vertica-ml-python1.0 Features

Flexible as Python, Fast and Scalable as Vertica

Ouali Badr

November 28, 2019

Executive Summary

This document is the `vertica-ml-python` API overview. It briefly explains each function and its importance. `vertica-ml-python` allows the user to use his Vertica Database with Python without loading the data in his personal machine first. All the functions execute requests directly in the database in order to gain in efficiency. It combines Vertica aggregations and Python flexibility to create objects similar to the ones available in `pandas` and `sklearn` with the power of a columnar oriented analytic database: Vertica.

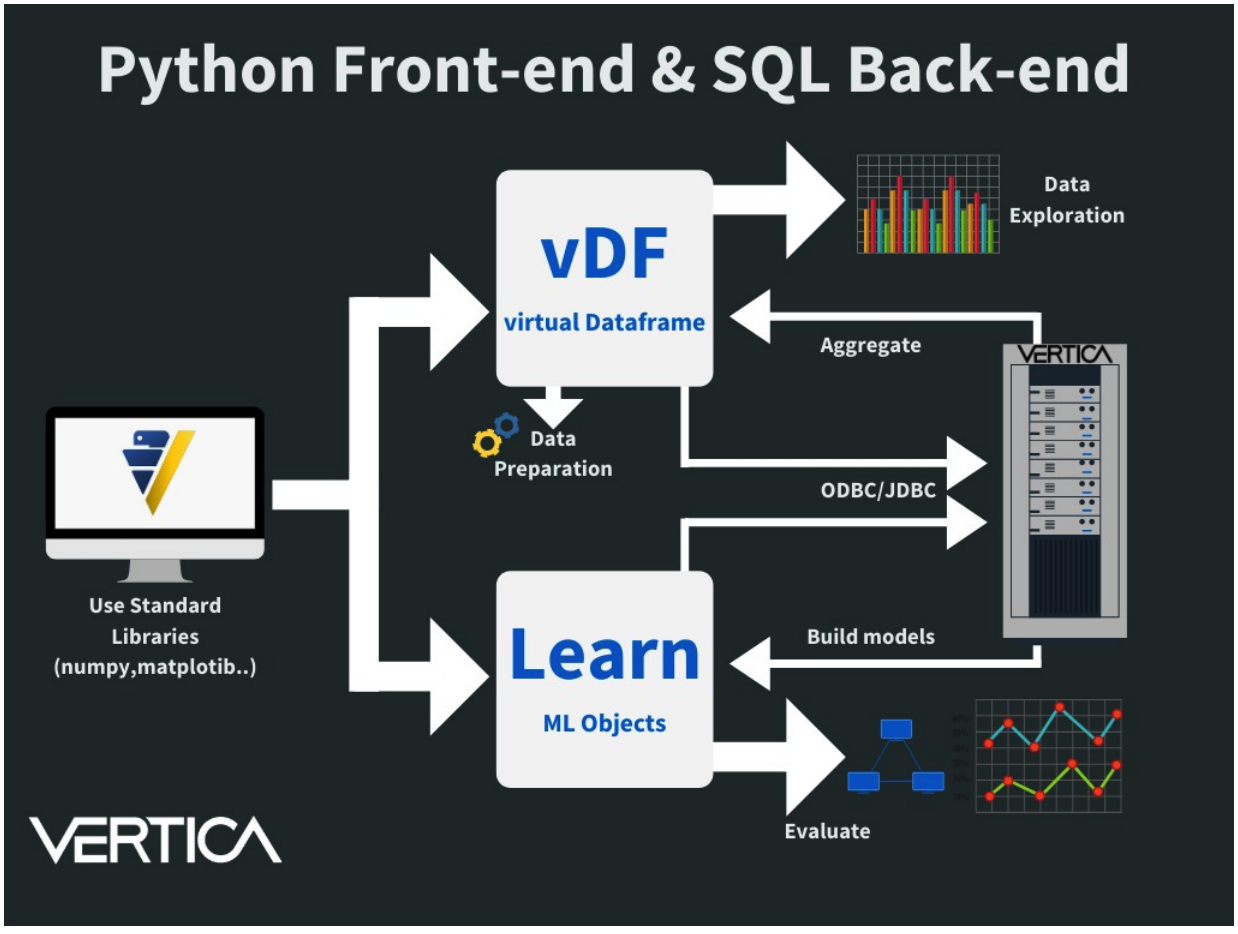
`vertica-ml-python` allows users to use the vDataframe (Virtual Dataframe). This object keeps in memory all the users modifications in order to use optimised SQL queries to compute all the necessary aggregations. Thanks to this object, the initial relation is intact and will never be modified. The purpose is to explore, preprocess and clean the object without changing the initial relation.

What contains vertica-ml-python?

This API contains many functions for:

- Data Exploration, Preprocessing and Cleaning: `vertica_ml_python.vdataframe`
- Machine Learning (Regression, Classification, Clustering): `vertica_ml_python.learn`

`vertica-ml-python` helps to explore, preprocess and clean the data without changing the initial relation. It uses scalable Machine Learning Algorithms such as Logistic Regression, Random Forest, SVM and much more... It allows also to evaluate and to optimise models (Classification/Regression Reports, ROC/PRC curves, Parameters tuning...).



Contents

"Science knows no country, because knowledge belongs to humanity, and is the torch which illuminates the world."

Louis Pasteur

The 'Big Data' (Tb of data) is now one of the main topics in the Data Science World. Data Scientists are now very important for any organisation. Becoming Data-Driven is mandatory to survive. Vertica is the first real analytic columnar Database and is still the fastest in the market. However, SQL is not enough flexible to be very popular for Data Scientists. Python flexibility is priceless and provides to any user a very nice experience. The level of abstraction is so high that it is enough to think about a function to notice that it already exists. Many Data Science APIs were created during the last 15 years and were directly adopted by the Data Science community (examples: pandas and scikit-learn).

However, Python is only working in-memory for a single node process. Even if some famous highly distributed programming languages exist to face this challenge, they are still in-memory and most of the time they can not process on all the data. Besides, moving the data can become very expensive. Data Scientists must also find a way to deploy their data preparation and their models. We are far away from easiness and the entire process can become time expensive. The idea behind VERTICA ML PYTHON is simple: Combining the Scalability of VERTICA with the Flexibility of Python to give to the community what they need *Bringing the logic to the data and not the opposite*. This version 1.0 is the work of 3 years of new ideas and improvement.

In this document the functions are placed in different categories to help the user to find the appropriate one for each problematic. A column 'Importance' was created to guide the user to the most useful functions (the one, he should consider often). The 'Importance' is going from ★ (not really important) to ★★★★★ (very important). For example, it is preferable to use the method `.describe(...)` or `.statistics(...)` to print all the descriptive statistics at the same moment rather than using the methods `.max(...)`, `.min(...)`, `.median(...)`...

Any Data Scientist will need to follow a part or the entire Data Science cycle to consider solving a data problem. Data Exploration and Business Understanding are very important to realise the most important step which is Data preparation. I decided to group the methods and functions using the Data Science cycle thematics to help any Data Scientist to find fast the proper function. If a function is missing, I'll be very happy to help you code it.

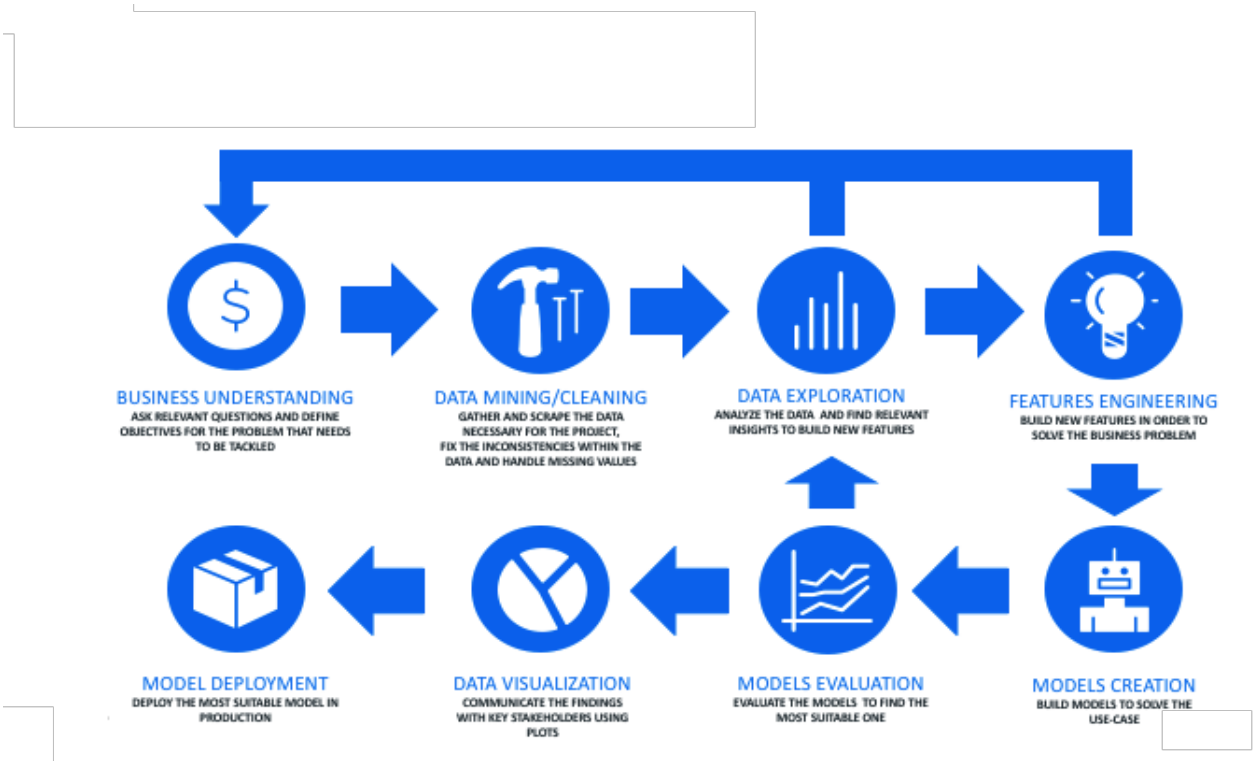


Figure 1: Data Science Life Cycle

1 Data Exploration and Understanding

1.1 Computation / Descriptive Stats

Function / Method	Library	Definition	Importance
vDataframe.aggregate(...)	vDataframe	Aggregate the vdf	★★★★★
vDataframe[].aggregate(...)	vDataframe	Aggregate the column	★★★
vDataframe.all(...)	vDataframe	Compute the bool and aggregation	★
vDataframe.any(...)	vDataframe	Compute the bool or aggregation	★
vDataframe.beta(...)	vDataframe	Compute the beta matrix	★★★
vDataframe.corr(...)	vDataframe	Compute the correlation matrix	★★★★★
vDataframe.cov(...)	vDataframe	Compute the covariance matrix	★★★
vDataframe.count(...)	vDataframe	Compute the count	★★★

Function / Method	Library	Definition	Importance
vDataframe[].count(...)	vDataframe	Compute the count	★★
vDataframe.describe(...)	vDataframe	Descriptive Statistics	★★★★★
vDataframe[].describe(...)	vDataframe	Descriptive Statistics	★★★★
vDataframe.kurtosis(...)	vDataframe	Compute the kurtosis	★★★
vDataframe[].kurtosis(...)	vDataframe	Compute the kurtosis	★★
vDataframe.mad(...)	vDataframe	Compute the median absolute deviation	★★★
vDataframe[].mad(...)	vDataframe	Compute the median absolute deviation	★★
vDataframe.max(...)	vDataframe	Compute the max	★★★
vDataframe[].max(...)	vDataframe	Compute the max	★★
vDataframe[].mode(...)	vDataframe	Compute the mode	★★
vDataframe.median(...)	vDataframe	Compute the median	★★★
vDataframe[].median(...)	vDataframe	Compute the median	★★
vDataframe.min(...)	vDataframe	Compute the min	★★★
vDataframe[].min(...)	vDataframe	Compute the min	★★
vDataframe[].numh(...)	vDataframe	Compute the column best histogram interval	★
vDataframe[].nunique(...)	vDataframe	Compute the column cardinality	★★
vDataframe.prod(...)	vDataframe	Compute the prod	★★★
vDataframe[].prod(....)	vDataframe	Compute the prod	★★
vDataframe.quantile(...)	vDataframe	Compute the quantiles	★★★★
vDataframe[].quantile(...)	vDataframe	Compute the quantile	★★
vDataframe.sem(...)	vDataframe	Compute the standard error of the mean	★★★
vDataframe[].sem(...)	vDataframe	Compute the standard error of the mean	★★
vDataframe.skewness(...)	vDataframe	Compute the skewness	★★★
vDataframe[].skewness(...)	vDataframe	Compute the skewness	★★
vDataframe.statistics(...)	vDataframe	Descriptive Statistics	★★★★★
vDataframe.std(...)	vDataframe	Compute the standard deviation	★★★
vDataframe[].std(...)	vDataframe	Compute the standard deviation	★★
vDataframe.sum(...)	vDataframe	Compute the sum	★★★
vDataframe[].sum(...)	vDataframe	Compute the sum	★★
vDataframe[].topk(...)	vDataframe	Compute the K most occurrent categories	★★★
vDataframe[].value_counts(...)	vDataframe	Categories count	★★
vDataframe.var(...)	vDataframe	Compute the variance	★★★
vDataframe[].var(...)	vDataframe	Compute the variance	★★

1.2 Plotting

Function / Method	Library	Definition	Importance
vDataframe.bar(...)	vDataframe	Draw the bar chart	★★★★
vDataframe[].bar(...)	vDataframe	Draw the bar chart	★★★★
vDataframe.boxplot(...)	vDataframe	Draw the box plot	★
vDataframe[].boxplot(...)	vDataframe	Draw the box plot	★

Function / Method	Library	Definition	Importance
vDataframe[].donut(...)	vDataframe	Draw the donut chart	★★
vDataframe[].density(...)	vDataframe	Draw the density plot	★★
vDataframe.hexbin(...)	vDataframe	Draw the hexbin plot	★★★
vDataframe.hist(...)	vDataframe	Draw the histogram	★★★★
vDataframe[].hist(...)	vDataframe	Draw the histogram	★★★
vDataframe[].pie(...)	vDataframe	Draw the pie chart	★★
vDataframe.pivot_table(...)	vDataframe	Draw the pivot table	★★★
vDataframe.plot(...)	vDataframe	Plot the time series	★★★★
vDataframe[].plot(...)	vDataframe	Plot the time series	★★★★★
vDataframe.scatter(...)	vDataframe	Draw the scatter plot	★★★★★
vDataframe.scatter_matrix(...)	vDataframe	Draw the scatter matrix	★★

1.3 Utilities

Function / Method	Library	Definition	Importance
vDataframe.dtypes(...)	vDataframe	Returns the data types	★★★
vDataframe[].dtype(...)	vDataframe	Returns the column data type	★
vDataframe[].distinct(...)	vDataframe	Returns the column unique elements	★★
vDataframe.head(...)	vDataframe	Print the head	★
vDataframe[].head(...)	vDataframe	Print the head	★
vDataframe.isin(...)	vDataframe	Verify if the elements are in the object	★
vDataframe[].isin(...)	vDataframe	Verify if the elements are in the column	★
vDataframe.nlargest(...)	vDataframe	Returns the n smallest column elements	★
vDataframe.nsmallest(...)	vDataframe	Returns the n largest column elements	★
vDataframe[].rename(...)	vDataframe	Rename the column	★
vDataframe.tail(...)	vDataframe	Print a part of the object	★★
vDataframe[].tail(...)	vDataframe	Print a part of the object	★★

2 Data Preparation

2.1 Binary Operator Functions

Function / Method	Library	Definition	Importance
vDataframe[].add(...)	vDataframe	Add an element to the column	★
vDataframe[].div(...)	vDataframe	Divide the column by an element	★
vDataframe[].equals(...)	vDataframe	Verify an equality	★
vDataframe[].ge(...)	vDataframe	Verify a ge inequality	★
vDataframe[].gt(...)	vDataframe	Verify a gt inequality	★

Function / Method	Library	Definition	Importance
vDataframe[].le(...)	vDataframe	Verify a le inequality	★
vDataframe[].lt(...)	vDataframe	Verify a lt inequality	★
vDataframe[].mod(...)	vDataframe	Apply the modulo on the column	★
vDataframe[].mul(...)	vDataframe	Multiply the column by an element	★
vDataframe[].neq(...)	vDataframe	Verify an equality	★
vDataframe[].pow(...)	vDataframe	Apply the power function on the column	★
vDataframe[].round(...)	vDataframe	Round the column elements	★
vDataframe[].sub(...)	vDataframe	Subtract an element from the column	★

2.2 Combining / Joining / Sorting

Function / Method	Library	Definition	Importance
vDataframe.append(...)	vDataframe	Merge a vDataframe with another relation	★★
vDataframe.groupby(...)	vDataframe	Group the vDataframe elements	★★★★★
vDataframe.join(...)	vDataframe	Join a vDataframe with another relation	★★★★★
vDataframe.rank(...)	vDataframe	Compute a specific rank	★
vDataframe.sort(...)	vDataframe	Sort the data	★★★

2.3 Data Type Conversion

Function / Method	Library	Definition	Importance
vDataframe.astype(...)	vDataframe	Convert to the selected types	★★★
vDataframe[].astype(...)	vDataframe	Convert the column to the selected types	★★★
vDataframe[].to_enum(...)	vDataframe	Convert the column to categorical	★★★
vDataframe[].to_timestamp(...)	vDataframe	Convert the column to timestamp	★

2.4 Decomposition / Normalization / Preprocessing

Function / Method	Library	Definition	Importance
Balance	learn.preprocessing	Balance the data	★★★
vDataframe[].decode(...)	vDataframe	User-defined encoding	★★★★
vDataframe.drop(...)	vDataframe	Drop the columns	★★
vDataframe[].drop(...)	vDataframe	Drop the column	★★
vDataframe[].get_dummies(...)	vDataframe	One Hot Encoder Encoding	★★★★★
vDataframe[].label_encode(...)	vDataframe	Apply a Label Encoding on the column	★★★
vDataframe[].mean_encode(...)	vDataframe	Apply a Mean Encoding on the column	★★★

Function / Method	Library	Definition	Importance
vDataframe.normalize(...)	vDataframe	Normalize the data	★★
vDataframe[].normalize(...)	vDataframe	Normalize the column	★★★★★
Normalizer	learn.preprocessing	Normalize the data	★★
OneHotEncoder	learn.preprocessing	One Hot Encoder Encoding	★★
PCA	learn.decomposition	Apply the PCA	★★★★★
vDataframe.sample(...)	vDataframe	Sample the data	★
vDataframe.save(...)	vDataframe	Save the vDataframe structure	★★★★
SVD	learn.decomposition	Apply the SVD	★★

2.5 Features Engineering / Computation

Function / Method	Library	Definition	Importance
vDataframe.abs(...)	vDataframe	Apply the Absolute function	★
vDataframe[].abs(...)	vDataframe	Apply the Absolute function	★
vDataframe[].add_copy(...)	vDataframe	Add a copy of the column	★★★
vDataframe.apply(...)	vDataframe	Apply functions	★★★★
vDataframe.apply[](...)	vDataframe	Apply a function	★★★★★
vDataframe.applymap(...)	vDataframe	Apply a function	★★
vDataframe.eval(...)	vDataframe	Evaluate an expression	★★★★★

2.6 Handling Missing Values, Outliers and Duplicates / Filtering

Function / Method	Library	Definition	Importance
vDataframe[].clip(...)	vDataframe	Clip the column	★★
vDataframe.drop_duplicates(...)	vDataframe	Drop the duplicates	★★
vDataframe.dropna(...)	vDataframe	Drop the missing values	★★
vDataframe[].dropna(...)	vDataframe	Drop the column missing values	★★★★
vDataframe[].drop_outliers(...)	vDataframe	Drop the column outliers	★★★
vDataframe.duplicated(...)	vDataframe	Returns the duplicates	★★★★
vDataframe.filter(...)	vDataframe	Filtering the data	★★★★★
vDataframe.fillna(...)	vDataframe	Fill the missing values	★★
vDataframe[].fillna(...)	vDataframe	Fill the column missing values	★★★★★
vDataframe[].fill_outliers(...)	vDataframe	Fill the column outliers	★★
vDataframe.select(...)	vDataframe	Select only specific columns	★

3 Machine Learning

3.1 Anomaly Detection / Clustering

Function / Method	Library	Definition	Importance
DBSCAN	learn.cluster	DBSCAN Algorithm	★★★★★
KMeans	learn.cluster	KMeans Algorithm	★★★★★
LocalOutlierFactor	learn.neighbors	Local Outlier Factor	★★★★★

3.2 Classification

3.2.1 Binary Classifiers

Function / Method	Library	Definition	Importance
LinearSVC	learn.svm	Linear SVM	★★
LogisticRegression	learn.linear_model	Logistic Regression	★★★★★

3.2.2 Evaluation

Function / Method	Library	Definition	Importance
accuracy_score	learn.metrics	Compute the Accuracy	★★
auc	learn.metrics	Compute the ROC AUC	★★
classification_report	learn.metrics	Compute many classification metrics	★★★★★
confusion_matrix	learn.metrics	Compute the Confusion Matrix	★★
critical_success_index	learn.metrics	Compute the Critical Success Index	★★
f1_score	learn.metrics	Compute the F1 Score	★★
informedness	learn.metrics	Compute the Informedness	★★
lift_chart	learn.plot	Draw the Lift Chart	★★
log_loss	learn.metrics	Compute the Log Loss	★★
markedness	learn.metrics	Compute the Markedness	★★
matthews_corrcoef	learn.metrics	Compute the Matthews correlation coefficient	★★
multilabel_confusion_matrix	learn.metrics	Compute the Multi Label Confusion Matrix	★★
negative_predictive_score	learn.metrics	Compute the Negative Predictive Score	★★
prc_auc	learn.metrics	Compute the PRC AUC	★★
prc_curve	learn.plot	Draw the PRC Curve	★★
precision_score	learn.metrics	Compute the Precision	★★
recall_score	learn.metrics	Compute the Recall	★★
roc_curve	learn.plot	Draw the ROC Curve	★★
specificity_score	learn.metrics	Compute the Specificity	★★

3.2.3 Multinomial Classifiers

Function / Method	Library	Definition	Importance
DecisionTreeClassifier	learn.tree	Decision Tree	★
DummyTreeClassifier	learn.tree	Dummy Classifier	★
KNeighborsClassifier	learn.neighbors	K Nearest Neighbors	★★★
MultinomialNB	learn.naive_bayes	Naive Bayes	★★
NearestCentroid	learn.neighbors	Nearest Centroid	★★
RandomForestClassifier	learn.ensemble	Random Forest	★★★★★

3.3 Model Selection

Function / Method	Library	Definition	Importance
best_k	learn.model_selection	Find the best KMeans K	★★★
cross_validate	learn.model_selection	K-Fold Cross Validation	★★★★★
elbow	learn.plot	Draw the Elbow curve	★★★
train_test_split	learn.model_selection	Split the data	★★
NearestCentroid	learn.model_selection	Nearest Centroid	★★
RandomForestClassifier	learn.ensemble	Random Forest	★★★★★

3.4 Regression

3.4.1 Algorithms

Function / Method	Library	Definition	Importance
DecisionTreeRegressor	learn.tree	Decision Tree	★
DummyTreeRegressor	learn.tree	Dummy Regressor	★
ElasticNet	learn.linear_model	Elastic Net Regression	★★★★★
KNeighborsRegressor	learn.neighbors	K Nearest Neighbors	★★★
Lasso	learn.linear_model	Lasso Regression	★★
LinearRegression	learn.linear_model	Linear Regression	★★
LinearSVR	learn.svm	Linear SVM	★★
RandomForestRegressor	learn.ensemble	Random Forest	★★★★★
Ridge	learn.linear_model	Ridge Regression	★★

3.4.2 Evaluation

Function / Method	Library	Definition	Importance
explained_variance	learn.metrics	Compute the explained variance	★★
max_error	learn.metrics	Compute the max residual error.	★★
median_absolute_error	learn.metrics	Compute the median absolute error	★★
mean_absolute_error	learn.metrics	Compute the average absolute error	★★
mean_squared_error	learn.metrics	Compute the average squared error	★★
mean_squared_log_error	learn.metrics	Compute the average squared log error	★★
regression_report	learn.metrics	Compute many regression metrics	★★★★★
r2_score	learn.metrics	Compute the R2 Score	★★

4 Time Series

Function / Method	Library	Definition	Importance
vDataframe.asfreq(...)	vDataframe	Interpolate, slice and agg the time series	★★★★★
vDataframe.at_time(...)	vDataframe	Filter the data at the specific time	★
vDataframe.between_time(...)	vDataframe	Filter the data at the specific time range	★★
vDataframe.cummax(...)	vDataframe	Compute the cumulative max	★★★
vDataframe.cummin(...)	vDataframe	Compute the cumulative min	★
vDataframe.cumprod(...)	vDataframe	Compute the cumulative prod	★
vDataframe.cumsum(...)	vDataframe	Compute the cumulative sum	★★★
vDataframe[].date_part(...)	vDataframe	Extract the date field	★★
vDataframe[].ema(...)	vDataframe	Exponential Moving Average	★★★
vDataframe.first(...)	vDataframe	Filter by keeping only the first elements	★★
vDataframe.last(...)	vDataframe	Filter by keeping only the last elements	★★
vDataframe[].next(...)	vDataframe	Compute and apply the lead	★★
vDataframe[].pct_change(...)	vDataframe	Time series percent change	★★
vDataframe[].prev(...)	vDataframe	Compute and apply the lag	★★
vDataframe.rolling(...)	vDataframe	Compute a moving window	★★★★★
vDataframe.sessionize(...)	vDataframe	Build a session ID	★★★★★
vDataframe[].slice(...)	vDataframe	Slice the time series	★★

5 Utilities

Function / Method	Library	Definition	Importance
vDataframe.catcol(...)	vDataframe	Returns the categorical columns	★
vDataframe[].category(...)	vDataframe	Returns the column category	★
vDataframe.copy(...)	vDataframe	Returns a vDataframe copy	★★
drop_model	utilities	Drop the model	★
drop_table	utilities	Drop the table	★
drop_text_index	utilities	Drop the text index	★
drop_view	utilities	Drop the view	★
vDataframe.dsn_restart(...)	vDataframe	Restart the DSN	★
vDataframe.empty(...)	vDataframe	Returns if the vDataframe is empty	★
vDataframe.expected_store_usage(...)	vDataframe	Returns the expected store usage	★★
vDataframe.get_columns(...)	vDataframe	Returns the vDataframe columns	★★
vDataframe.help(...)	vDataframe	Get info about the API	★
vDataframe.info(...)	vDataframe	Get vDataframe history	★★
vDataframe[].isdate(...)	vDataframe	Verify if the column is a date	★
vDataframe[].isnum(...)	vDataframe	Verify if the column is numerical	★
vDataframe.load(...)	vDataframe	Load a saving	★★★
load_iris	learn.datasets	Load the Iris dataset	★
load_model	utilities	Load the model	★★★
load_smart_meters	learn.datasets	Load the Smart Meters dataset	★
load_titanic	learn.datasets	Load the Titanic dataset	★
load_winequality	learn.datasets	Load the Wine Quality dataset	★
vDataframe.memory_usage(...)	vDataframe	Returns the object memory usage	★
vDataframe.numcol(...)	vDataframe	Returns the numerical columns	★
pandas_to_vertica	utilities	Store a pandas.vDataframe in Vertica	★
vDataframe.set_cursor(...)	vDataframe	Set a new cursor	★★
vDataframe.set_dsn(...)	vDataframe	Set a new DSN	★★
vDataframe.shape(...)	vDataframe	Returns the vDataframe shape	★★
vDataframe.sql_on_off(...)	vDataframe	Display the queries	★★★
vDataframe.time_on_off(...)	vDataframe	Display the computation time	★★
vDataframe.to_csv(...)	vDataframe	Write a csv file of the data	★★★
vDataframe.to_db(...)	vDataframe	Save the vDataframe in Vertica	★★★
vDataframe.to_pandas(...)	vDataframe	Convert to a pandas.vDataframe	★
vDataframe.to_vdf(...)	vDataframe	Write the vdf file of the vDataframe	★★★
to_tablesample	utilities	Query result to tablesample	★★
vDataframe.version(...)	vDataframe	Returns Vertica version	★
vertica_cursor	utilities	Create a Vertica cursor	★★★

5.1 Working with Text

Function / Method	Library	Definition	Importance
vDataframe[].add_prefix(...)	vDataframe	Add a prefix to the column	★
vDataframe[].add_suffix(...)	vDataframe	Add a suffix to the column	★
CountVectorizer	learn.preprocessing	Compute the Text Dictionary	★★★★★
vDataframe[].str_contains(...)	vDataframe	Verify if the pattern is in the column	★
vDataframe[].str_count(...)	vDataframe	Count the pattern occurences	★★
vDataframe[].str_replace(...)	vDataframe	Replace the pattern	★★
vDataframe[].str_slice(...)	vDataframe	Slice the column	★