



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

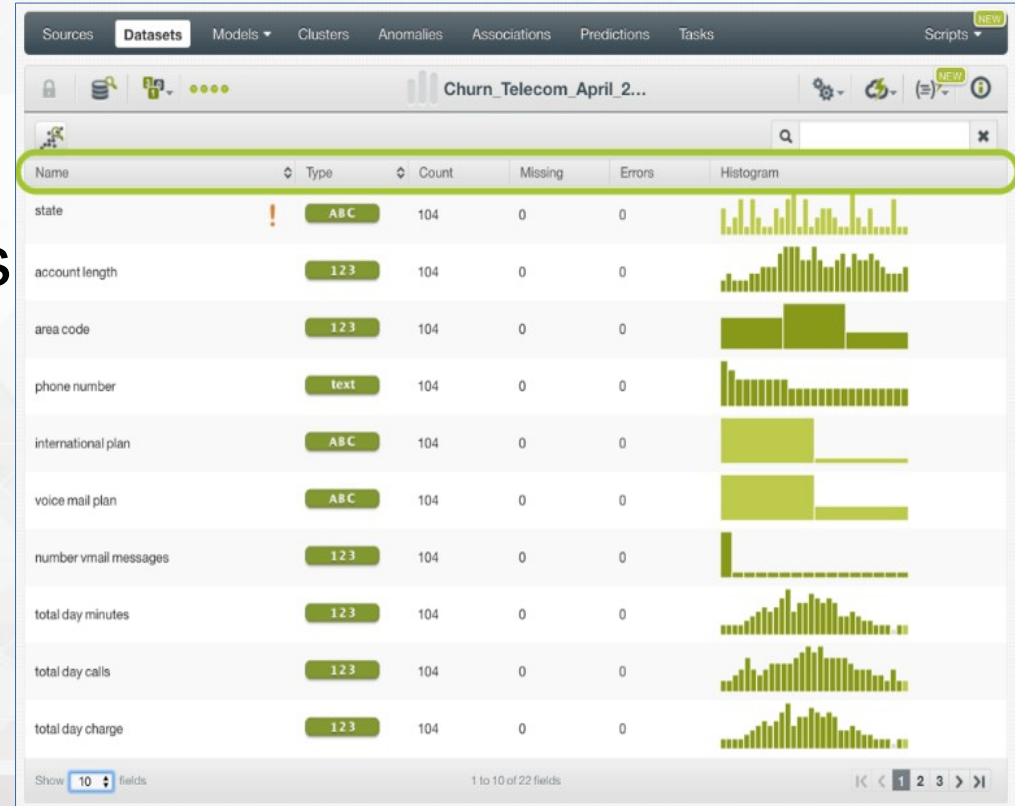
Entendendo Datasets

- Versão estruturada dos dados
- Visualização de estatísticas e estado dos campos

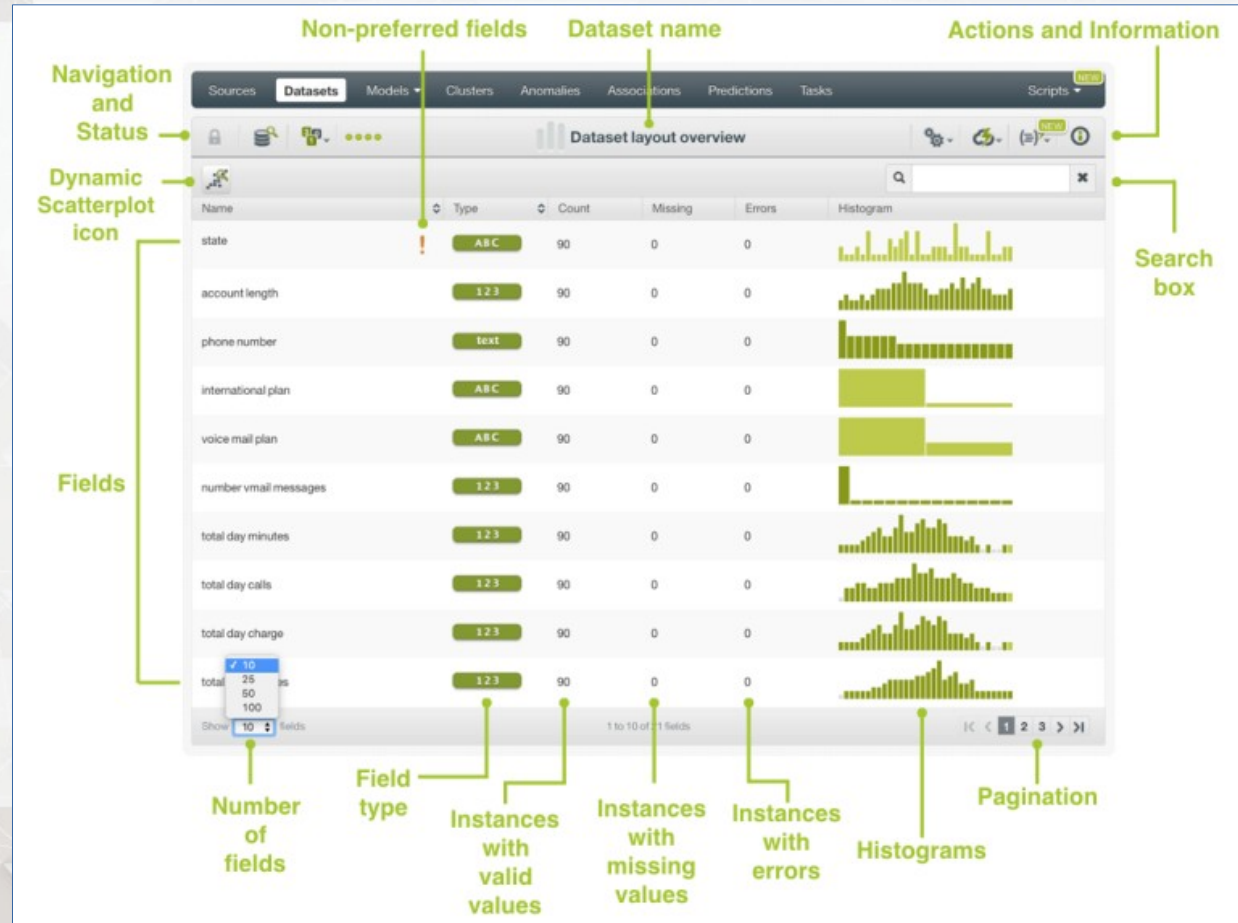
~ Count

~ Missing

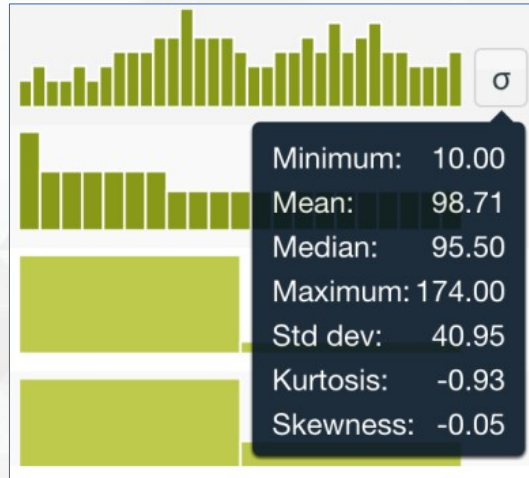
~ Errors



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

The screenshot displays a web application interface for managing datasets. The top navigation bar includes tabs for Sources, Datasets (active), Ensembles, Clusters, Anomalies, Associations, Predictions, Tasks, and Scripts. Below the navigation bar, a list of fields is shown with columns for Name, Type, and a status icon. The fields listed are: Name, Work status, Marital status, Children, Age, Gender, and Race. A modal window titled "Update field 000001 details" is open, showing the configuration for the "Marital status" field. The modal includes fields for Name, Label, and Description. The Name field contains "Marital status", the Label field contains "married, divorced, separated, widowed", and the Description field contains "This field indicates the marital status of each instance." There is a checkbox for "Preferred / non-preferred field" and a target icon for "Objective field". A tooltip message "This field is not the objective field" is visible near the target icon. The modal also has "Save" and "Cancel" buttons. On the right side of the interface, there is a histogram chart. The bottom of the interface shows a pagination bar with "1 to 6 of 6 fields" and navigation controls.

Annotations in the image:

- Name**: Points to the "Name" column header in the field list.
- Label**: Points to the "Label" field in the modal.
- Edit icon**: Points to the edit icon in the field list.
- Description**: Points to the "Description" field in the modal.
- Preferred / non-preferred field**: Points to the checkbox in the modal.
- Objective field**: Points to the target icon in the modal.

Scatterplot

- Analisa uma amostra (até 500 instâncias)

~ Padrões

~ Correlação entre campos

~ Datapoints anômalos

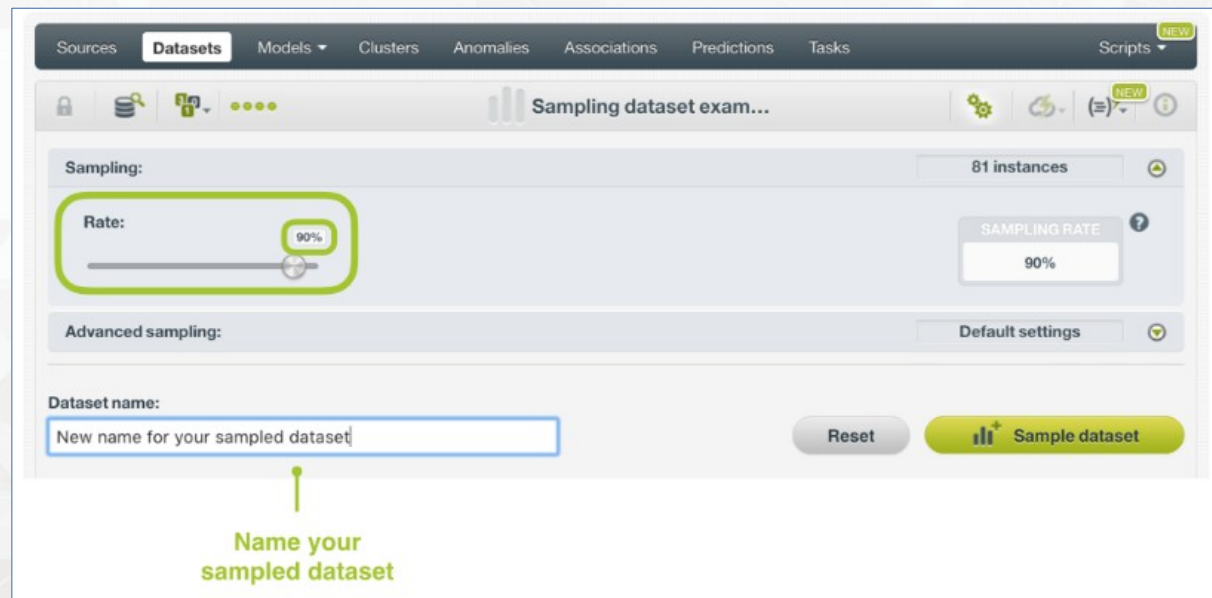


Configurações avançadas

- 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

New field's name Select operation Select field Complete operation

Sources Datasets Models Clusters Anomalies Associations Predictions Tasks Scripts

Adding fields example

Name: Operation: Field: Split points:

New field name = Discretize by percentiles number vmail messages Median

+ New field

Help

More detailed help: [Reference manual](#) | [Quick reference](#)

Discretize by percentiles

Discretize
Assigns to a numeric field labels, according to the quantile its value belongs to.
Operator: `percentile-label`
Parameters: `<field identifier>`, `<list of labels for each group>`

JSON example:
`{"percentile-label", "000000", "Q1", "Q2", "Q3", "Q4"}`

Lisp example:
`(percentile-label "000000" "Q1" "Q2" "Q3" "Q4")`

Dataset name:

New name for your extended dataset

Reset Create dataset

Name your extended dataset

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

The background features a light gray geometric pattern of triangles. At the top, there is a horizontal band with a yellow and gray geometric design, including a stylized 'U' shape. The word 'Obrigado' is centered in a large, black, sans-serif font.

Obrigado

leandro@utfpr.edu.br

<http://lapti.ct.utfpr.edu.br>

<lapti>