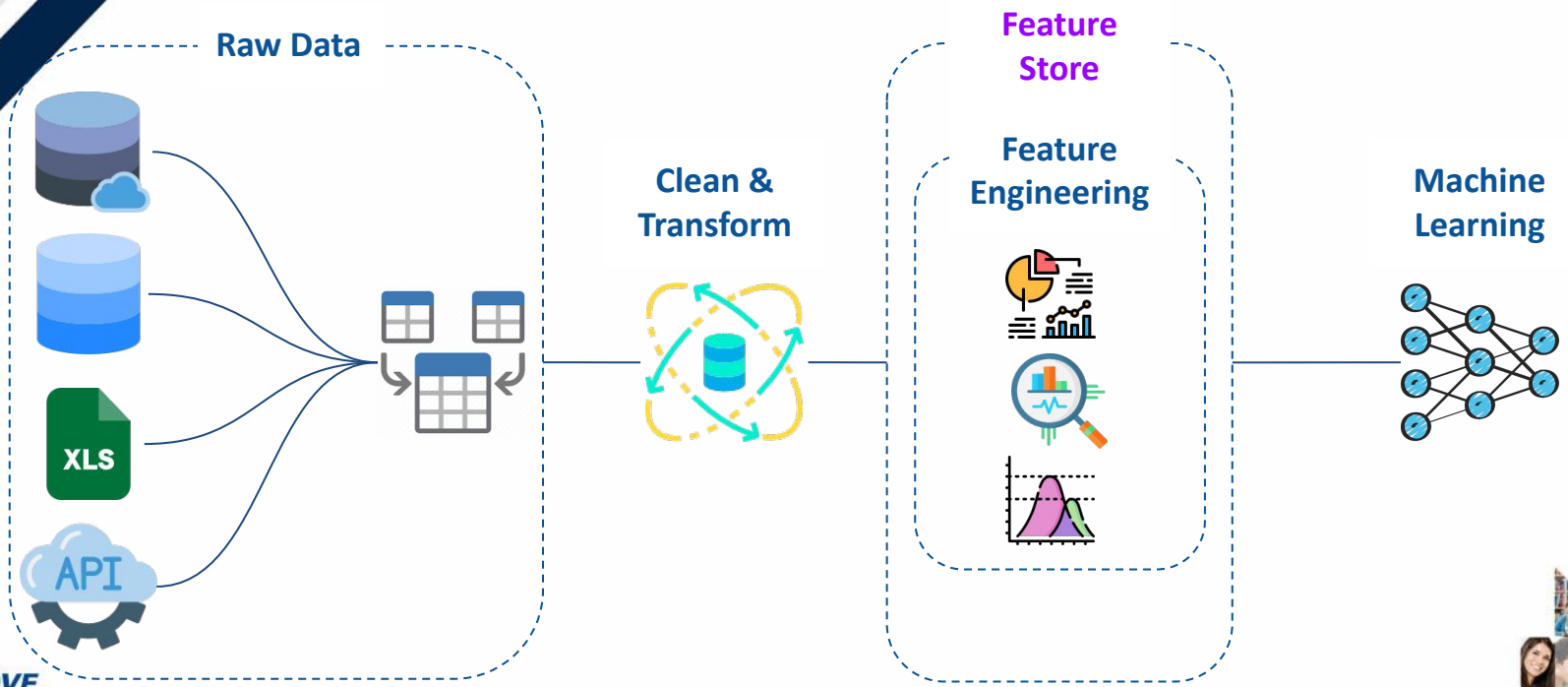
Feature Engineering

Feature engineering tem o objetivo de tornar os dados mais adequados ao problema em questão.

Feature engineering geralmente é usada para:

- **Melhorar performance de um modelo preditivo**
- **Reduzir as necessidades computacionais e/ou de dados**
- **Melhorar a interpretabilidade dos resultados do modelo**

Feature Engineering



UNINOVE
Universidade Nove de Julho
A UNINOVE COM VOCÊ É 10

A caixa de ferramentas para dados

Features de data

Abordagens

Features baseada em data

Raw Data

01/09/2022 16:30:03

- Dia: 01
- Mês: 09
- Ano: 2022
- Semestre: 2º
- Semana: 35
- Dia da semana: 06
- Final de semana: Sim/Não
- Faixa Horária: tarde

UNINOVE
Universidade Nove de Julho
A UNINOVE COM VOCÊ É 10

Features numéricas

Abordagens

Imputar dados ausentes

- Ignorar as linhas com dados ausentes
- Eliminar colunas com muitos dados ausentes (>20%)
- Estratégias

  - Media : Abordagem básica, pode sofrer com outliers
  - Mediana : Mais robusta e tende a sofrer menos com dados extremos
  - Moda: Abordagem simples, funcional em alguns casos
  - Inferência através de modelo: Sofisticado, mas tome cuidado com viés do modelo

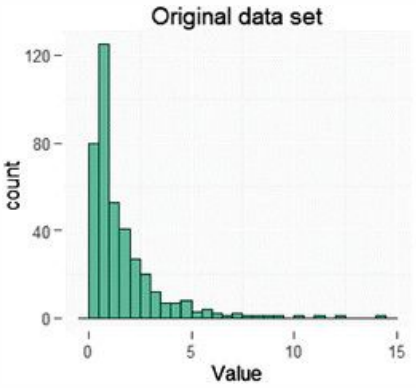
Abordagens

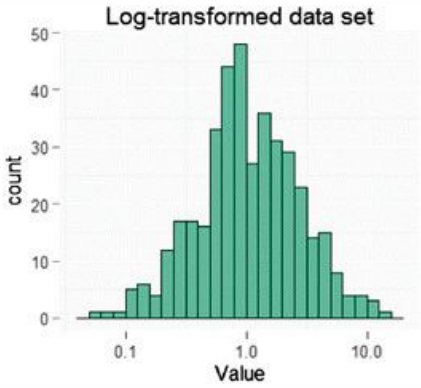
Features de valores contínuos - Logaritmo

Raw Data

Menor valor R$0,00

Maior valor R$15.000,00





Logaritmo

É o inverso da função exponencial. Exemplo: O Log de 1000 na base 10 é 3 e 10 elevado ao cubo é 1000.

UNINOVE
Universidade Nove de Julho
A UNINOVE COM VOCÊ É 10

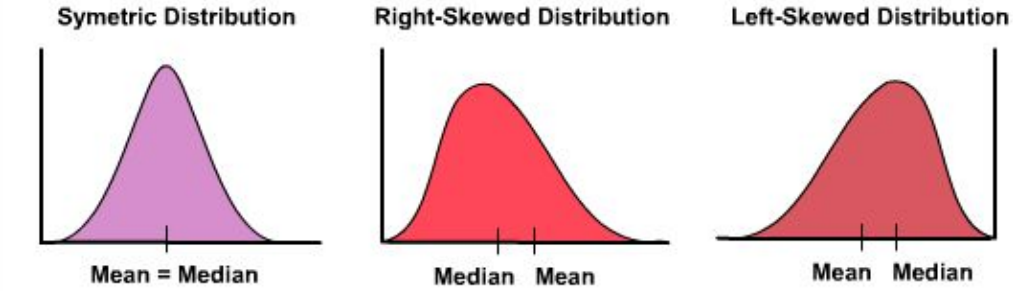
Abordagens

Features de valores contínuos - Normalização

Raw Data

Menor valor R$0,00

Maior valor R$15.000,00



Normalização

Transforma os dados dentro do intervalo de 0 e 1, caso tenham valores negativos -1 e 1.

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

UNINOVE
Universidade Nove de Julho
A UNINOVE COM VOCÊ É 10

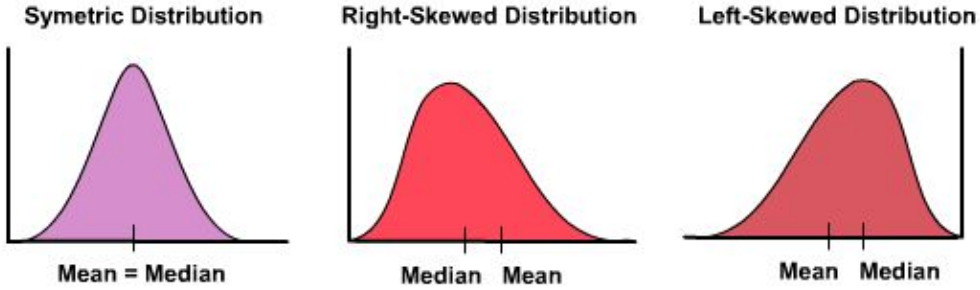
Abordagens

Features de valores contínuos - Padronização

Raw Data

Menor valor R$0,00

Maior valor R$15.000,00



Padronização

Transforma os dados para valores cuja média seja 0 e desvio padrão 1.

$$z = \frac{x - \mu}{\sigma}$$

UNINOVE
Universidade Nove de Julho
A UNINOVE COM VOCÊ É 10

Features categóricas

Abordagens

Features baseada em dados categóricos

Raw Data

ID	Data	Tipo transação
19854	01/09/2022	Aplicativo
19855	01/09/2022	Ecommerce
19856	01/09/2022	Loja
19857	01/09/2022	Outros

ID	Data	Aplicativo	Ecommerce	Loja	Outros
19854	01/09/2022	1	0	0	0
19855	01/09/2022	0	1	0	0
19856	01/09/2022	0	0	1	0
19857	01/09/2022	0	0	0	1

Técnica conhecida como get Dummies ou OHE(One-hot Encoding)

Features de espaço

Abordagens

Features baseada em localização

Raw Data

ID	latitude início	longitude início	latitude fim	longitude fim
19854	-23.503552	-46.661222	-32.986459	-53.945363
19855	-24.106386	-46.823158	-23.503552	-46.823158
19856	-29.631086	-49.318258	-32.986459	-49.318258
19857	-32.986459	-53.945363	-23.503552	-46.661222

ID	latitude início	longitude início	latitude fim	longitude fim	Distance
19854	-23.503552	-46.661222	-32.986459	-53.945363	52KM
19855	-24.106386	-46.823158	-23.503552	-46.823158	35KM
19856	-29.631086	-49.318258	-32.986459	-49.318258	82KM
19857	-32.986459	-53.945363	-23.503552	-46.661222	102KM

Podemos calcular distância, tempo e gerar outras features com base na localização.