Package: correlationfunnel (via r-universe)

June 14, 2024

Type Package

Title Speed Up Exploratory Data Analysis (EDA) with the Correlation Funnel

Version 0.2.0

Description Speeds up exploratory data analysis (EDA) by providing a succinct workflow and interactive visualization tools for understanding which features have relationships to target (response). Uses binary correlation analysis to determine relationship. Default correlation method is the Pearson method. Lian Duan, W Nick Street, Yanchi Liu, Songhua Xu, and Brook Wu (2014) <doi:10.1145/2637484>.

URL https://github.com/business-science/correlationfunnel, https://business-science.github.io/correlationfunnel/

BugReports https://github.com/business-science/correlationfunnel/issues

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Depends R (>= 3.1)

Imports ggplot2, rlang, recipes, magrittr, plotly, tibble, dplyr (>= 1.0.0), tidyr (>= 1.0.0), stats, utils, ggrepel, stringr, forcats, purrr, cli, crayon, rstudioapi

Suggests scales, knitr, rmarkdown, covr, lubridate, testthat (>= 2.1.0)

RoxygenNote 7.1.0

Roxygen list(markdown = TRUE)

VignetteBuilder knitr

Repository https://business-science.r-universe.dev

RemoteUrl https://github.com/business-science/correlationfunnel

RemoteRef HEAD

RemoteSha e592ef37ff50f8cbdb2763681f6a0a03aad72611

Contents

`binarize` *Turn data with numeric, categorical features into binary data.*

Description

`binarize` returns the binary data coverted from data in normal (numeric and categorical) format.

Usage

```
binarize(
  data,
  n_bins = 4,
  thresh_infreq = 0.01,
  name_infreq = "-OTHER",
  one_hot = TRUE
)
```

Arguments

`data`	A `tibble` or `data.frame`
`n_bins`	The number of bins to for converting continuous (numeric features) into discrete features (bins)
`thresh_infreq`	The threshold for converting categorical (character or factor features) into an "Other" Category.
`name_infreq`	The name for infrequently appearing categories to be lumped into. Set to "-OTHER" by default.
`one_hot`	If set to `TRUE`, binarization returns number of new columns = number of levels. If `FALSE`, binarization returns number of new columns = number of levels - 1 (dummy encoding).

Details

The Goal:

The binned format helps correlation analysis to identify non-linear trends between a predictor (binned values) and a response (the target)

What Binarize Does:

The `binarize()` function takes data in a "normal" format and converts to a binary format that is useful as a preparation step before using `correlate()`:

Numeric Features: The "Normal Data" format has numeric features that are continuous values in numeric format (`double` or `integer`). The `binarize()` function converts these to bins (categories) and then discretizes the bins using a one-hot encoding process.

Categorical Features: The "Normal Data" format has categorical features that are `character` or `factor` format. The `binarize()` function converts these to binary features using a one-hot encoding process.

Value

A `tbl`

Examples

```
library(dplyr)
library(correlationfunnel)

marketing_campaign_tbl %>%
    select(-ID) %>%
    binarize()
```

`correlate` *Correlate a response (target) to features in a data set.*

Description

`correlate` returns a correlation between a target column and the features in a data set.

Usage

```
correlate(data, target, ...)
```

Arguments

`data`	A `tibble` or `data.frame`
`target`	The feature that contains the response (Target) that you want to measure relationship.
`...`	Other arguments passed to cor

Details

The `correlate()` function provides a convient wrapper around the cor function where the `target` is the column containing the Y variable. The function is intended to be used with `binarize()`, which enables creation of the binary correlation analysis, which is the feed data for the `plot_correlation_funnel()` visualization.

The default method is the Pearson correlation, which is the Correlation Coefficient from L. Duan et al., 2014. This represents the linear relationship between two dichotomous features (binary variables). Learn more about the binary correlation approach in the Vignette covering the Methodology, Key Considerations and FAQs.

Value

A `tbl`

References

Lian Duan, W. Nick Street, Yanchi Liu, Songhua Xu, and Brook Wu. 2014. Selecting the right correlation measure for binary data. ACM Trans. Knowl. Discov. Data 9, 2, Article 13 (September 2014), 28 pages. DOI: http://dx.doi.org/10.1145/2637484

See Also

`binarize()`, `plot_correlation_funnel()`

Examples

```
library(dplyr)
library(correlationfunnel)

marketing_campaign_tbl %>%
    select(-ID) %>%
    binarize() %>%
    correlate(TERM_DEPOSIT__yes)
```

---

customer_churn_tbl *Customer Churn Data Set for a Telecommunications Company*

Description

A dataset containing data related to telecom customers that have enrolled in various products and services

Usage

```
customer_churn_tbl
```

Format

An object of class `spec_tbl_df` (inherits from `tbl_df`, `tbl`, `data.frame`) with 7043 rows and 21 columns.

Telecom Customer Data:

- customerID (chr): CUSTOMER ID
- gender (chr): Customer's gender ("Female", "Male")
- SeniorCitizen (dbl): 1 = Senior Citzen, 0 = Not Senior Citizen
- Partner (chr): Whether the customer has a partner or not (Yes, No)
- Dependents (chr): Whether the customer has dependents or not (Yes, No)
- tenure (dbl): Number of months the customer has stayed with the company
- PhoneService (chr): Whether the customer has a phone service or not (Yes, No)
- MultipleLines (chr): Whether the customer has multiple lines or not (Yes, No, No phone service)
- InternetService (chr): Customer's internet service provider (DSL, Fiber optic, No)
- OnlineSecurity (chr): Whether the customer has online security or not (Yes, No, No internet service)
- OnlineBackup (chr): Whether the customer has online backup or not (Yes, No, No internet service)
- DeviceProtection (chr): Whether the customer has device protection or not (Yes, No, No internet service)
- TechSupport (chr): Whether the customer has tech support or not (Yes, No, No internet service)
- StreamingTV (chr): Whether the customer has streaming TV or not (Yes, No, No internet service)
- StreamingMovies (chr): Whether the customer has streaming movies or not (Yes, No, No internet service)
- Contract (chr): The contract term of the customer (Month-to-month, One year, Two year)
- PaperlessBilling (chr): Whether the customer has paperless billing or not (Yes, No)
- PaymentMethod (chr): The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
- MonthlyCharges (dbl): The amount charged to the customer monthly
- TotalCharges (dbl): The total amount charged to the customer
- Churn (chr): Outcome. Whether the customer churned or not (Yes or No)

Source

IBM Sample Datasets

`marketing_campaign_tbl`
Marketing Data for a Bank

Description

A dataset containing data related to bank clients, last contact of the current marketing campaign, and attributes related to a previous marketing campaign.

Usage

```
marketing_campaign_tbl
```

Format

An object of class `tbl_df` (inherits from `tbl`, `data.frame`) with 45211 rows and 18 columns.

Bank Client Data:

- ID (chr): CUSTOMER ID
- AGE (dbl): Customer's age
- JOB (chr): Type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","s "blue-collar","self-employed","retired","technician","services")
- MARITAL (chr): marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
- EDUCATION (chr): categorical: "unknown","secondary","primary","tertiary"
- DEFAULT (chr): Has credit in default? (binary: "yes","no")
- BALANCE (dbl): Average yearly balance, in euros (numeric)
- HOUSING (chr): Has housing loan? (binary: "yes","no")
- LOAN (chr): Has personal loan? (binary: "yes","no")

Features related to the last contact during the current marketing campaign:

- CONTACT (chr): Contact communication type (categorical: "unknown","telephone","cellular")
- DAY (dbl): Last contact day of the month (numeric)
- MONTH (chr): Last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
- DURATION (dbl): Last contact duration, in seconds (numeric)

Additional Attributes:

- CAMPAIGN (dbl): Number of contacts performed during this campaign and for this client (numeric, includes last contact)
- PDAYS (dbl): Number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
- PREVIOUS (dbl): Number of contacts performed before this campaign and for this client (numeric)
- POUTCOME (chr): Outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

Target Variable (Response):

- TERM_DEPOSIT (chr): Has the client subscribed a term deposit? (binary: "yes","no")

Source

Moro et al., 2014 S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

`plot_correlation_funnel`

Plot a Correlation Funnel

Description

`plot_correlation_funnel` returns a correlation funnel visualization in either static (`ggplot2`) or interactive (`plotly`) formats.

Usage

```
plot_correlation_funnel(
  data,
  interactive = FALSE,
  limits = c(-1, 1),
  alpha = 1
)
```

Arguments

`data`	A `tibble` or `data.frame`
`interactive`	Returns either a static (`ggplot2`) visualization or an interactive (`plotly`) visualization
`limits`	Sets the X-Axis limits for the correlation space
`alpha`	Sets the transparency of the points on the plot.

Value

A static ggplot2 plot or an interactive plotly plot

See Also

binarize(), correlate()

Examples

```
library(dplyr)
library(correlationfunnel)

marketing_campaign_tbl %>%
    select(-ID) %>%
    binarize() %>%
    correlate(TERM_DEPOSIT__yes) %>%
    plot_correlation_funnel()
```

Index