BigML Datasets

2023

<lapti>

Programa

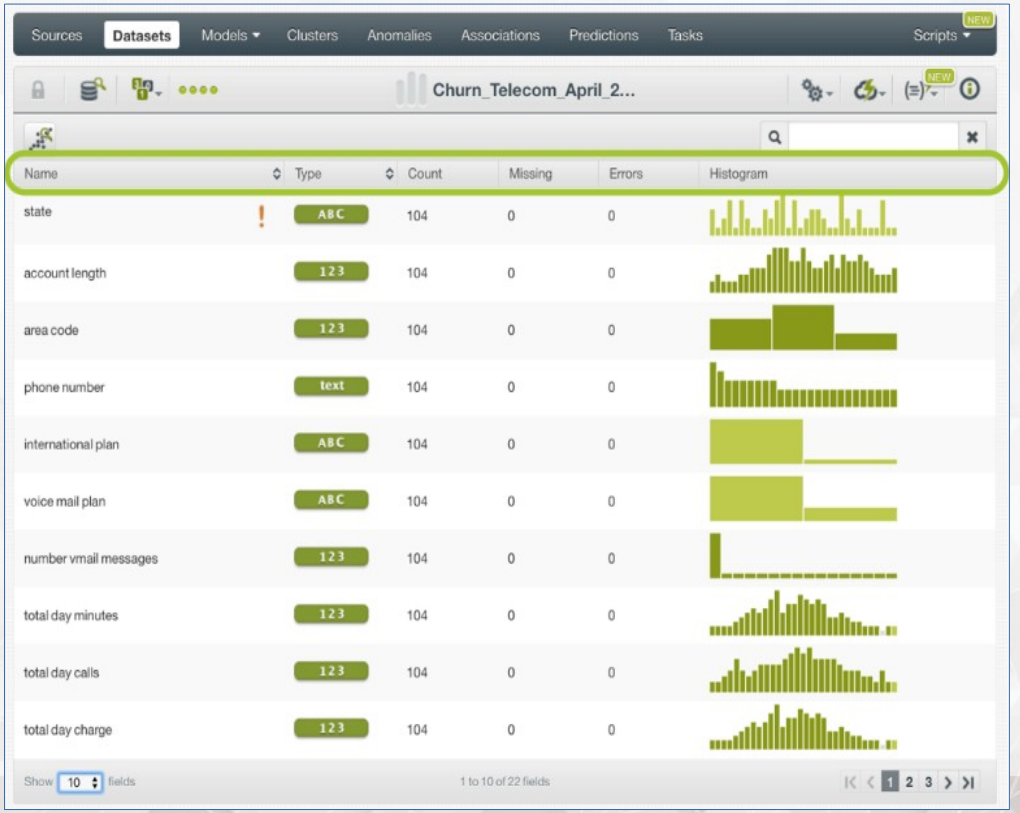
- Introdução
  - Workflow típico: 1-click creation
  - Propósitos de datasets em BigML
    - Exploração
    - Pre-flight check
- Features básicas
  - Outras maneiras para se criar datasets
  - Train/test split
  - Mais exploração
- Features avançadas
  - Filtros
  - Feature engineering com Flatline

O que é um Dataset?

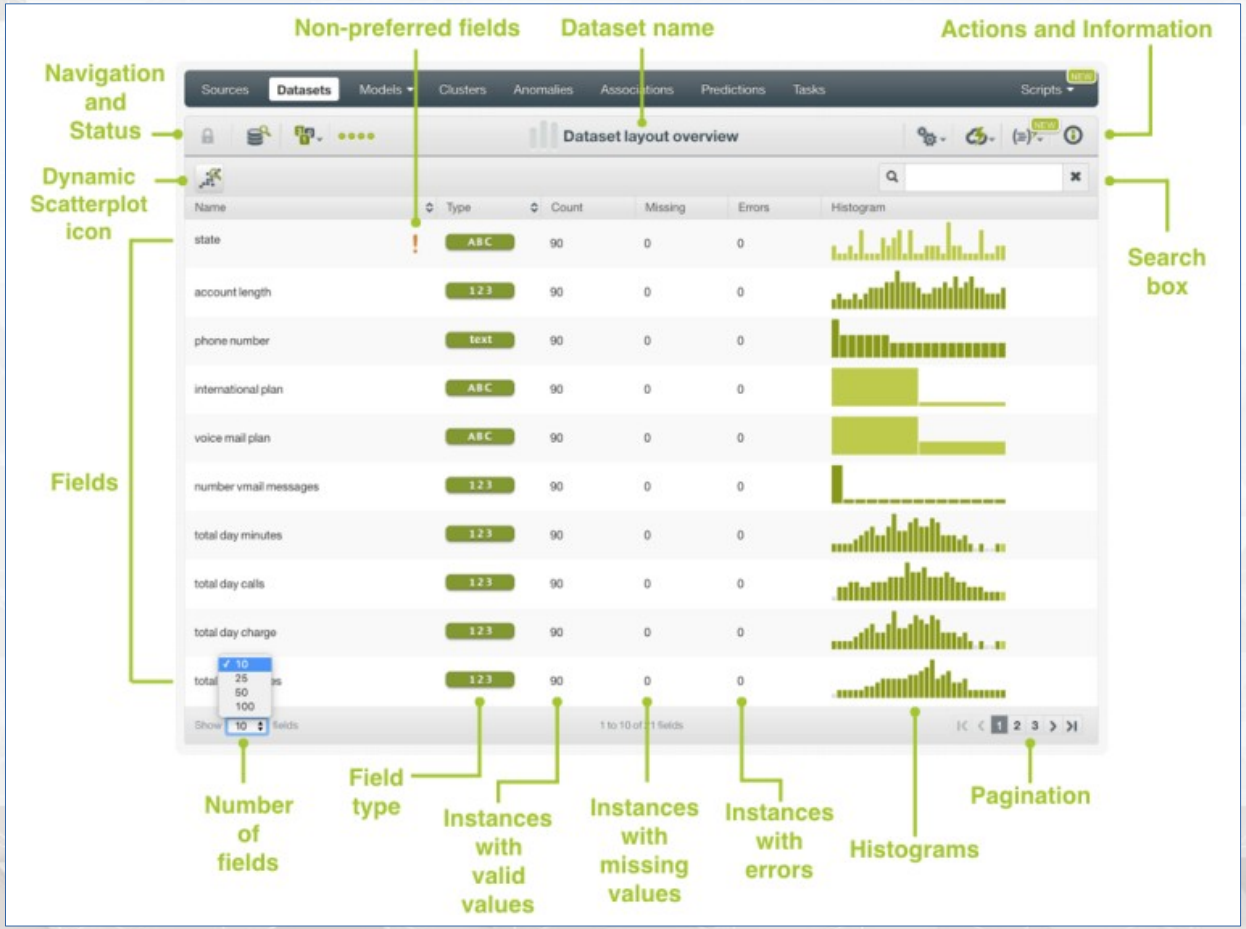
- Datasets são os blocos fundamentais em BigML
  - Decision Trees, Clusters, etc, todos derivam de datasets
  - Fontes evoluem somente para datasets
- Versão estruturadas dos dados
- Permite “wrangling” dos dados
- Exploração de dados / pre-flight check
  - Missing / Erros
  - Estatísticas básicas
  - Campos não utilizáveis (non-preferred)
  - Objetivo default para 1-click actions

Entendendo Datasets

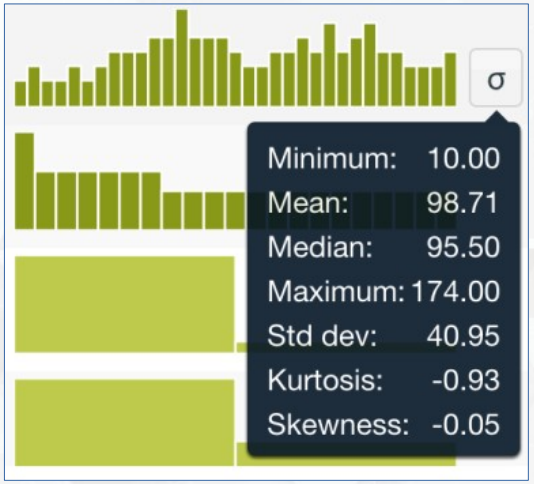
- Versão estruturada dos dados
- Visualização de estatísticas e estado dos campos
  - Count
  - Missing
  - Errors

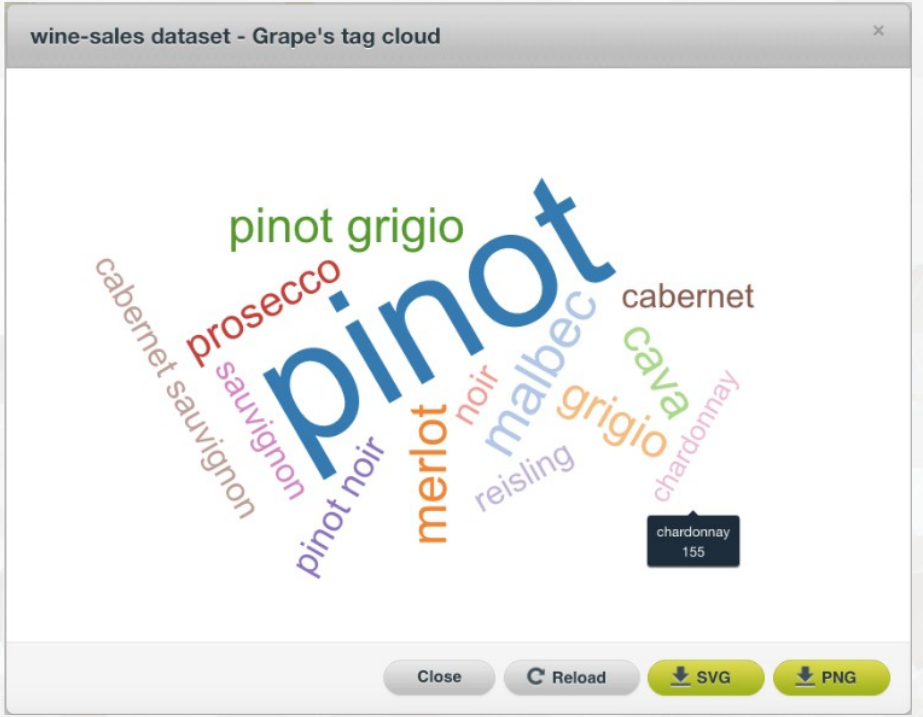


Entendendo Datasets



Estatísticas dos campos





Datasets

- Imutáveis
  - Como outros resources em BigML
  - Possui um ID acessível a partir da API
- Criando datasets
  - A partir de uma source
  - A partir de outro dataset
    - Sampling, filter, training/test
    - De uma saída de lote
      - Batch output, de uma previsão de outro algoritmo
- Dynamic scatterplot

Opções de configuração

- Relativo ao source de origem
  - Nome
  - Tamanho
    - Relativo ao source
  - Inclusão e exclusão de campos
- Navegação

Opções de configuração

- Configuração de campos

Scatterplot

- Analisa uma amostra (até 500 instâncias)
  - Padrões
  - Correlação entre campos
  - Datapoints anômalos

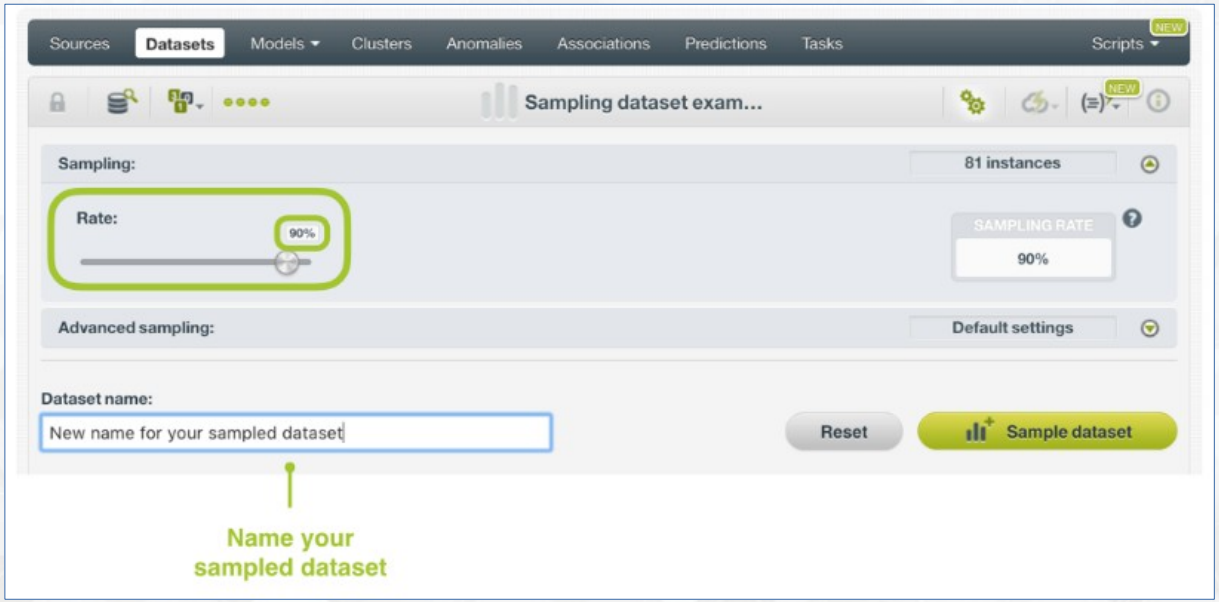
Configurações avançadas

- Filtros
- Sampling
- Feature engineering

- *Obs: decidir entre executar essas ações dentro de um ambiente de AutoML ou em um ambiente apropriado de manipulação de dados*
  - *Data wrangler*
  - *Cleaning, sanitizing (curadoria)*

Sampling e Filtering

- Amostragem
  - Volume
  - Modo de amostragem
- Filtragem
  - Por expressões usando campos
  - Usando linguagem de tratamento
    - Flatline

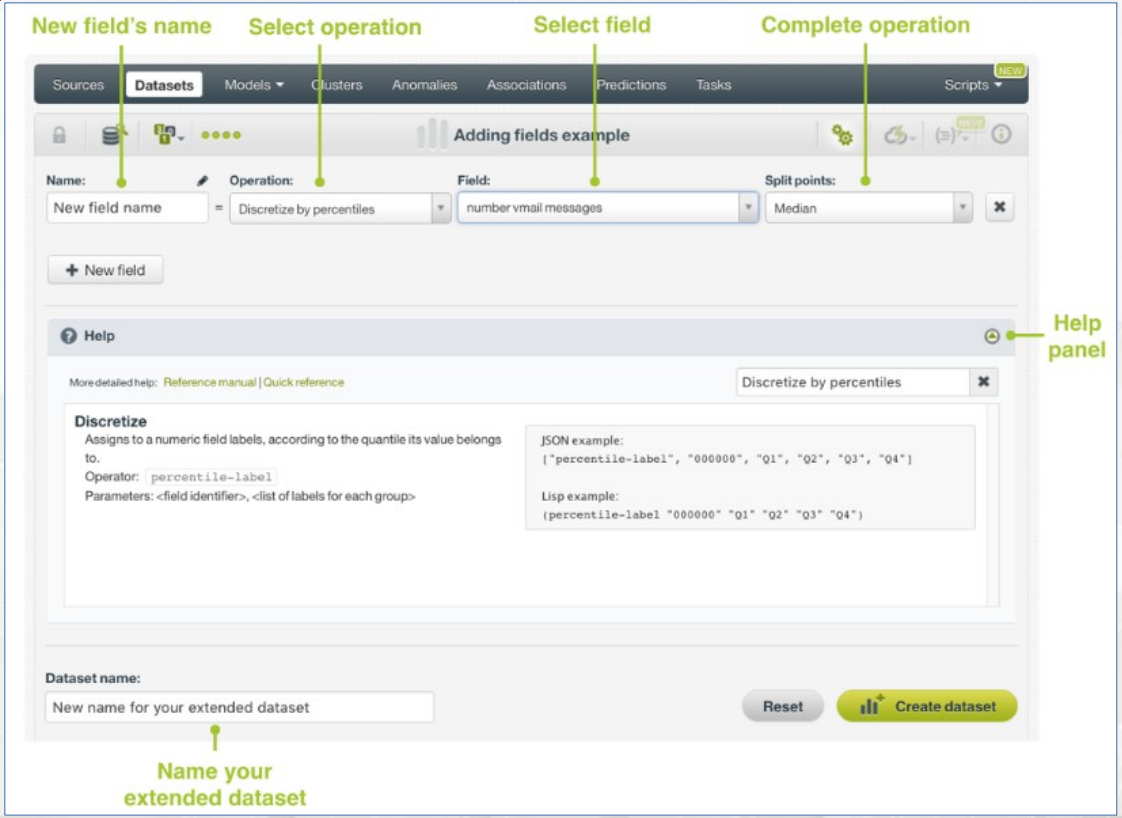


Sampling para treinamento/teste

- Dividir o dataset em porções diferentes
  - Para validar os futuros modelos
  - 1-Click random split
  - Configurações avançadas
    - Divisão aleatória
      - Opções adicionais
    - Divisão linear

Transformação de datasets

- Adição de campos
  - Operações sobre campos existentes para geração de novos
- Discretização
  - Por percentis, segmentação
- Substituição de valores faltando
- Normalização
- Sliding windows
  - Agregação de campos com base temporal



Join de datasets

- Semelhante a join em SQL
  - Tão semelhante que é interessante ponderar sobre seu uso...

employee_id	name	...	department_id
1	John	...	5
2	Rose	...	8
3	Rick	...	2
4	Pat	...	5
6	Patrick	...	9
...	...	...	...
1467	Mike	...	4

department_id	name	...	budget
1	HR	...	1,500,000
2	Accounting	...	1,900,000
3	Sales	...	3,400,000
4	Developing	...	5,000,000
5	Operations	...	2,500,000
...	...	...	...
9	Projects	...	1,900,000

Join by department_id

employee_id	name	...	department_id	name	budget
1	John	...	5	Operations	2,500,000
2	Rose	...	8	Systems	4,500,000
3	Rick	...	2	Accounting	1,900,000
4	Pat	...	5	Operations	2,500,000
6	Patrick	...	9	Projects	1,900,000
...	...	...	...	...	...
1467	Mike	...	4	Developing	5,000,000

Consumindo datasets

- Exportação para CSV ou Tableau (TDE)
- Usando via API Rest ou Bindings

```
curl "https://bigml.io/dataset?$BIGML_AUTH" \
     -X POST \
     -H 'content-type: application/json' \
     -d '{"source":"source/50a4527b3c1920186d000041", "name": "my dataset"}'
```

```
from bigml.api import BigML
api = BigML ()
dataset = api.create_dataset('source/50a4527b3c1920186d000041')
```

Limites

- Sem limite para número de campos
- Sem limite para número de instâncias
- No máximo 1.000 classes distintas por campo
  - Campos categóricos
- No máximo 1.000 termos no total
  - Em todos os campos texto
- No máximo 10.000 itens por campo
  - Em lista de itens

Então

- Para que servem datasets
  - Bloco fundamental
  - Pre-flight check, contadores, histograma, scatterplot
- Criando datasets
  - A partir de fontes
    - 1-click e sampling
  - Training/test split
  - Batch output
  - De outro dataset
    - Sampling, filtering, new features

Obrigado

leandro@utfpr.edu.br
http://lapti.ct.utfpr.edu.br

<lapti>