Vertica ML Python Library

Badr Ouali

Datascientist, Vertica

About Vertica-ML-Python

- It started as an internship project …

- It's a Python front end that exposes sci-kit like functionality and pushes all the (heavy) computation to Vertica

- Can be used for both ML and data preparation

- A 'pipeline' mechanism allow to sequentialize several data transformation operations in an object called RVD (Resilient Vertica Dataset)

- Offers multiple rendering capabilities

VERTICA

VERTICA PYTHON

MICRO FOCUS

Why Vertica-ML-Python?

- Vertica offers the possibility to do fast ML and data preparation using SQL

- Python is flexible and offers the possibility to create objects easily and to call very easy method (high level of abstraction).

- Vertica-ML-Python uses direct sql queries to the Vertica Database in order to do easily data preparation, exploration and Machine Learning. The methods are very simple to learn and can be called easily: the user is using Vertica as easily as Pandas library.
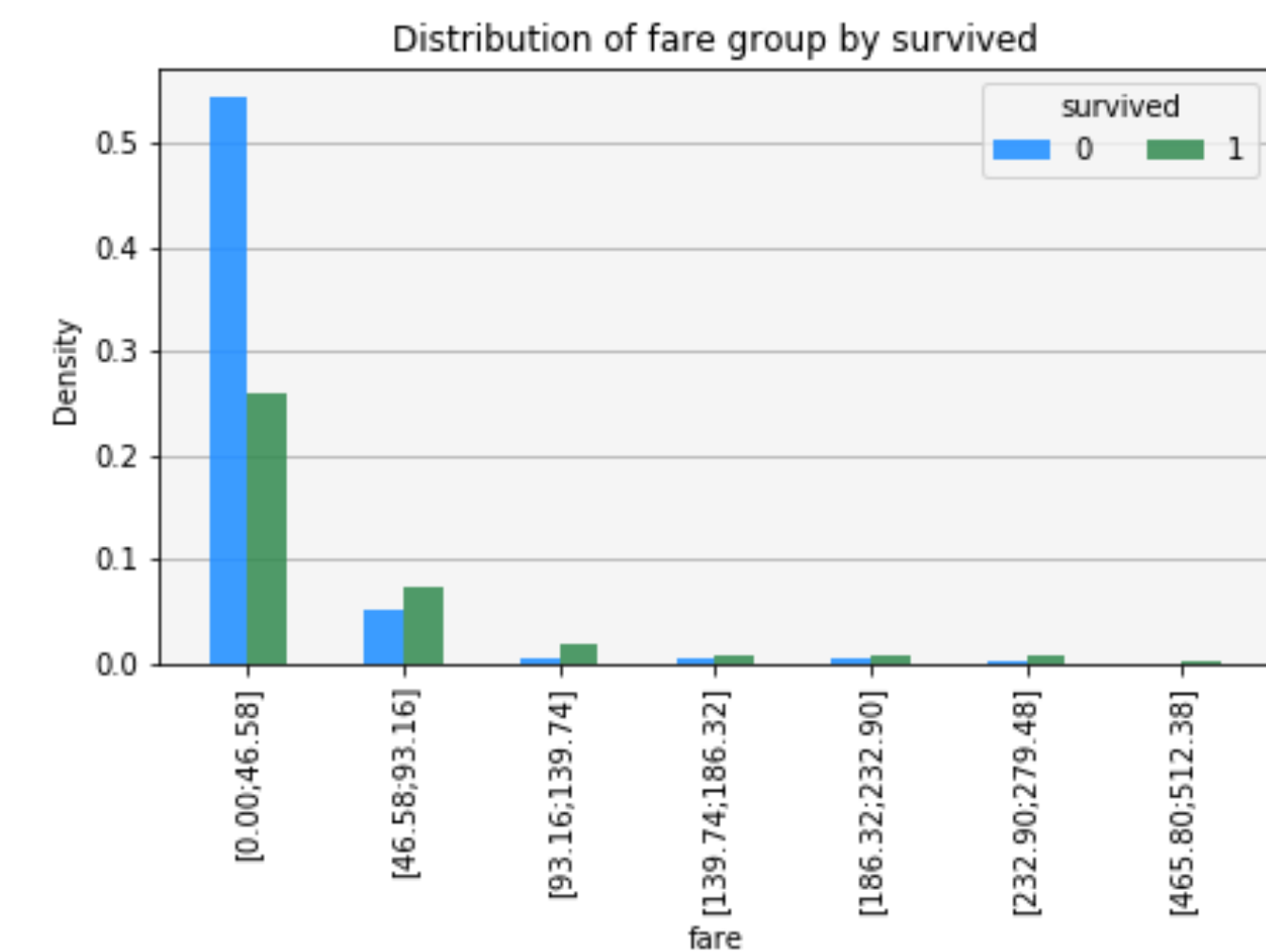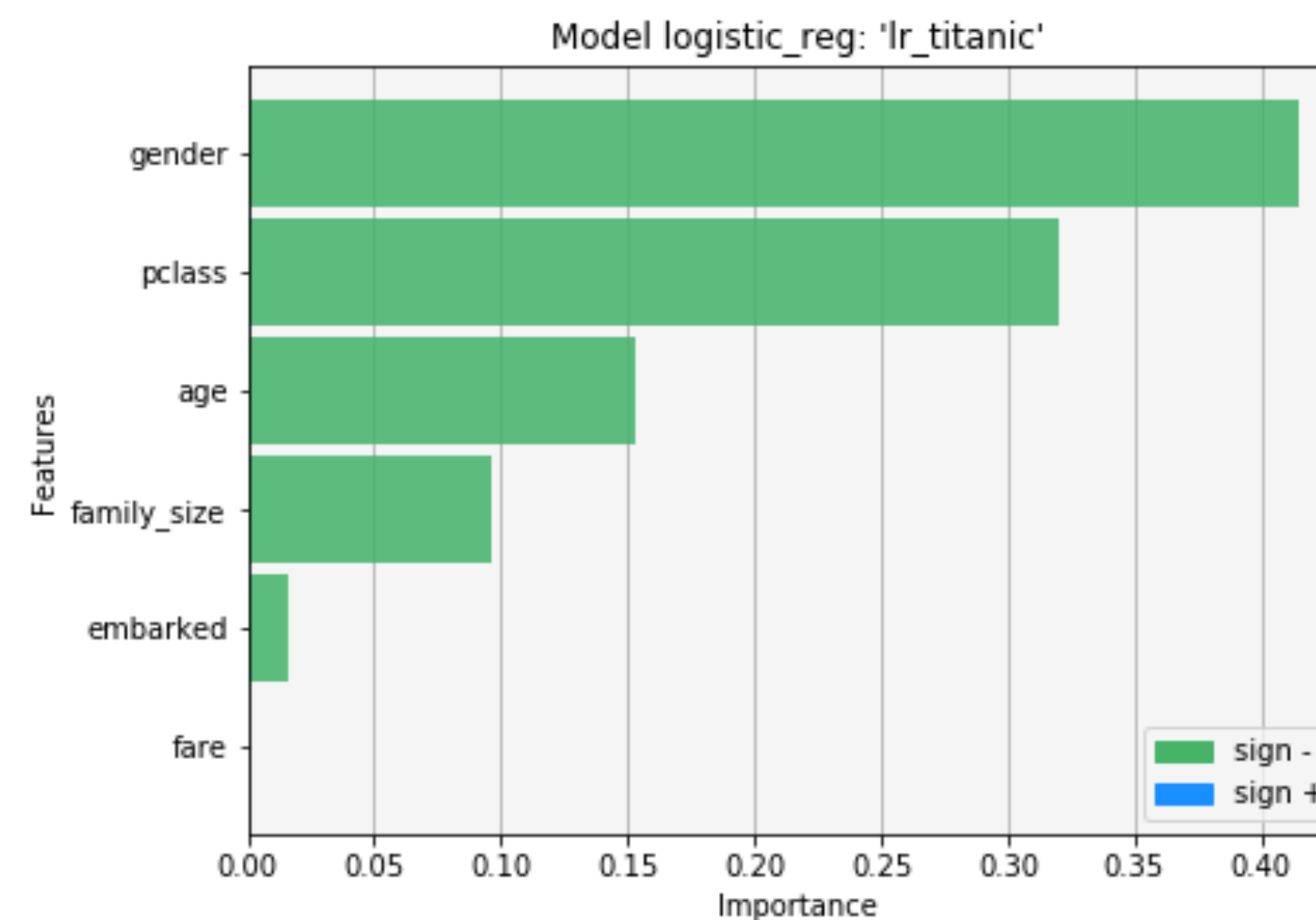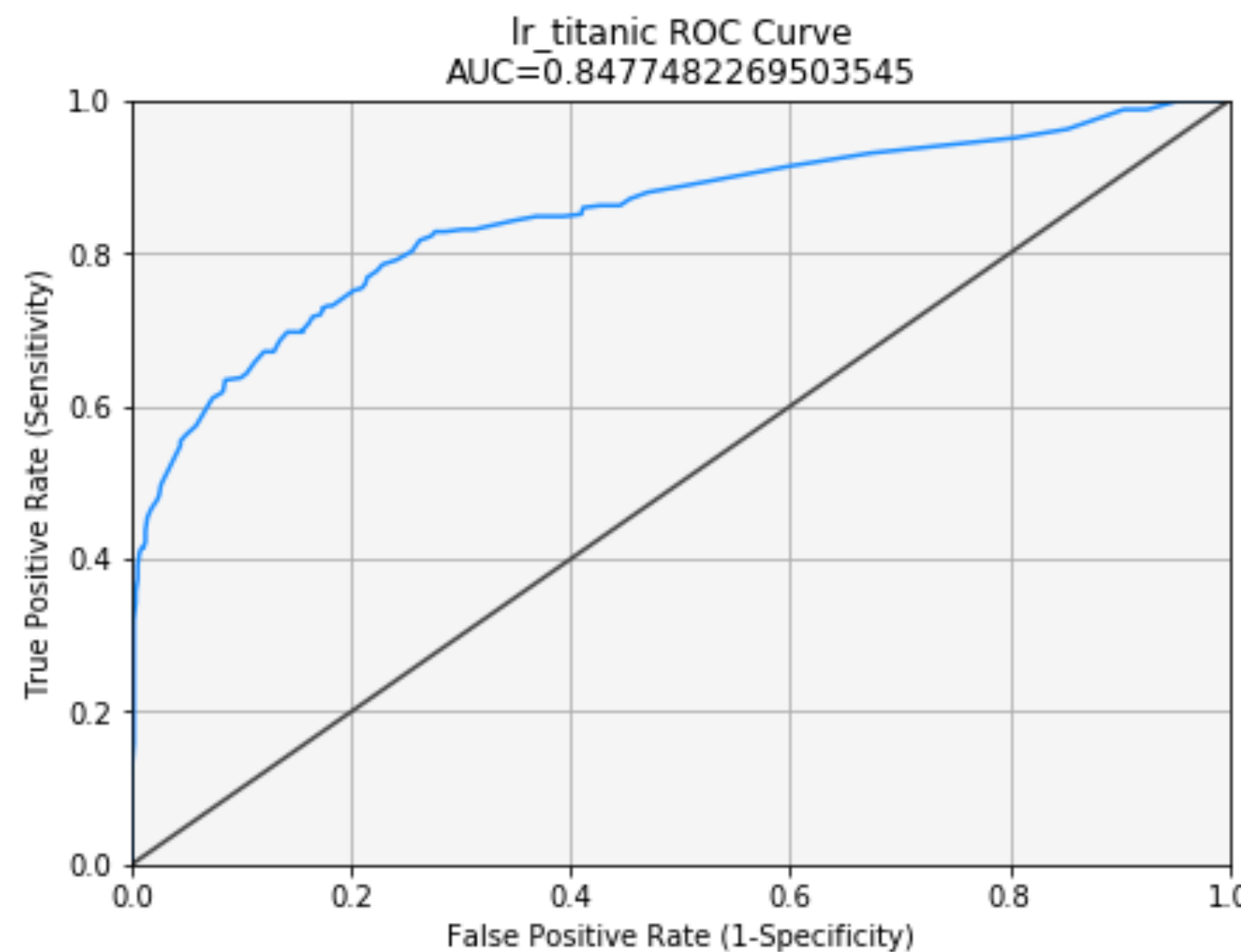

VERTICA PYTHON

VERTICA

MICRO FOCUS

What is Vertica-ML-Python?

- Vertica-ML-Python offers the possibility to create RVD (Resilient Vertica Dataset) which are similar in the use to the pandas.Dataframes but they use all the power of Vertica.

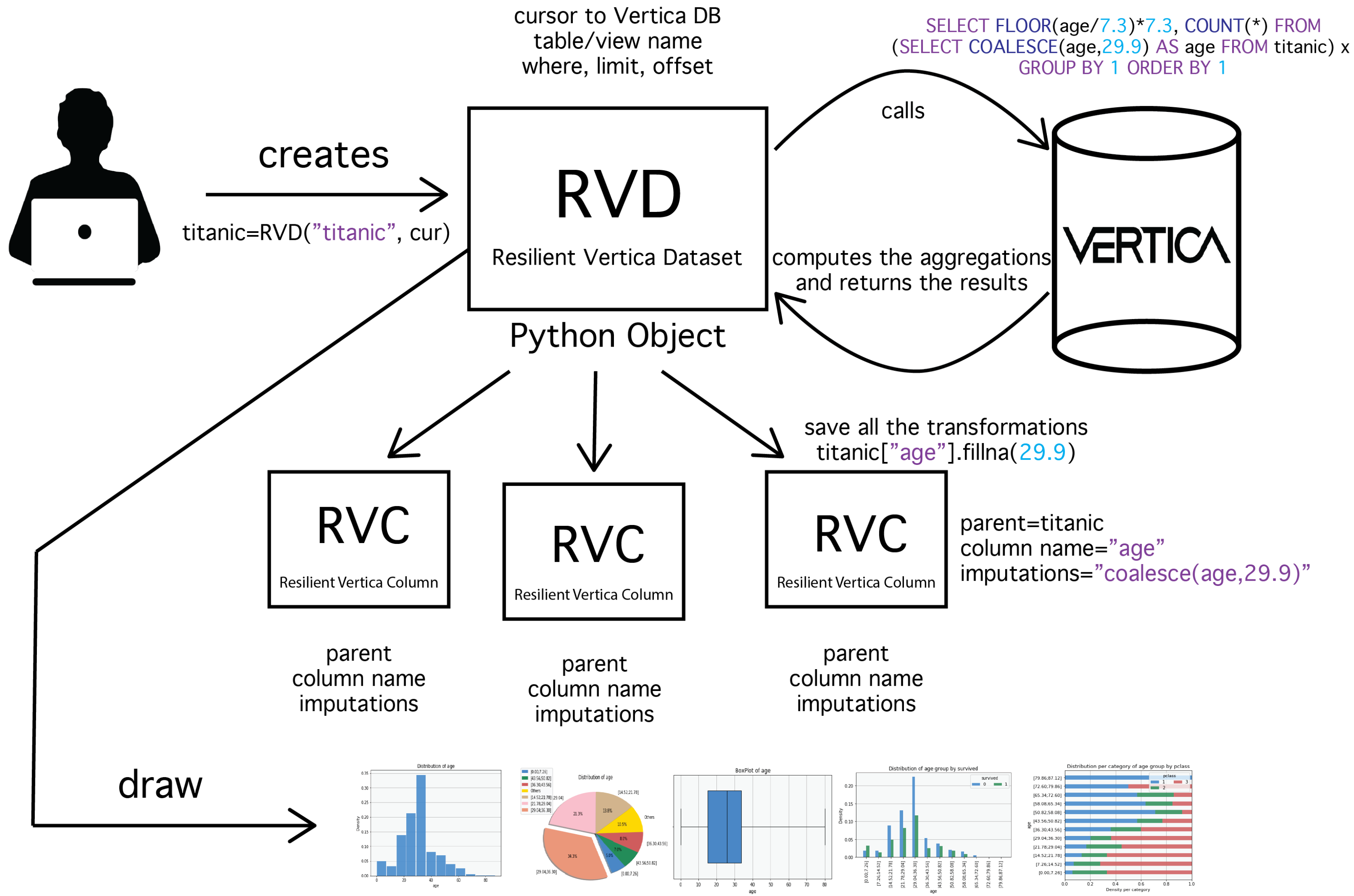- It offers also the possibility to create ML objects in order to see their performances easily.

VERTICA PYTHON

Resilient Vertica Dataset
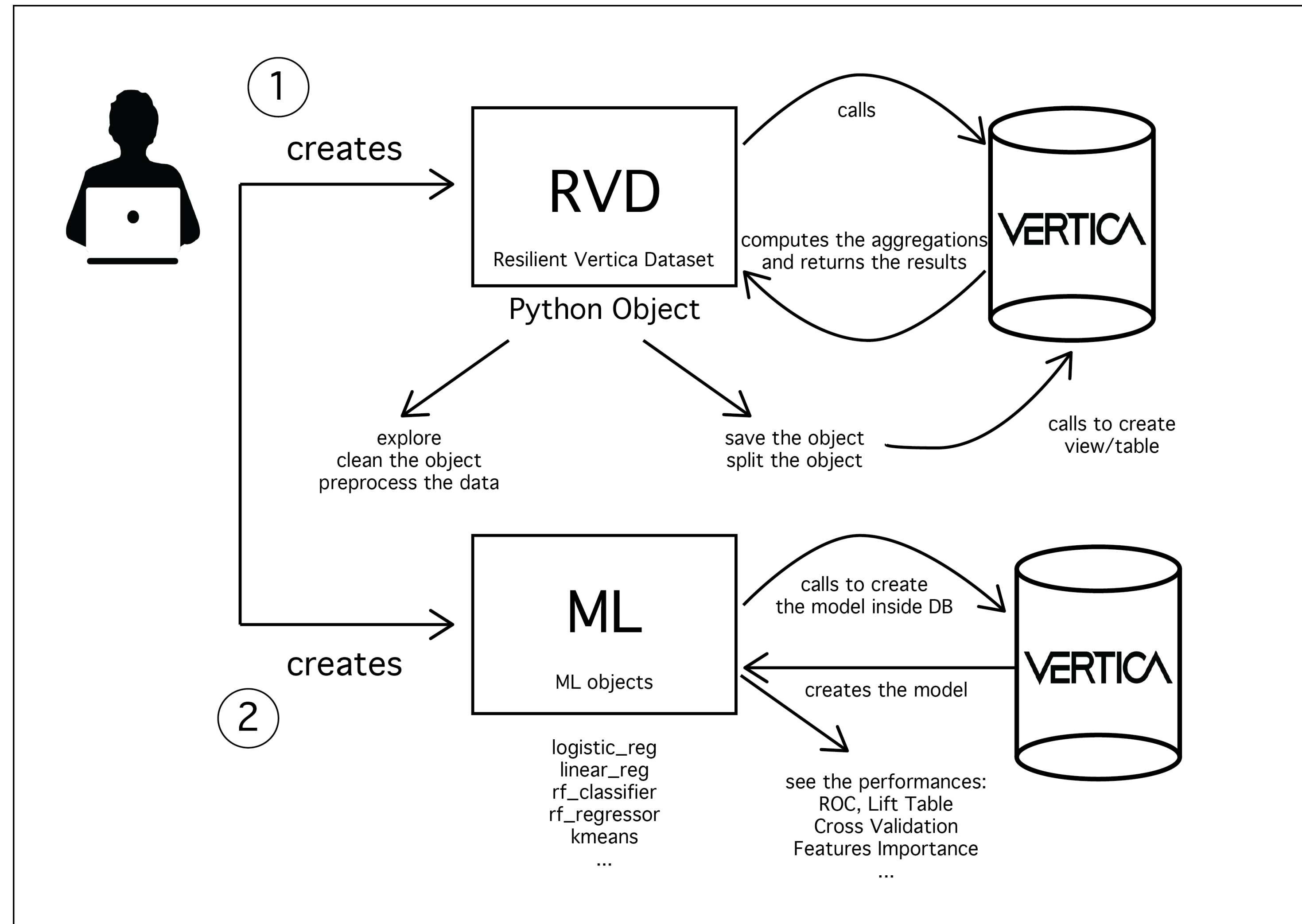
Resilient Vertica Dataset (RVD)

cursor to Vertica DB
table/view name
where, limit, offset

SELECT FLOOR(age/7.3)*7.3, COUNT(*) FROM
(SELECT COALESCE(age,29.9) AS age FROM titanic) x
GROUP BY 1 ORDER BY 1

creates

titanic=RVD("titanic", cur)

RVD

Resilient Vertica Dataset

calls

computes the aggregations
and returns the results

Python Object

save all the transformations
titanic["age"].fillna(29.9)

RVC

Resilient Vertica Column

RVC

Resilient Vertica Column

RVC

Resilient Vertica Column

parent=titanic
column name="age"
imputations="coalesce(age,29.9)"

parent
column name
imputations

parent
column name
imputations

parent
column name
imputations

draw

VERTICA PYTHON

Vertica Machine Learning

Vertica Machine Learning (VML)
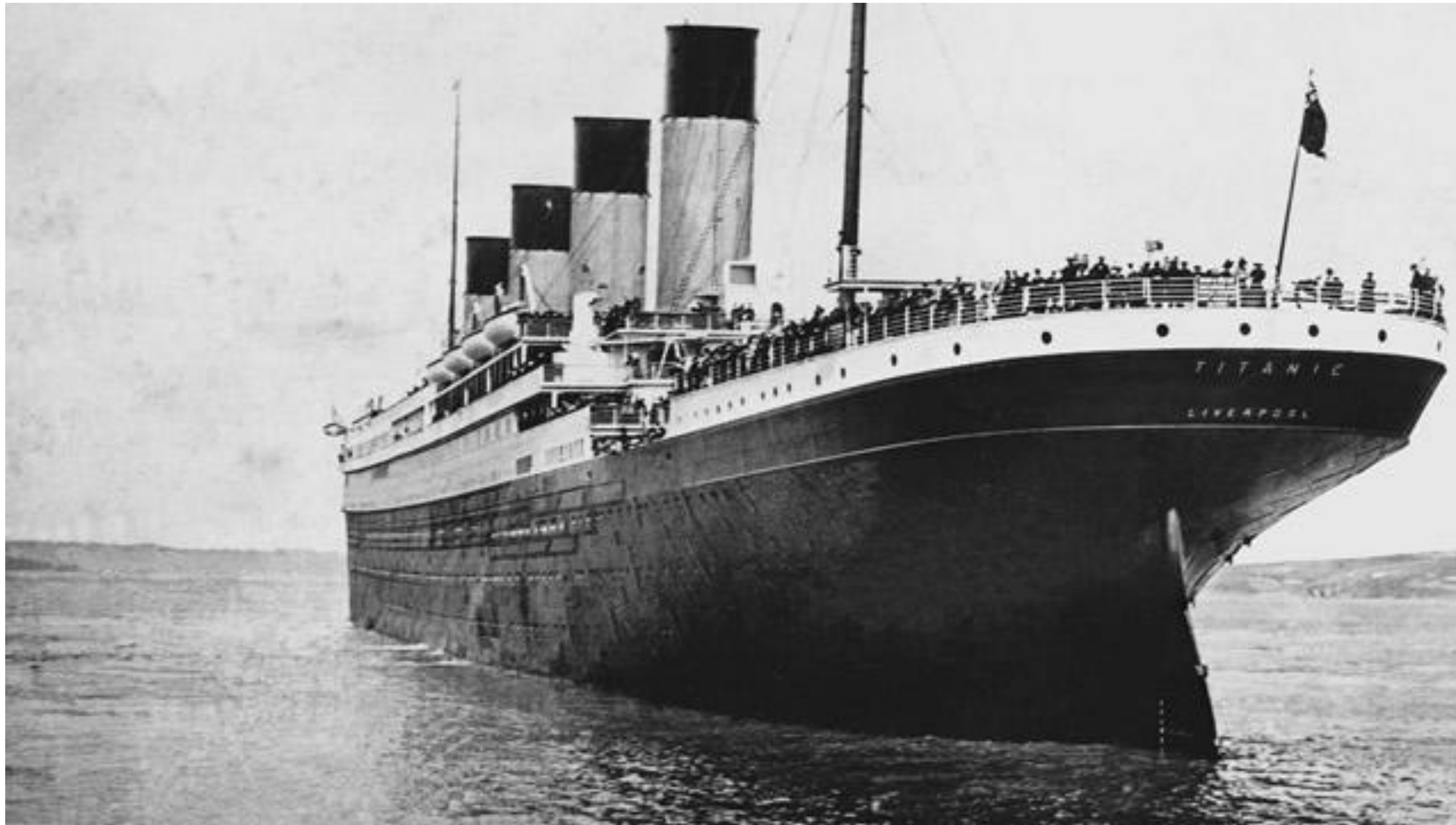
An example

A VPython example: Titanic dataset

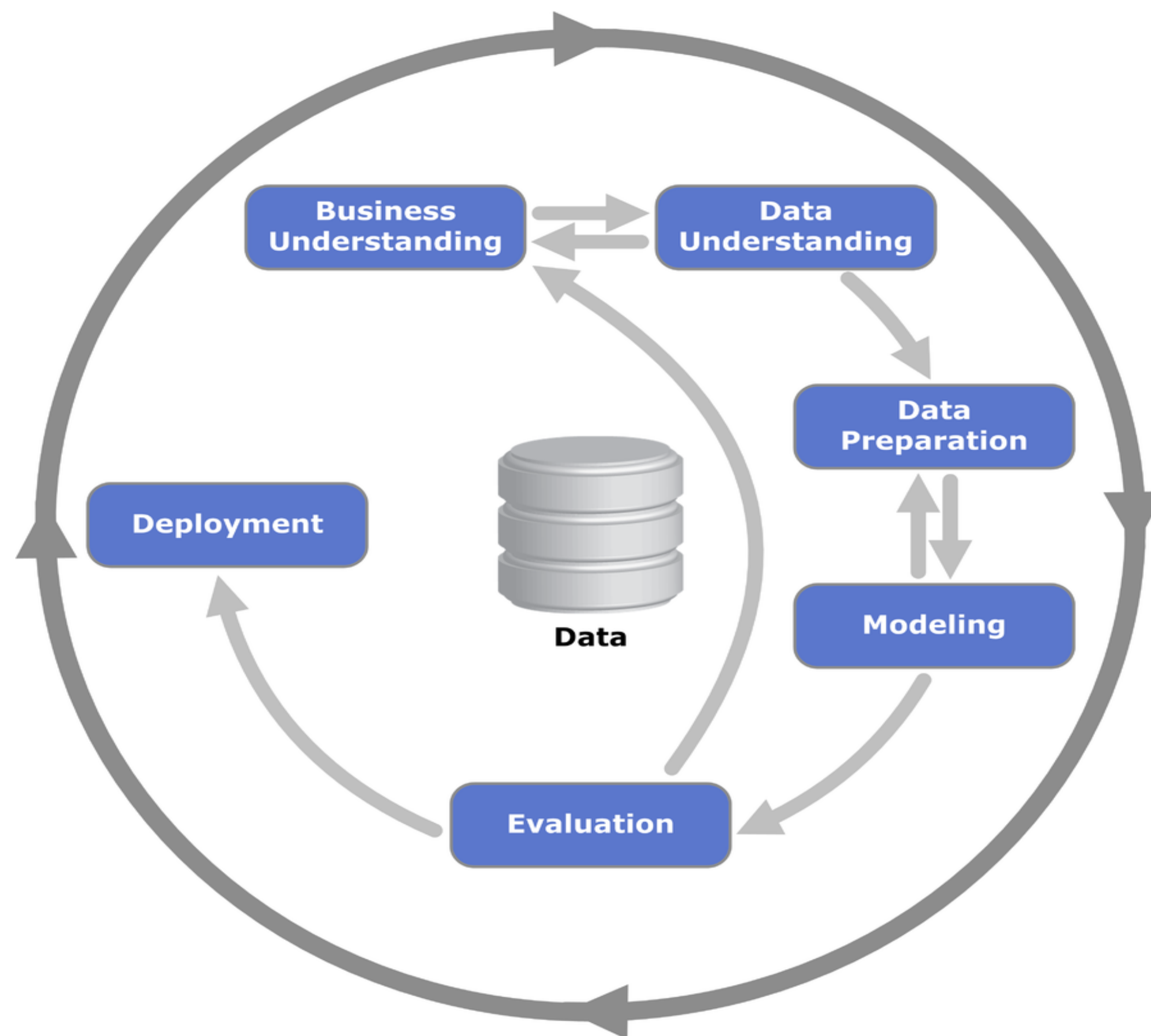# A VPython example: Titanic dataset

- We will follow all the Data Science cycle. See it on the tutorial folder.

THANK YOU