Vertica ML Python Workshop

Exercise 3: Correlation & Dependance

Ouali Badr

December 5, 2019

Executive Summary

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

VERTICA ML PYTHON allows the users to use Vertica advanced analytics and Machine Learning with a Python front-end Interface. In this exercise, you'll learn some basics to begin your fantastic Data Science Journey with the API. As a summary:

- Draw correlation matrix
- Understand the difference between dependance and correlation
- Understand how to use the correct correlation method

Contents

1 Presentation

Correlation is any statistical relationship, whether causal or not, between two variables. Handling correlated features is one of the most important part of Data Preparation. Correlation can help to understand the relationships between the variables and so to choose the proper Machine Learning algorithms. However, highly correlated predictors can add bias to the data which can lead to unexpected results. Most of the time, the purpose is to find the main source of the causation. Correlation doesn't mean causation but it can be taken as evidence for a possible causal relationship. There is no perfect correlation function. Many correlation functions exist and they can be used for different purposes. We can identify 5 main correlation functions:

- **Pearson:** Linear. It works for two numerical variables. It is the most famous correlation coefficient. It identifies the linear relationship between the variables.
- **Spearman:** Monotonic. It works for two numerical variables. It is bringing much more information than the Pearson coefficient. It is computing the monotonic relationship between the variables.
- **Kendall:** Complex. It works for two numerical variables. It is a more complete coefficient which can identify complex relationship between the variables.
- **Cramer's V:** Categorical. It works for two categorical variables. This coefficient will always be positive. It is as datasets will mainly be composed of categorical features.
- **Biserial Point:** Binary. It works for a numerical variable and a binary variable. We can then decompose a categorical variable using a One Hot Encoder and understanding the influence of each category on a numerical feature.

During this exercise, we will compute different correlation techniques. We will use the 'churn' telco dataset. This dataset contains many information of 7043 customers including:

- **customerID:** Customer ID
- **gender:** Whether the customer is a male or a female
- **SeniorCitizen:** Whether the customer is a senior citizen or not (1, 0)
- **Partner:** Whether the customer has a partner or not (Yes, No)
- **Dependents:** Whether the customer has dependents or not (Yes, No)
- **tenure:** Number of months the customer has stayed with the company
- **PhoneService:** Whether the customer has a phone service or not (Yes, No)
- **MultipleLines:** Whether the customer has multiple lines or not (Yes, No, No phone service)
- **InternetService:** Customer's internet service provider (DSL, Fiber optic, No)
- **OnlineSecurity:** Whether the customer has online security or not (Yes, No, No internet service)
- **OnlineBackup:** Whether the customer has online backup or not (Yes, No, No internet service)
- **DeviceProtection:** Whether the customer has device protection or not (Yes, No, No internet service)
- **TechSupport:** Whether the customer has tech support or not (Yes, No, No internet service)
- **StreamingTV:** Whether the customer has streaming TV or not (Yes, No, No internet service)
- **StreamingMovies:** Whether the customer has streaming movies or not (Yes, No, No internet service)
- **Contract:** The contract term of the customer (Month-to-month, One year, Two year)

- **PaperlessBilling:** Whether the customer has paperless billing or not (Yes, No)
- **PaymentMethod:** The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
- **MonthlyCharges:** The amount charged to the customer monthly
- **TotalCharges:** The total amount charged to the customer
- **Churn:** Whether the customer churned or not (Yes or No)

The purpose is to find identify relationships between the different variables to evaluate what could possibly influence a customer to churn.

2 Functions used during the Exercise

2.1 corr

Library: vertica_ml_python.vDataframe

```
vDataframe.corr(
        self,
        columns: list = [],
        method: str = "pearson",
        cmap: str = "",
        round_nb: int = 3,
        show: bool = True)
```

Compute the correlation matrix of the vDataframe.

Parameters

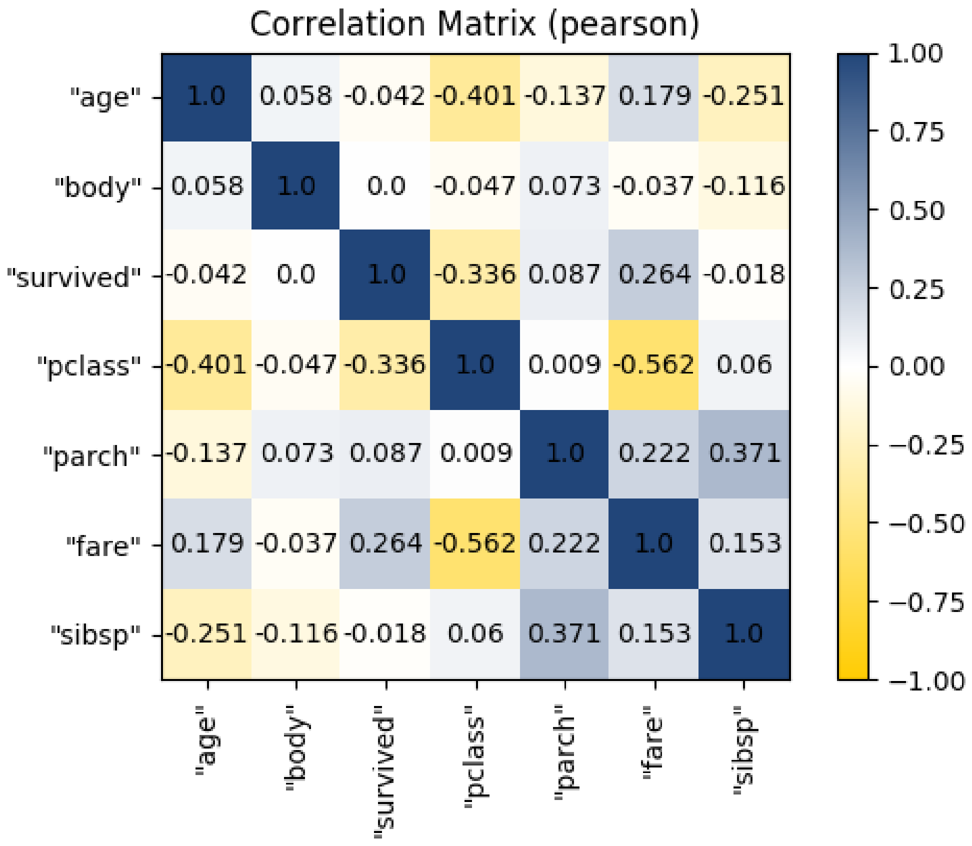
- **columns:** *<list>*, optional
  List of the vDataframe columns. If this parameter is empty, the method will consider all the numerical columns.
- **method:** *<str>*, optional
  The method must be in pearson (Pearson Coefficient) | kendall (Kendall Coefficient) | spearman (Spearman Coefficient) | biserial (Biserial Point) | cramer (Cramer'sV)
- **cmap:** *<str>*, optional
  Color Maps.
- **round_nb:** *<int>*, optional
  Integer used to round the numerical values.
- **show:** *<bool>*, optional
  Display the result using matplotlib.

Returns

The `tablesample` type containing the matrix (the information will be stored in the `values` attribute). You can convert this object to pandas using the `to_pandas` method or to vDataframe using the `to_vdf` method.

Example

```
from vertica_ml_python.learn.datasets import load_titanic
titanic = load_titanic(cur)
titanic.corr()
```



3 Questions

Turn on Jupyter with the 'jupyter notebook' command. Start the notebook exercise3.ipynb and answer to the following questions.

- **Question 1:** The linear correlation is computed using the 'pearson' method. Draw the Correlation Matrix using the Pearson coefficient. Explain the linear correlations and dependences between the different features.
- **Question 2:** There is no perfect way to deal with non-linear correlations between the different features. Spearman coefficients will identify Monotonic relationships between variables. Draw the Correlation Matrix using the Spearman coefficient. Explain the monotonic relationships between the different features.
- **Question 3:** Kendall coefficients can be very expensive to compute but they will bring very useful information. They can explain complex relationship between the different features. Draw the Correlation Matrix using the Kendall coefficient. Explain the relationships between the different features.

- **Question 4:** The point Biserial method explains the link between a binary variable and a numerical feature. It can be used only if one of the feature is binary, otherwise it will return 0. Draw the Correlation Matrix using the the point Biserial. In this use-case this method is not very efficient, however do not forget that it is always a possibility.
- **Question 5:** The last technique is the Cramer's V which can explain link between categorical columns. It will range only between 0 and 1. This technique is very important as it helps to understand the link between the categorical features which can represent a huge part of the dataset. Draw the Correlation Matrix using the Cramer's V coefficient. Explain the relationships between the different features.
- **Question 6:** Explain which features will be useful to predict churn and what type of ML algorithms you will probably not use for the prediction.