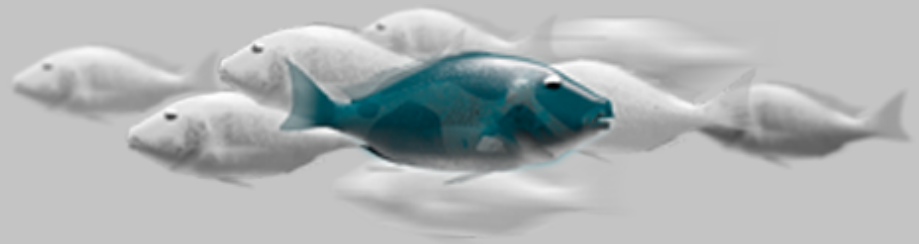
EDEM



Master Data Analytics | 3a Edición

Artificial Intelligence Intro V

Jose Peris Adsuara

Jose Peris Adsuara

Head of Ai @Bit2Me
Data analytics @Edem

bit 2 me

EDEM
Escuela de Empresarios

- **CTO // Digital Product Manager @aiwannapay.com**
  aiwanna
  Sep 2019 – Present · 2 yrs 4 mos

  The most complete SaaS solution for HORECA channel. Scan, order and pay from your smartphone.
  Artificial Intelligence for Small and Medium enterprises. Developing AI algorithms and display augmented analytics. Reduce your costs, increase sales and automate your repetitive tasks

- **Senior Consultant**
  straiqr.ai GmbH · Part-time
  Jun 2021 – Present · 7 mos

  Straiqr is a German company that is changing fashion rules through Metaverse, nft´s, digital fashion & AI
  We create digital fashion for the digital generation.We believe in a world where digital garment and clothing design is authentic and true to you. STRAIQR Intelligent Shopping is ready - now - in an easy to navigate store, where accurate delivery and sustainability is a consideration at every step of the way.

- **Teacher // Machine Learning and AI consulting and for MÁSTER EN DATA ANALYTICS & MASTER FINANZAS**
  EDEM Escuela de Empresarios
  Apr 2019 – Present · 2 yrs 9 mos
  Valencia y alrededores, España

  -Coordinating AI and ML contents , team building
  -Machine Learning and Deep Learning teacher:
  -Master Data Analytics
  -Master Finanzas
  -EMBA Executive
  -Webinar " IA aplicada a pymes"
  -Online

- **Data scientist // CEO & co-founder**
  tailor
  Jul 2017 – Sep 2019 · 2 yrs 3 mos
  Valencia, Spain

  Our mision is to improve the online shopping experience in fashion industry.
  Tailored by Big Data is an spanish Startup born in Valencia wich is reasearching in AI.
  Developing algorithms for fashion and trend industry .Our project has been developed under Garaje de Lanzadera incubator program & Órbita accelerator program
  https://lanzadera.es/
  http://www.programaorbita.com
  http://www.tailoredbybigdata.com
  Deep Learning - Machine Learning - Computer vision

2021:
-Winner Star Starups contest "Star days" hackathon

2019:
-Included in PlayBook platform, powered by Plug and Play Ventures

2018:
-Selected for Lanzadera start-up accelerator program
-Selected for Programa Órbita start-up accelerator program
-Selected as speaker in VI Mobile e-commece Madrid

// Index

Auto Exploratory Data Analysis

Auto EDA //

¿Why EDA is very important?

1 // We need a real understanding of the **business problem**

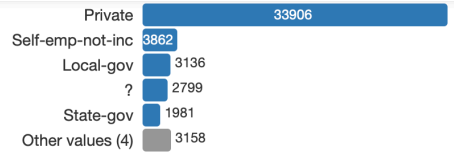
2 // Data is about stories we need the f**ull picture**.

3 // It gives us **hints** about variables, missing values, correlations

Auto EDA //

workclass
Categorical

Distinct	9
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	3.0 MiB

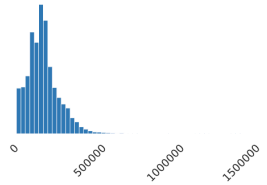


Toggle details

fnlwgt
Real number ($\mathbb{R}_{\geq 0}$)

Distinct	28523
Distinct (%)	58.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	189664.1346

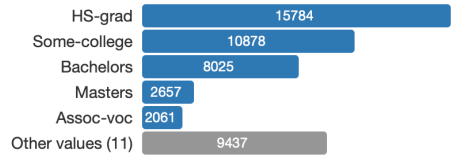
Minimum	12285
Maximum	1490400
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	381.7 KiB



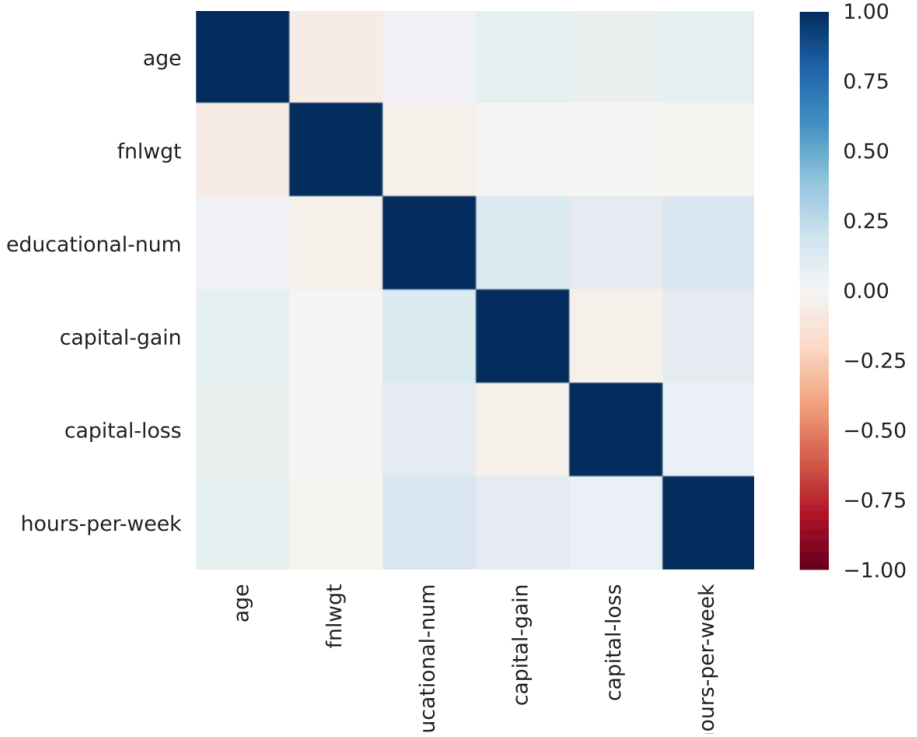
Toggle details

education
Categorical

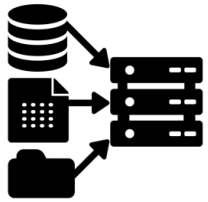
Distinct	16
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	3.0 MiB



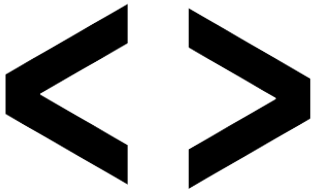
Toggle details



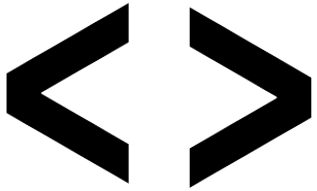
AUTO ML // Auto-EDA



titanic



Automated_EDA.ipynb



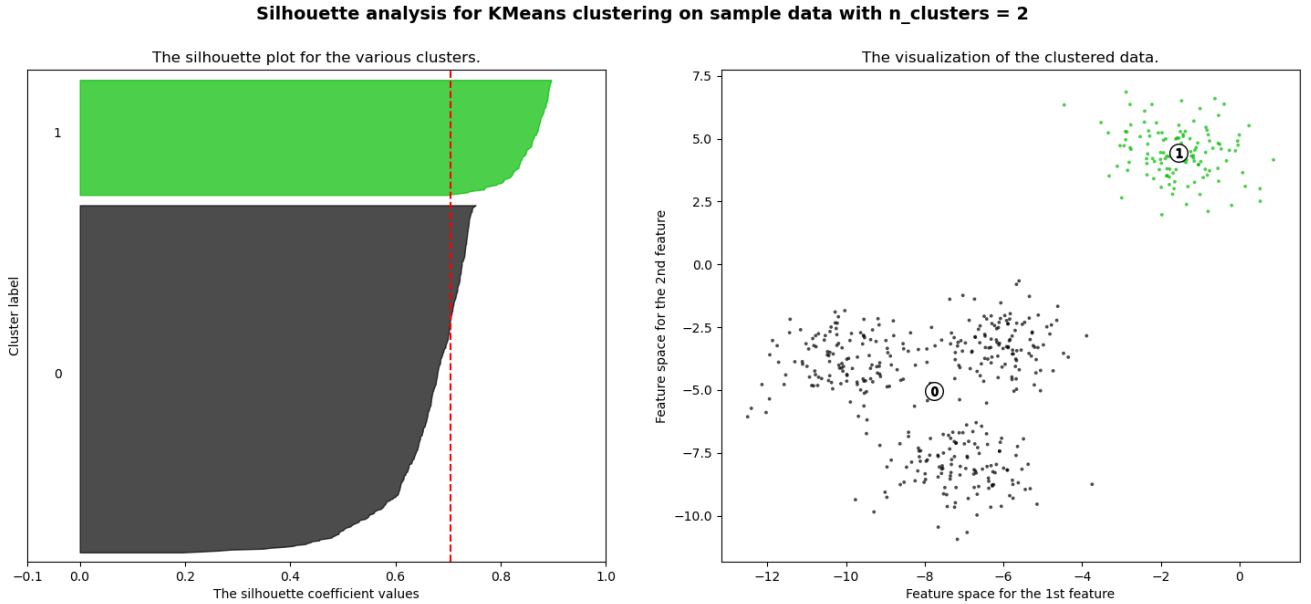
Automated_EDA2.ipynb

Clustering(K)

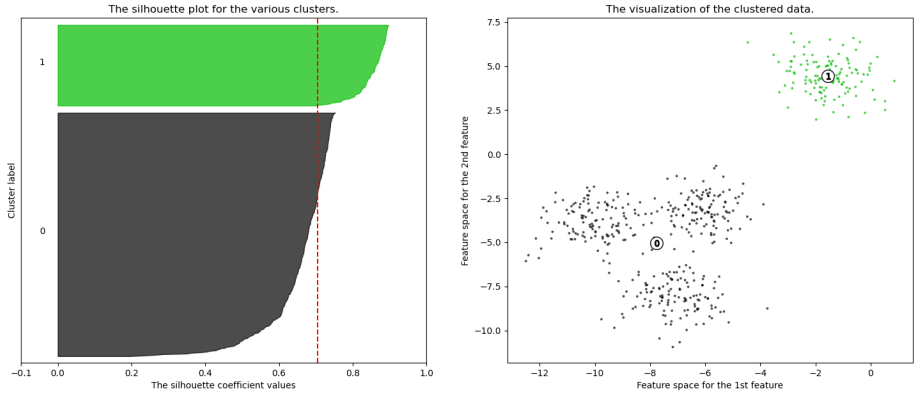
Silhoutte score

Silhouette analysis can be used t**o study the separation distance between the resulting clusters.** The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. **This measure has a range of [-1, 1].**

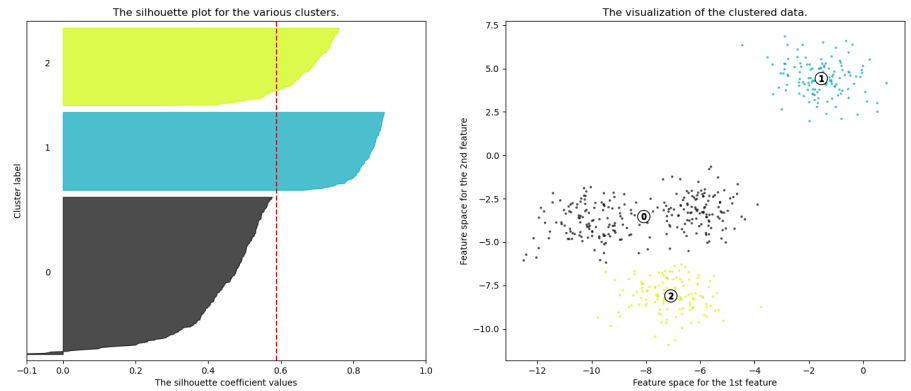
Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters , and negative values indicate that those samples might have been assigned to the wrong cluster.



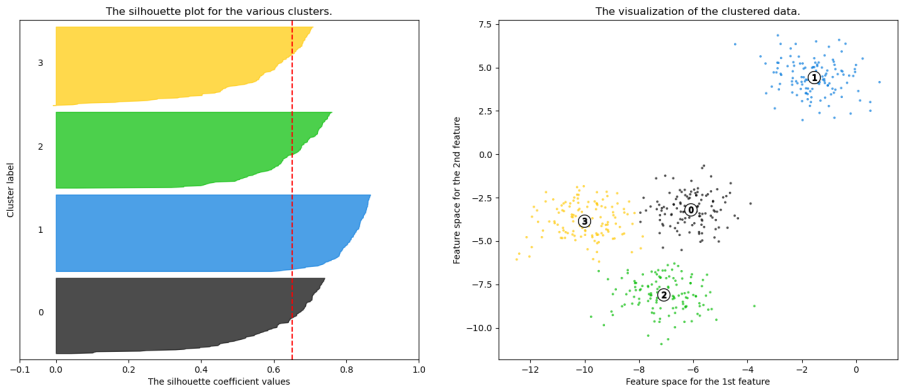
Silhouette analysis for KMeans clustering on sample data with n_clusters = 2



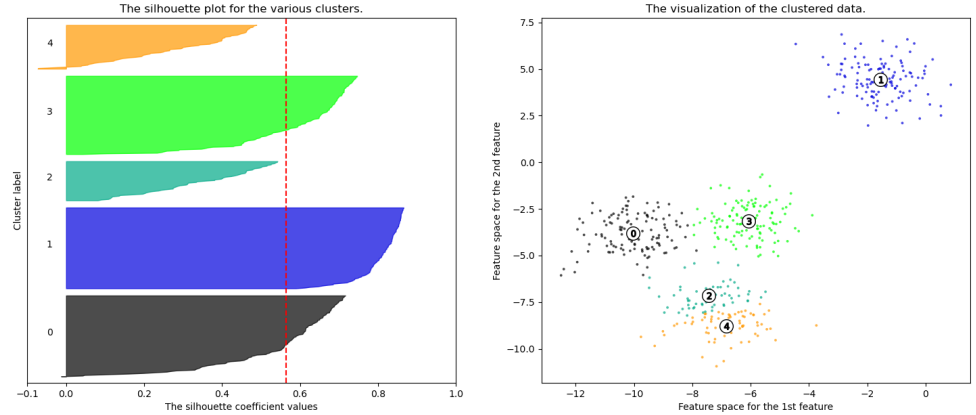
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3



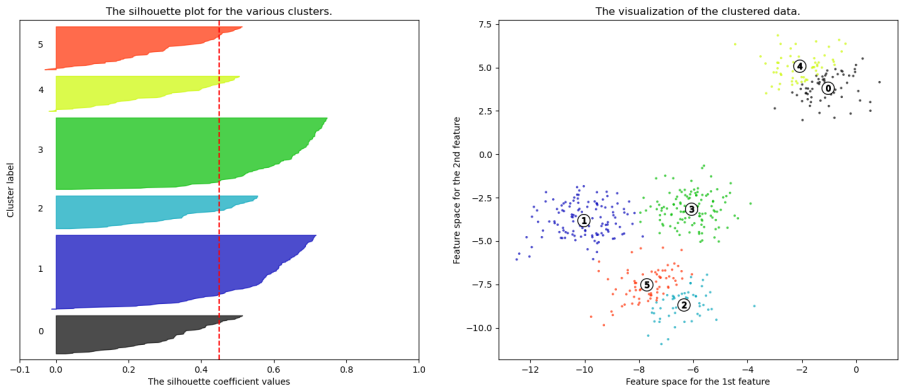
Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



Silhouette analysis for KMeans clustering on sample data with n_clusters = 5



Silhouette analysis for KMeans clustering on sample data with n_clusters = 6

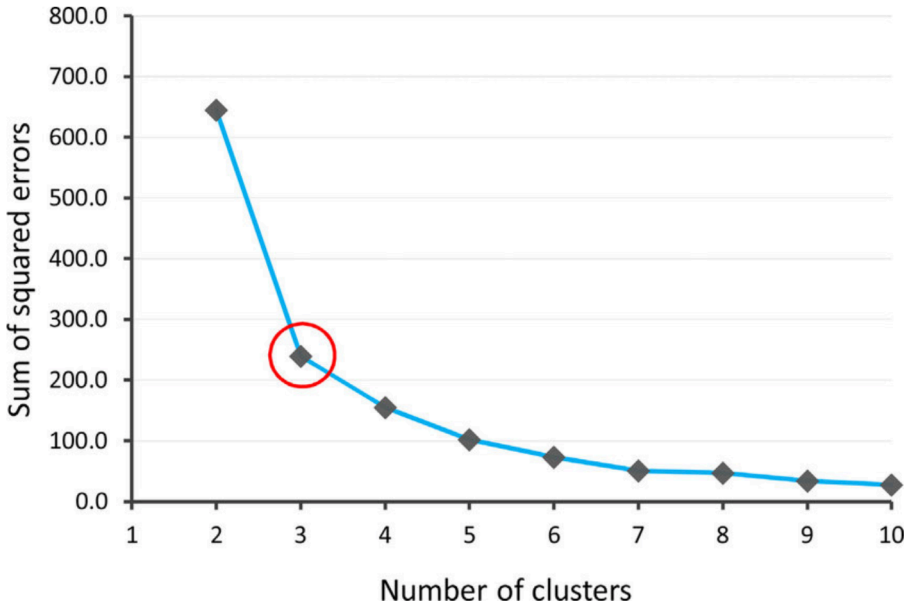


Out:

```
For n_clusters = 2 The average silhouette_score is : 0.7049787496083262
For n_clusters = 3 The average silhouette_score is : 0.5882004012129721
For n_clusters = 4 The average silhouette_score is : 0.6505186632729437
For n_clusters = 5 The average silhouette_score is : 0.56376469026194
For n_clusters = 6 The average silhouette_score is : 0.4504666294372765
```

The Elbow Method

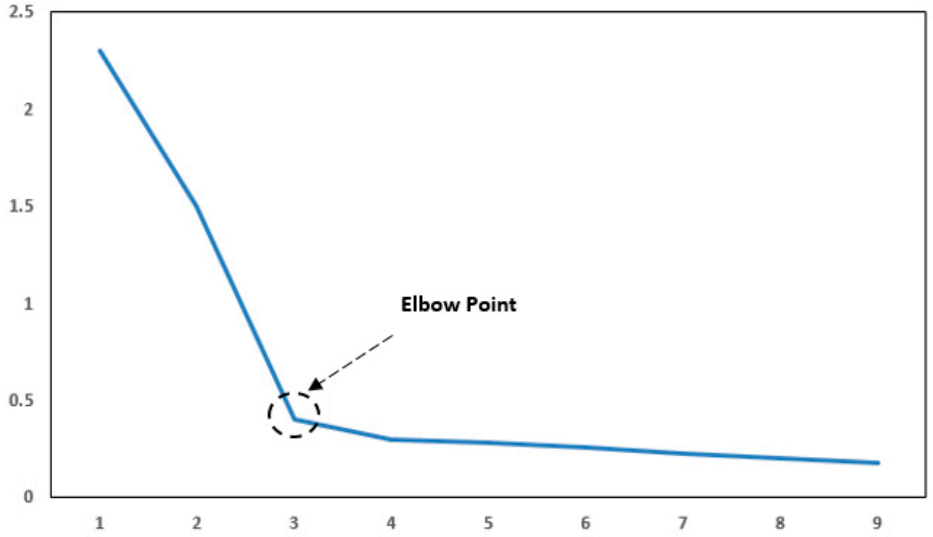
Therefore we have to come up with a technique that somehow **will help us decide how many clusters we should use for the K-Means model.**
The Elbow method is a very popular technique and the idea is to run k-means clustering for a range of clusters k (let's say from 1 to 10) and for each value, we are calculating the sum of squared distances from each point to its assigned center(distortions).

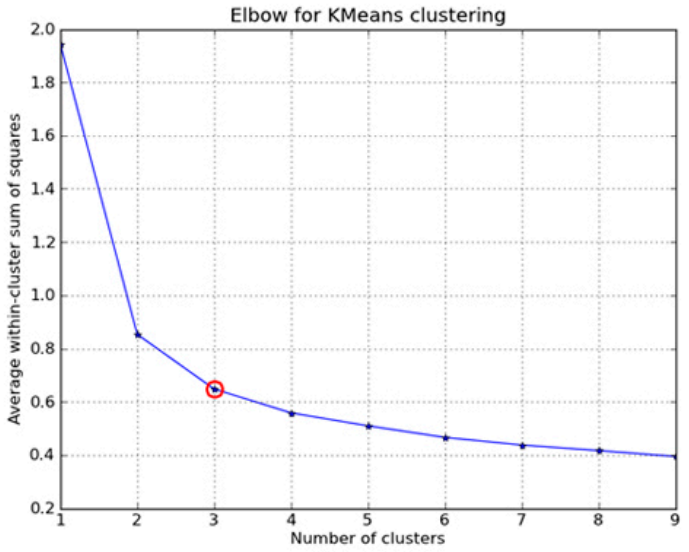


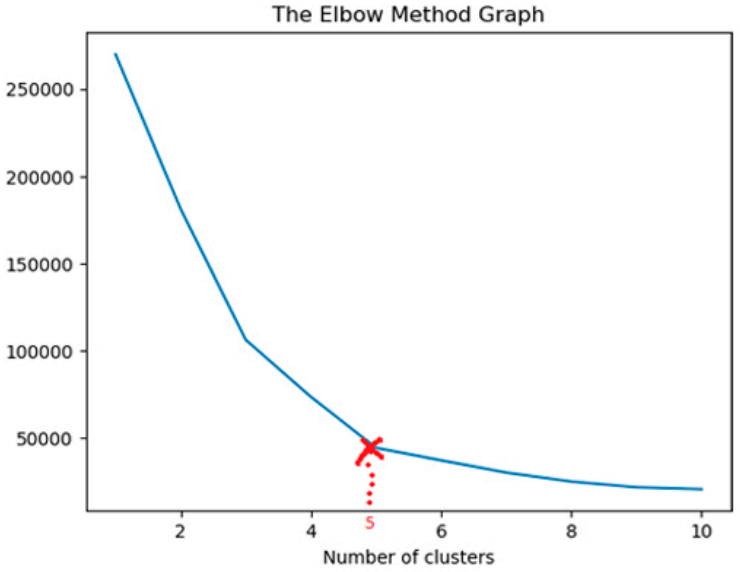
Result of the elbow method to determine optimum number of clusters.

The Elbow Method

When the distortions are plotted and the plot looks like an arm then the **“elbow”(the point of inflection on the curve) is the best value of k.**

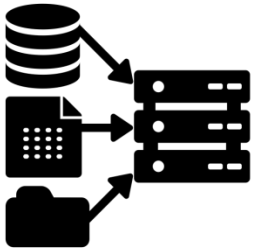




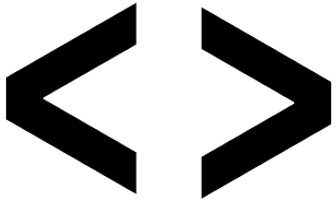


CLUSTERING // K

Google Colab The Elbow Method & Silhoutte score



Credit card clean.csv



Credit_card_k.pynb

Visual Programming

VISUAL PROGRAMMING // **ORANGE**

orange

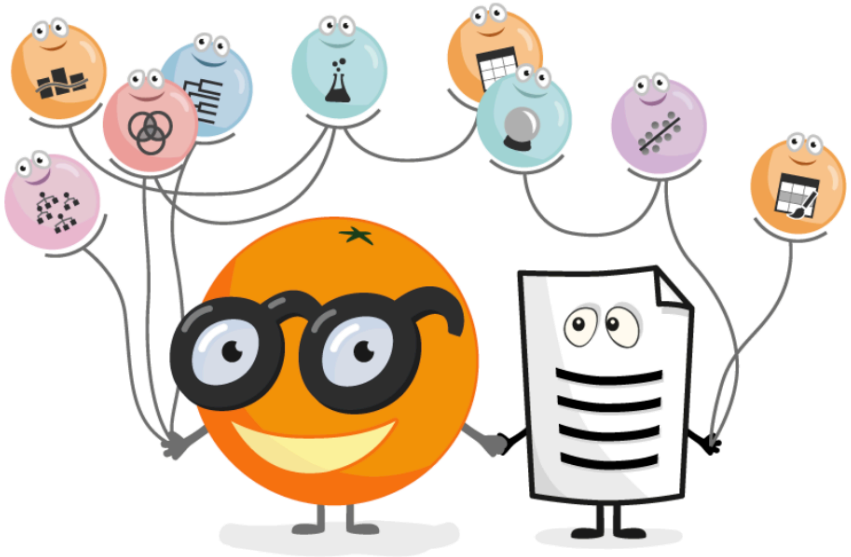
Screenshots Workflows Download Blog Docs Workshops

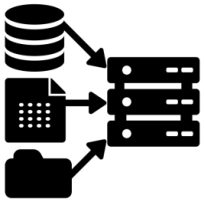
Data Mining
Fruitful and Fun

Open source machine learning and data visualization.

Build data analysis workflows visually, with a large, diverse toolbox.

Download Orange





Download Orange

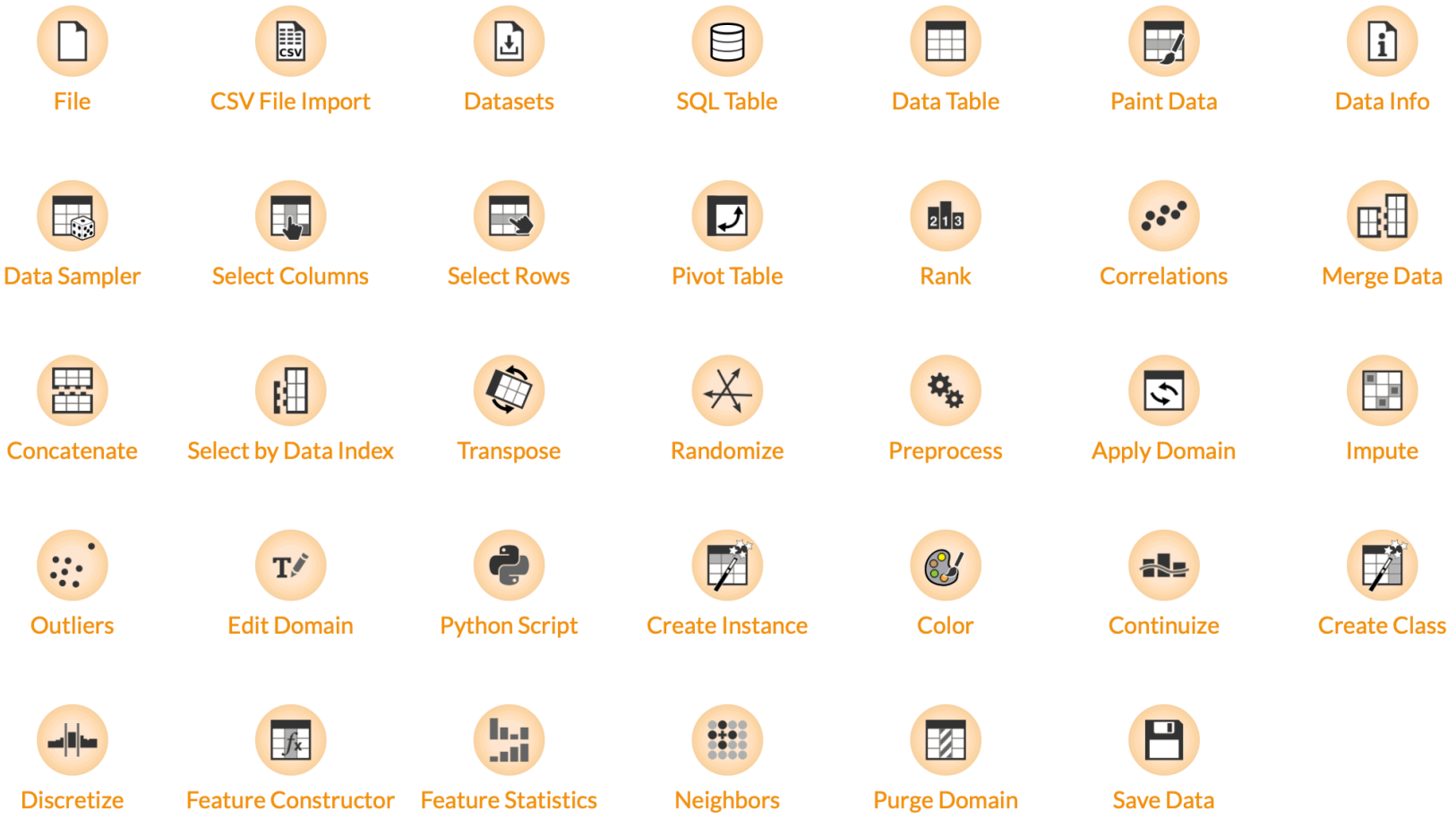
VISUAL PROGRAMMING // **ORANGE**

Orange is a component-based data mining software. It includes a range of **data visualization, exploration, preprocessing and modeling techniques**. It can be used through a nice and intuitive user interface or, for more advanced users, as a module for the Python programming language.

ANACONDA.ORG

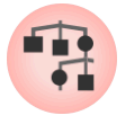
VISUAL PROGRAMMING // **ORANGE**

Widgets // **Data**



VISUAL PROGRAMMING // **ORANGE**

Widgets // **Visualize**


Tree Viewer


Box Plot


Violin Plot


Distributions


Scatter Plot


Line Plot


Bar Plot


Sieve Diagram


Mosaic Display


FreeViz


Linear Projection


Radviz

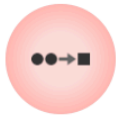

Heat Map

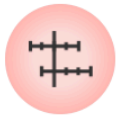

Venn Diagram


Silhouette Plot


Pythagorean Tree


Pythagorean Forest


CN2 Rule Viewer


Nomogram

VISUAL PROGRAMMING // **ORANGE**

Widgets // **Supervised Models**


Constant


CN2 Rule Induction


Calibrated Learner


kNN


Tree


Random Forest


Gradient Boosting


SVM


Linear Regression


Logistic Regression


Naive Bayes


AdaBoost


Neural Network


Stochastic Gradient Descent


Stacking


Save Model


Load Model

VISUAL PROGRAMMING // **ORANGE**

Widgets // **Unupervised Models**


Distance File


Distance Matrix

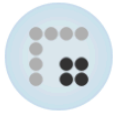

t-SNE


Distance Map


Hierarchical Clustering


k-Means


Louvain Clustering

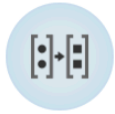

DBSCAN

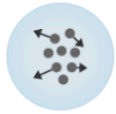

Manifold Learning


PCA


Correspondence Analysis


Distances


Distance Transformation


MDS


Save Distance Matrix


Self-Organizing Map

VISUAL PROGRAMMING // **ORANGE**

Widgets // **Evaluate**


Test and Score


Predictions


Confusion Matrix

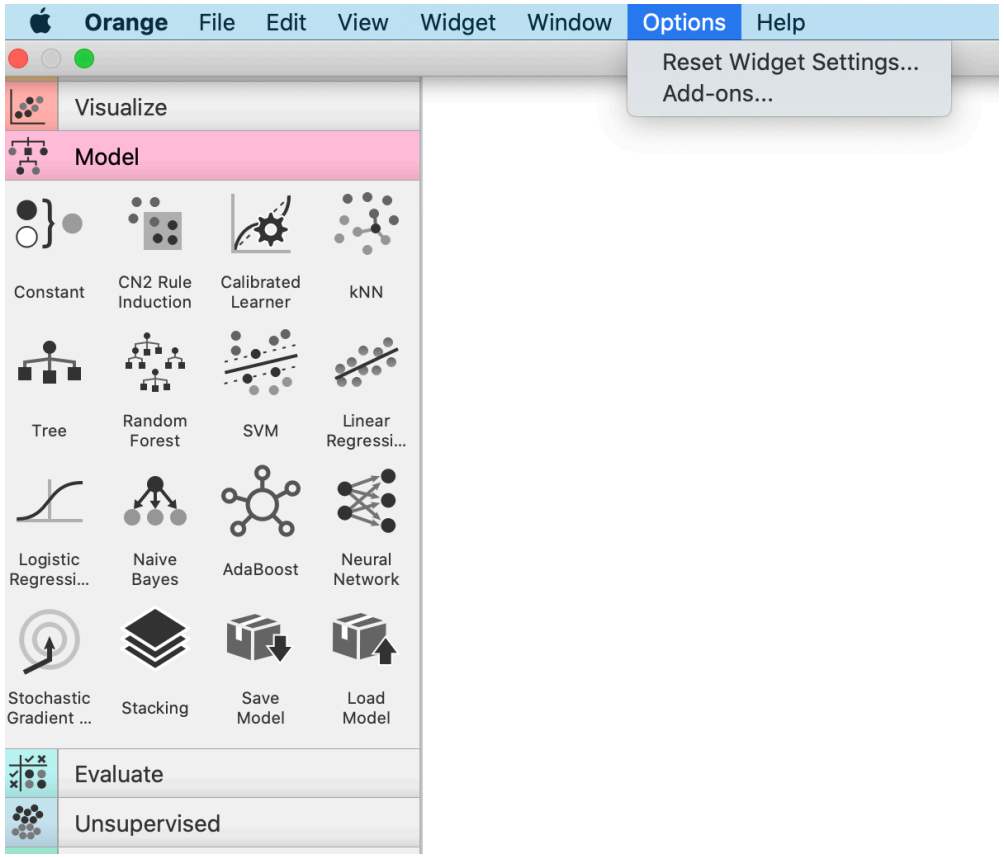

ROC Analysis

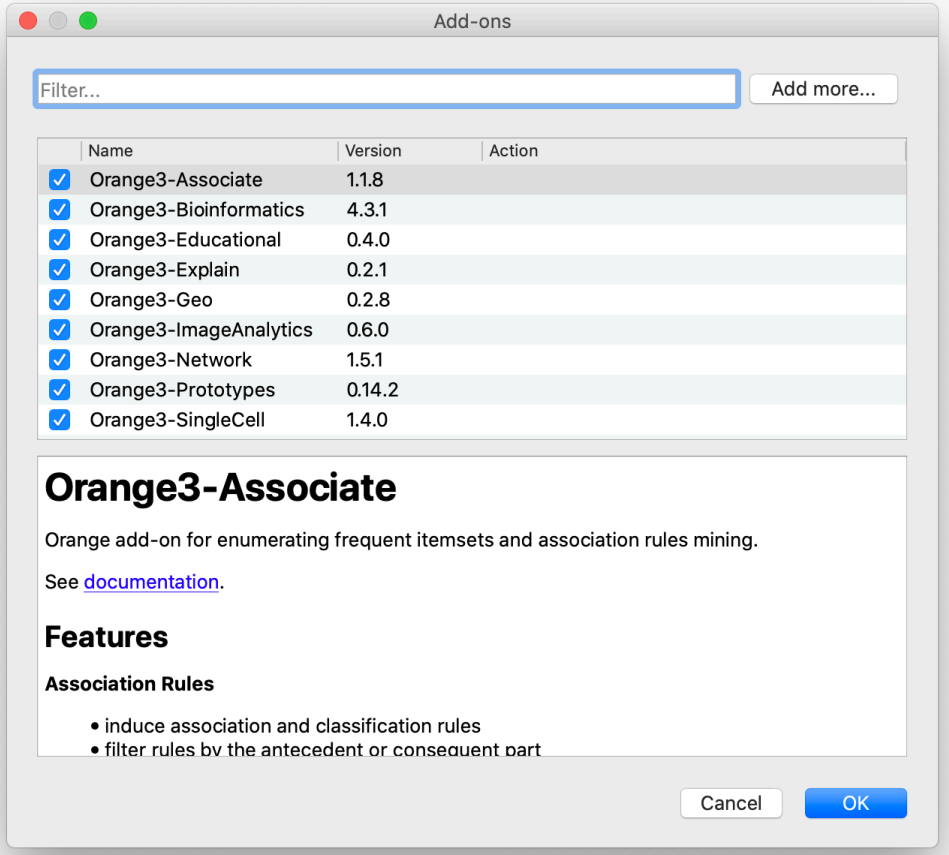

Lift Curve


Calibration Plot

VISUAL PROGRAMMING // **ORANGE**

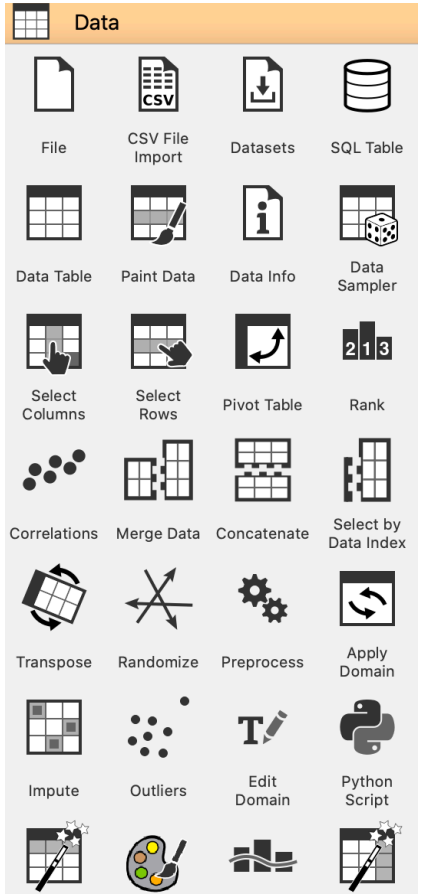
Options // **Add ons**

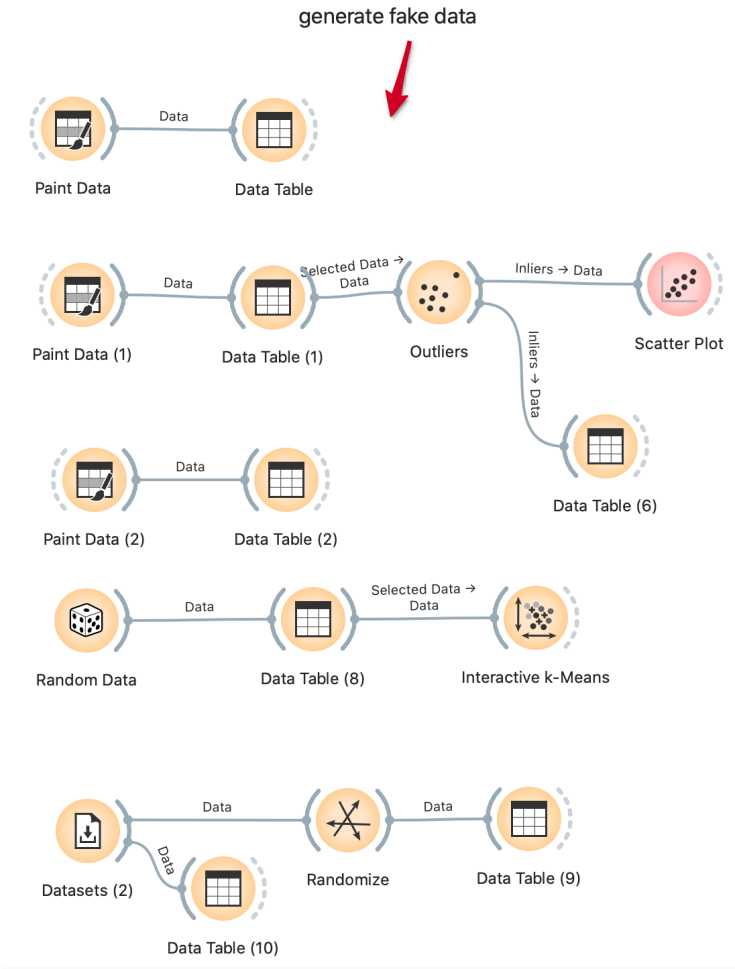


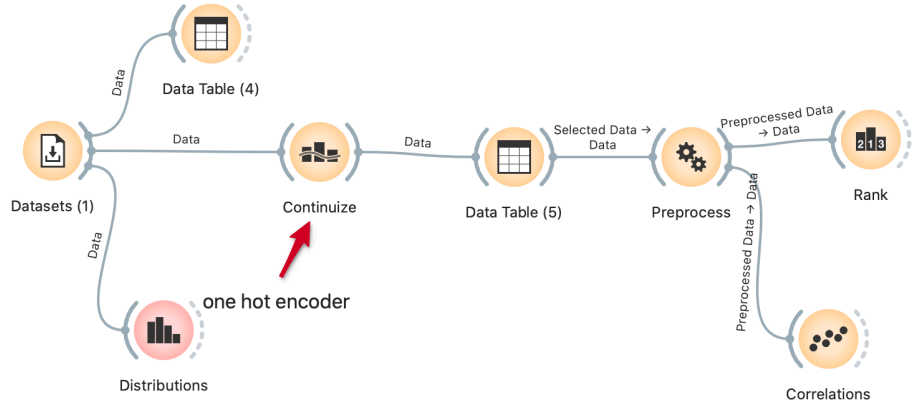


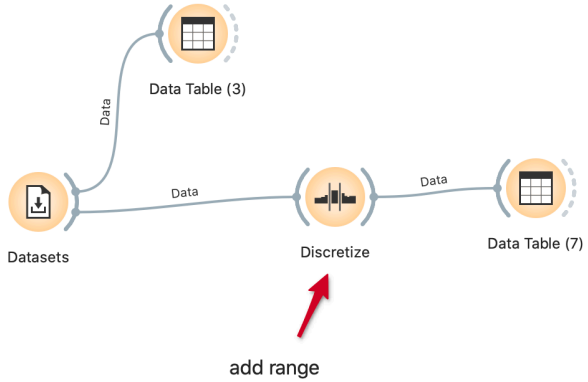
VISUAL PROGRAMMING // **ORANGE**

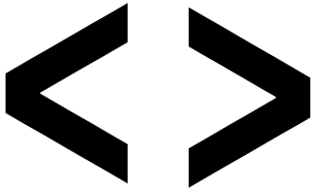
Data / Cool widgets







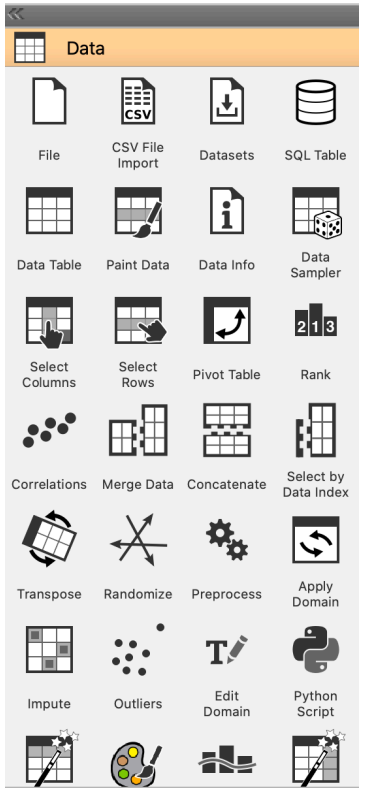


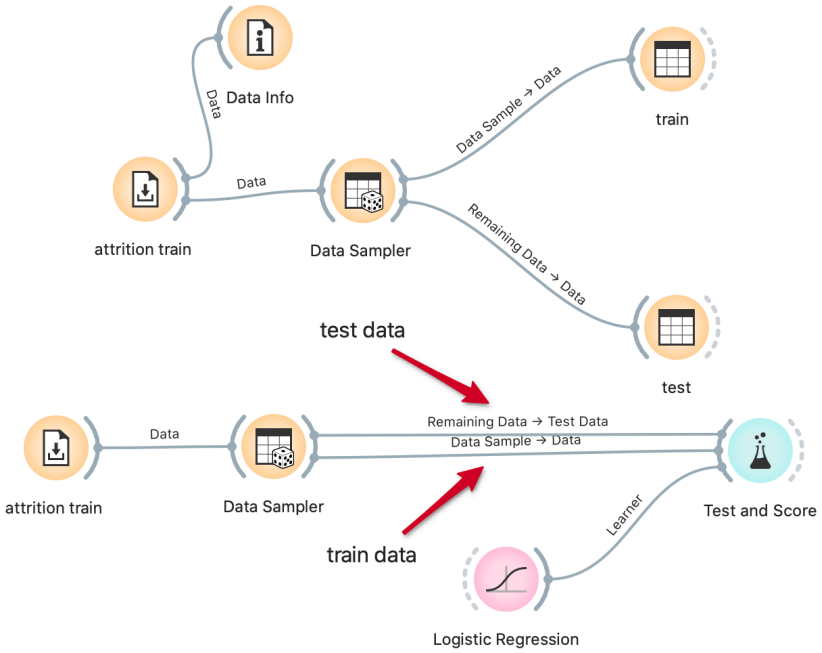


Cool widgets

VISUAL PROGRAMMING // **ORANGE**

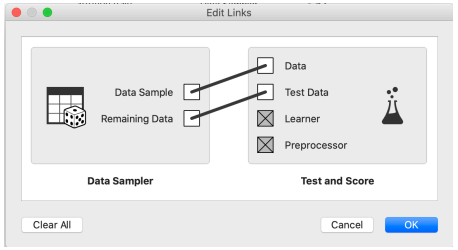
Data // sampler

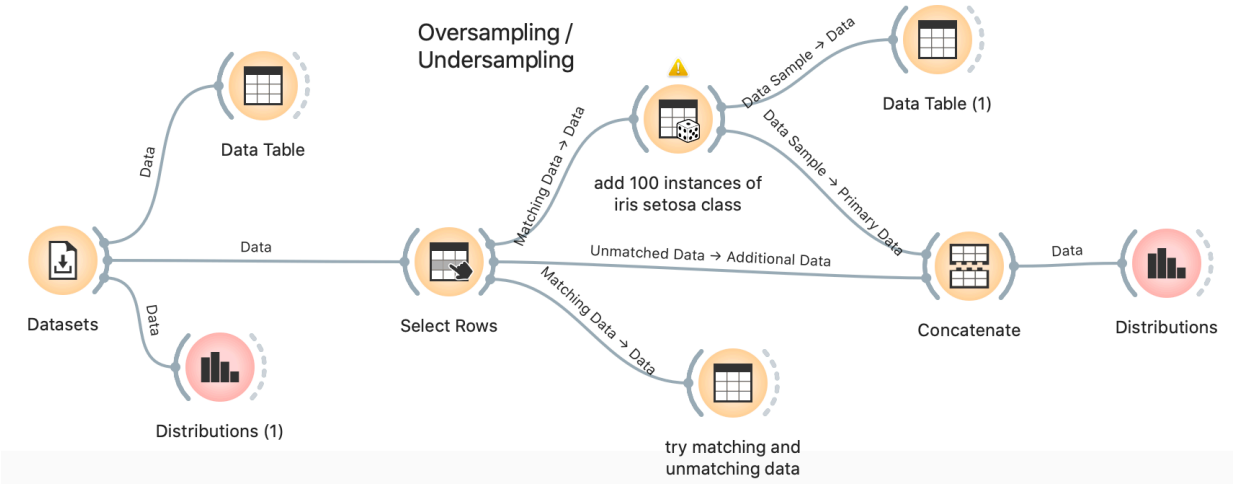


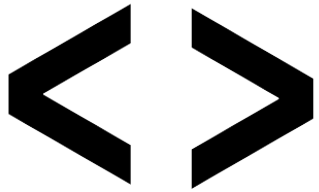


Over/Undersampling

Data Sampler can also be used to oversample a minority class or undersample majority class in the data. Let us show an example for oversampling





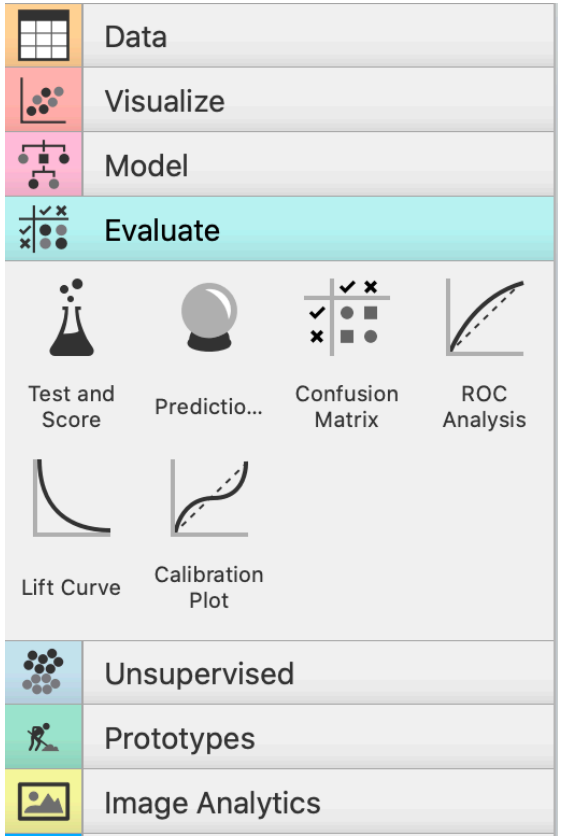


Data sampler

VISUAL PROGRAMMING // **ORANGE**

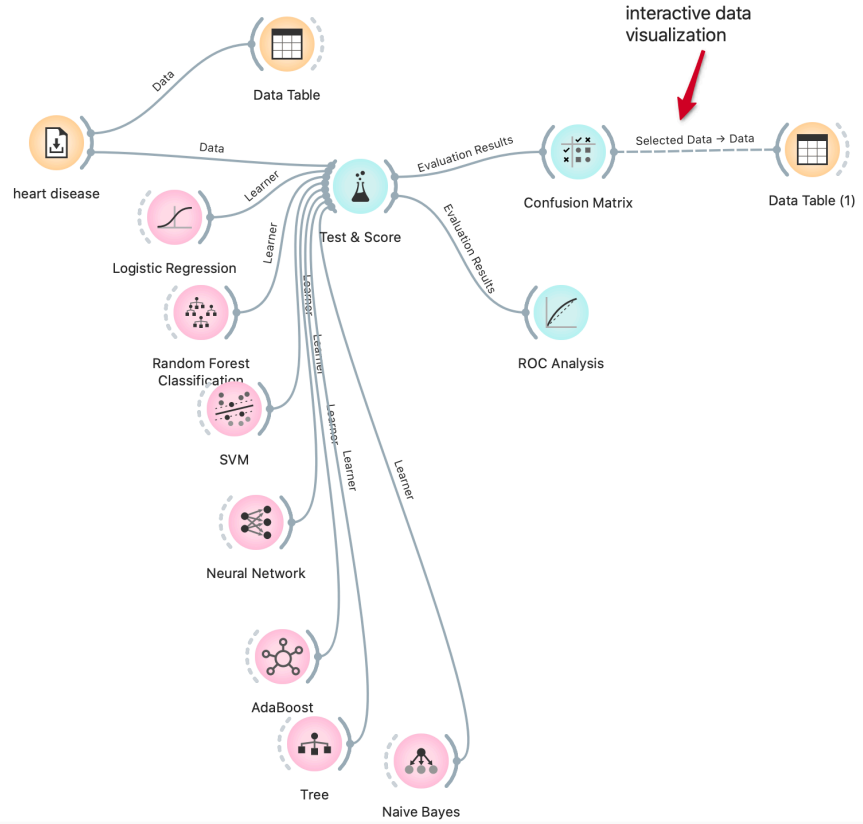
In this example you are testing the models **just with Training data**

Evaluate // test & score



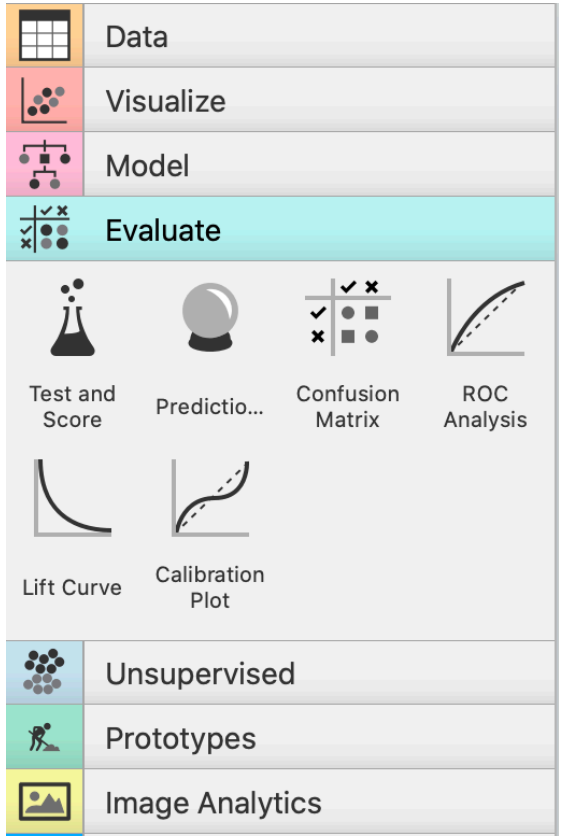
The widget tests learning algorithms. **Different sampling schemes are available, including using separate test data.** The widget does two things. First, it shows a table with different classifier performance measures, such as classification accuracy and area under the curve. Second, it outputs evaluation results, which can be used by other widgets for analyzing the performance of classifiers, such as ROC Analysis or Confusion Matrix.

Use case 1 // train data

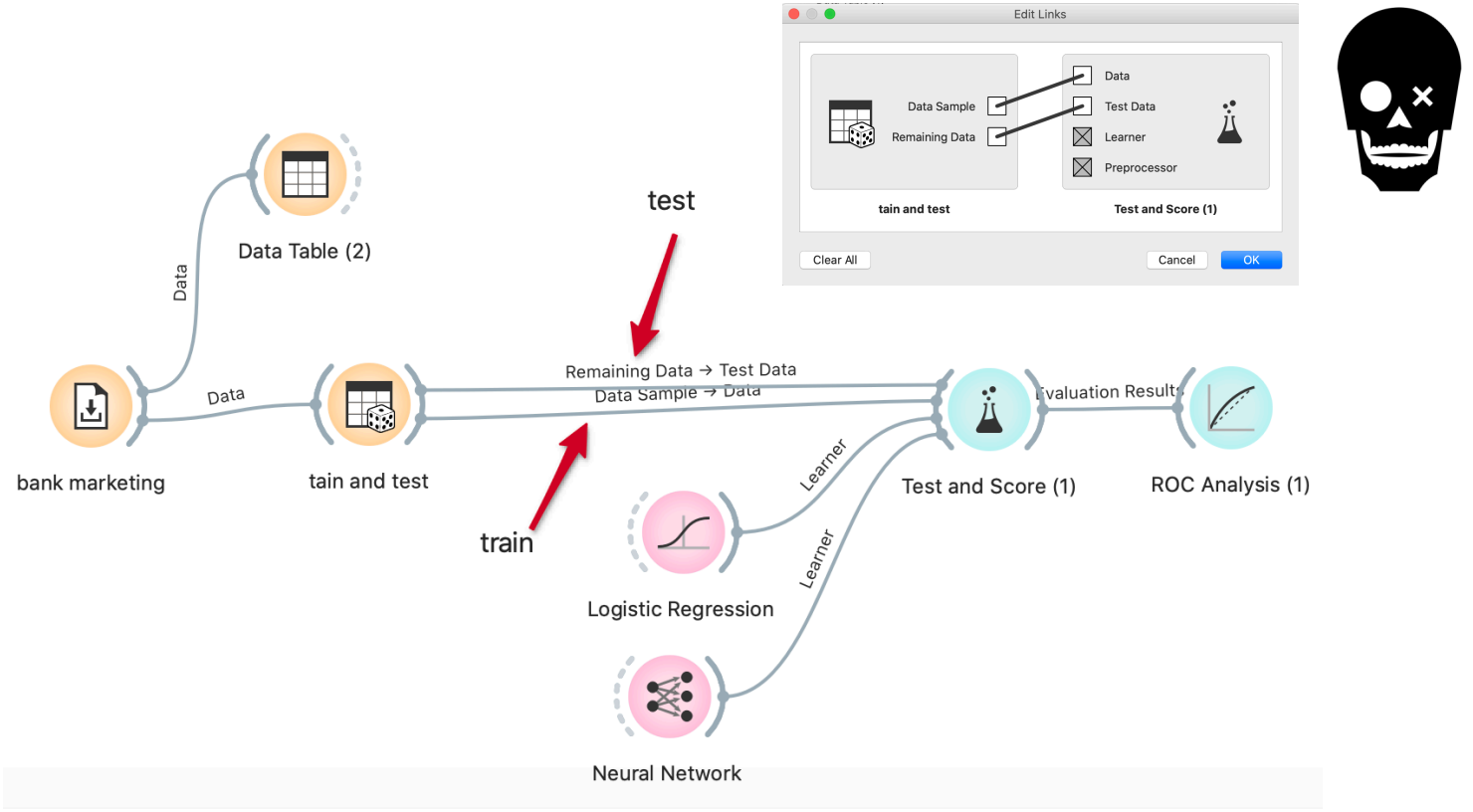


VISUAL PROGRAMMING // **ORANGE**

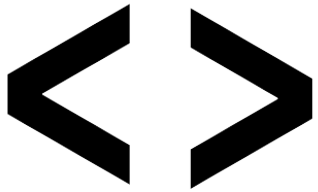
Evaluate // test & score



Use case 2 // train & test data



Workflows examples // Solved



Test & score

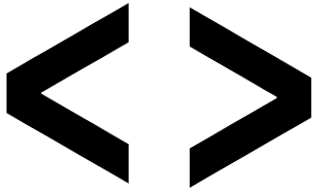
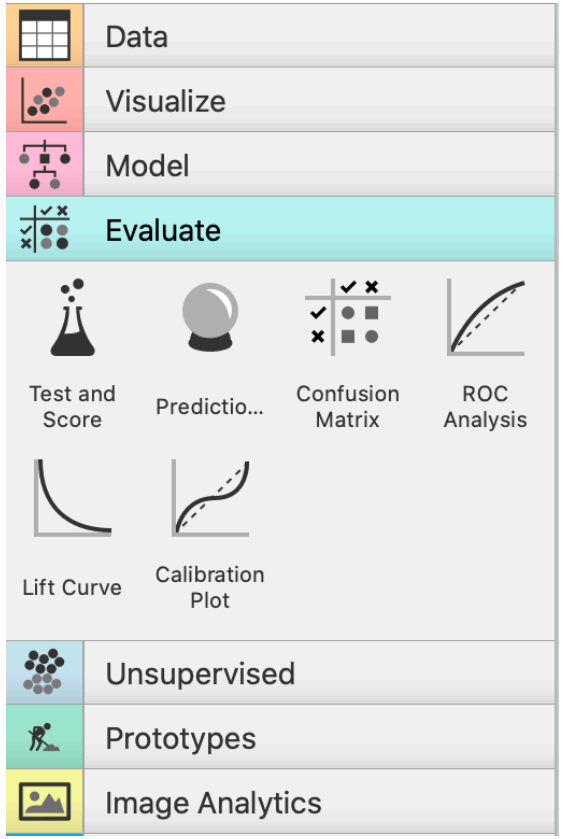


Image classification workflow

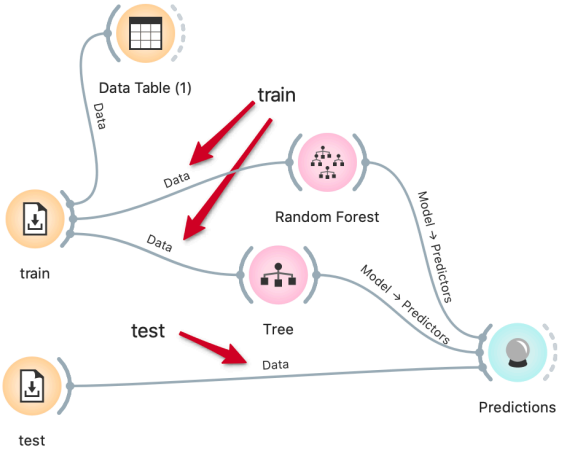
VISUAL PROGRAMMING // **ORANGE**

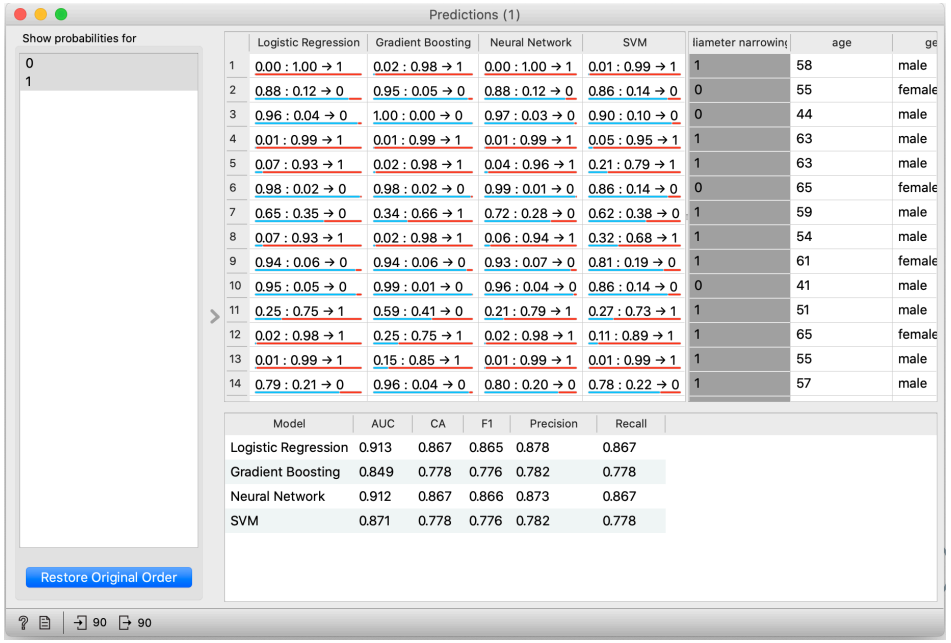
Evaluate // Predictions



The widget tests learning algorithms. **Different sampling schemes are available, including using separate test data.**

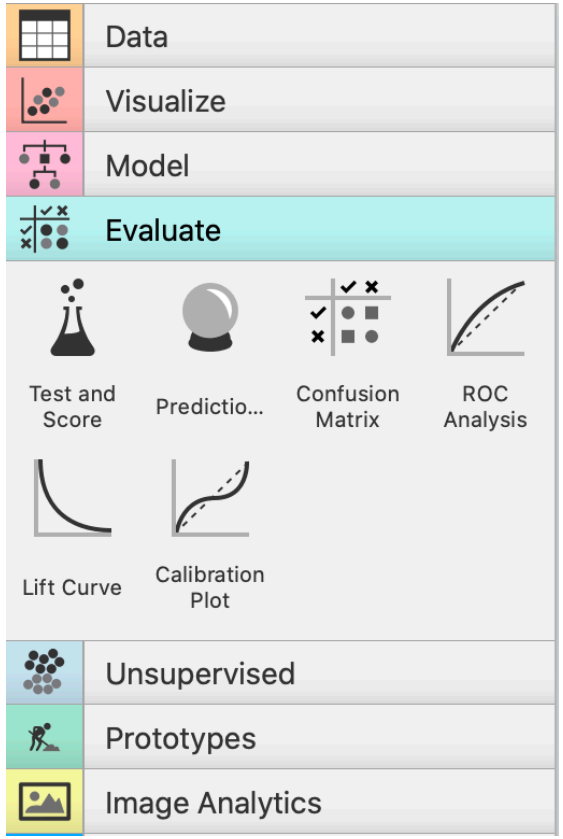
The widget does two things. First, it **shows a table with different classifier performance measures**, such as classification accuracy and area under the curve. Second, **it outputs evaluation results, which can be used by other widgets for analyzing the performance of classifiers**, such as ROC Analysis or Confusion Matrix.





VISUAL PROGRAMMING // **ORANGE**

Evaluate // Predictions



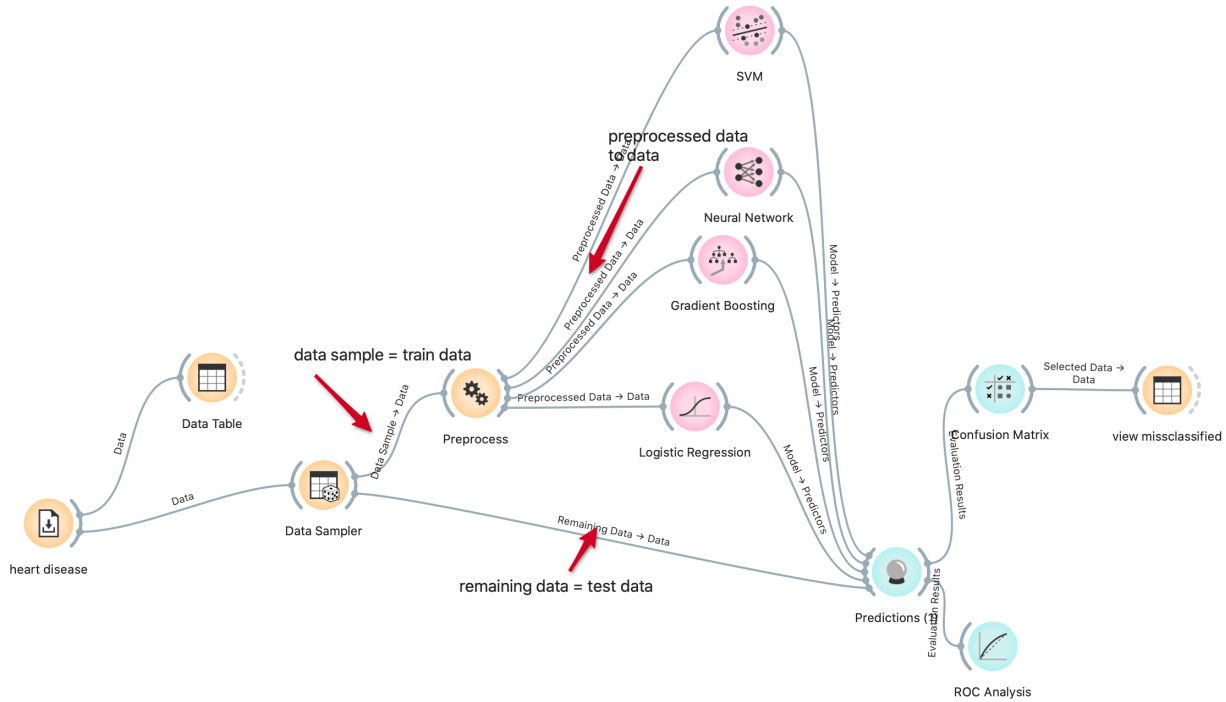
Shows models' predictions on the data.

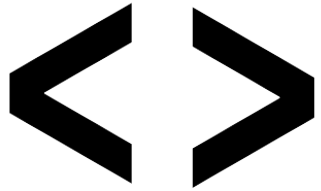
Inputs

- **Train data + model**
- **Test data:** predictors to be used on the data

Outputs

- **Predictions**: data with added predictions
- **Evaluation Results**: results of testing

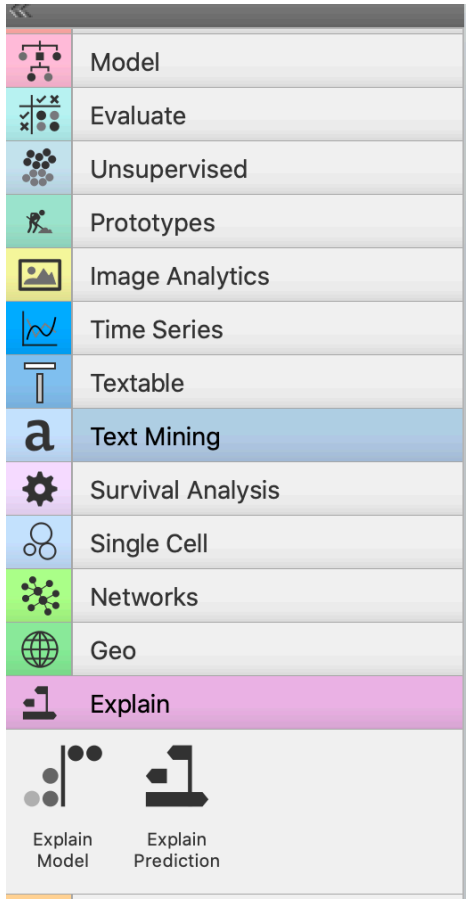




Predictors

VISUAL PROGRAMMING // **ORANGE**

Explain // model

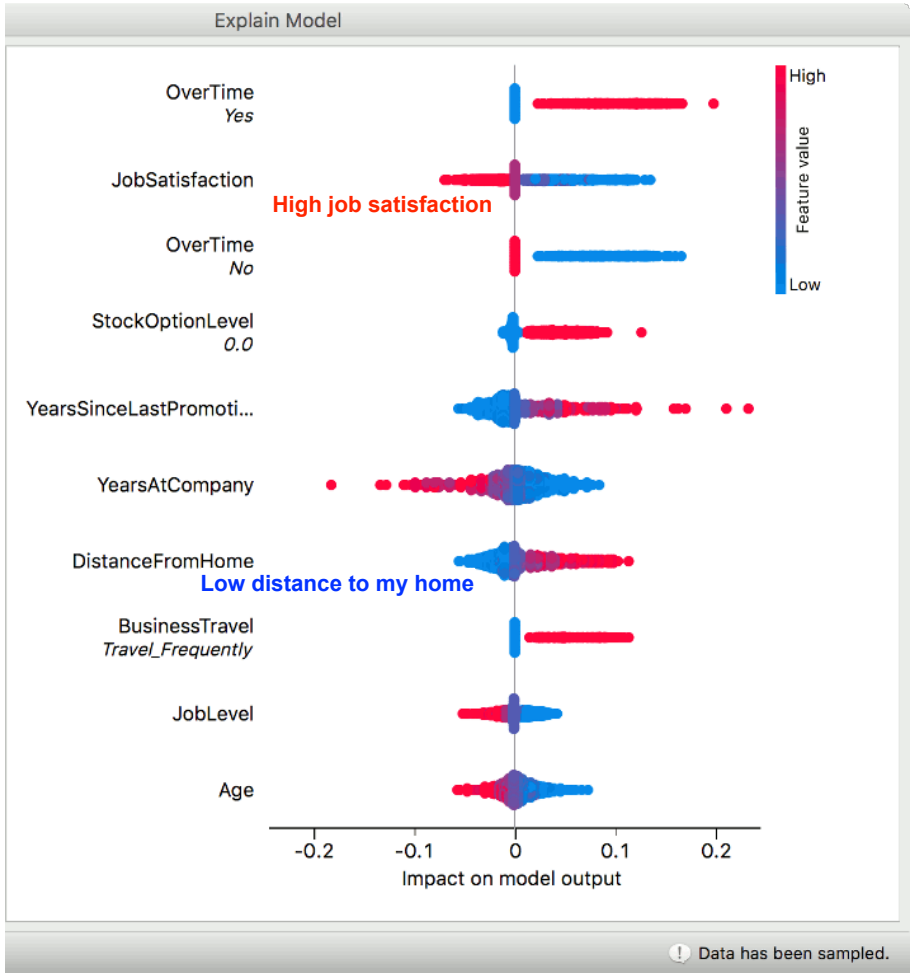


Explain Model. The widget lists top ranked variables, which means they contribute the most to the selected target variable. As we are trying to understand why people leave the company, we have set the target variable to Yes.

The highest ranked variable is OverTime - this is the variable with the highest impact on the prediction. Having a value Yes in the categorical attribute OverTime (red dots on the right) means the employee is likely to quit. Also, having low job satisfaction contributes to attrition (blue values on the right). The visualization shows the values which have a high impact on the prediction of the selected class on the right and those which vote against the selected class on the left. **The color of dot represents the value of the attribute (red for higher values and blue for lower).**

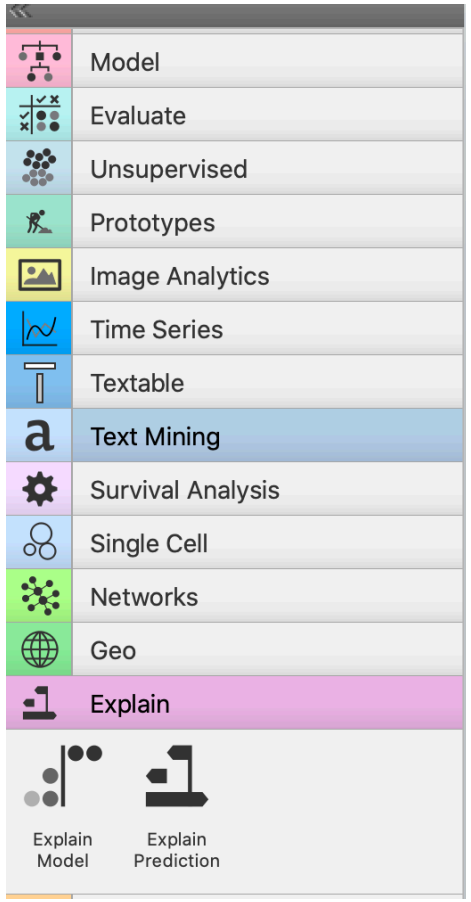
TARGET VARIABLE: YES

I don´t want to quit **I want to quit**



VISUAL PROGRAMMING // **ORANGE**

Explain // model



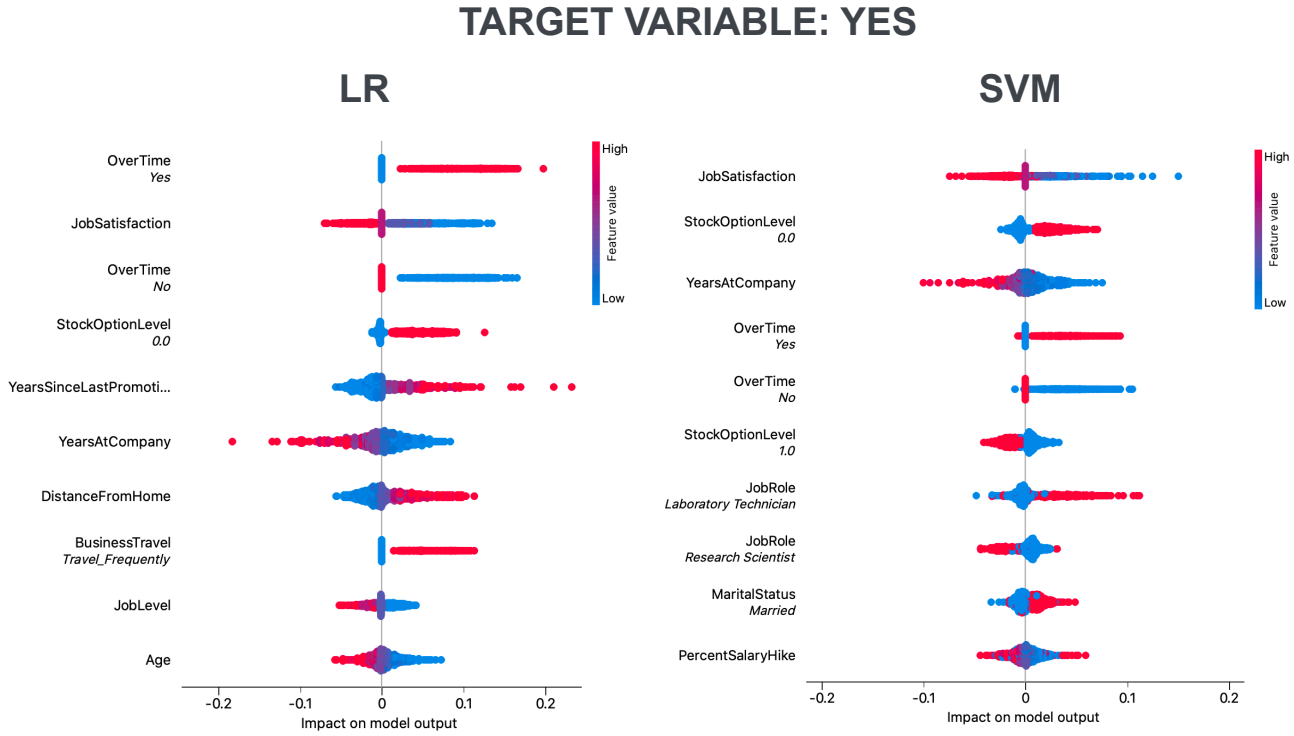
Inputs

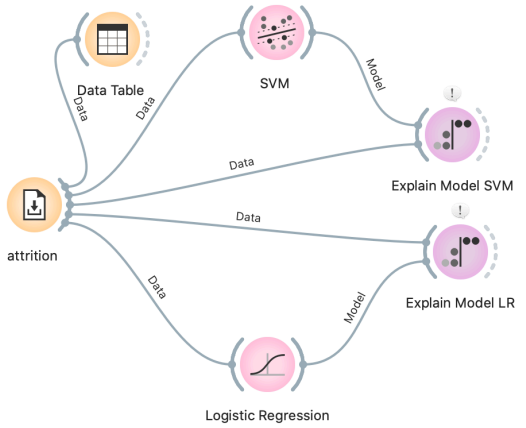
- Data: dataset used to compute the explanations
- Model: a model which widget explains

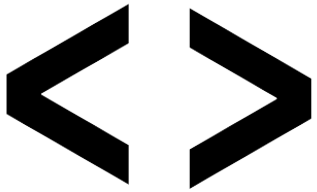
Outputs

- Selected data: data instance that belong to selected points in the plot
- Scores: The score of each attribute. Features that contribute more toward the final prediction have higher scores.

Explain Model widget explains classification and regression models with SHAP library. The widget gets a trained model and reference data on input. It uses the provided data to compute the contribution of each feature toward the prediction for a selected class.



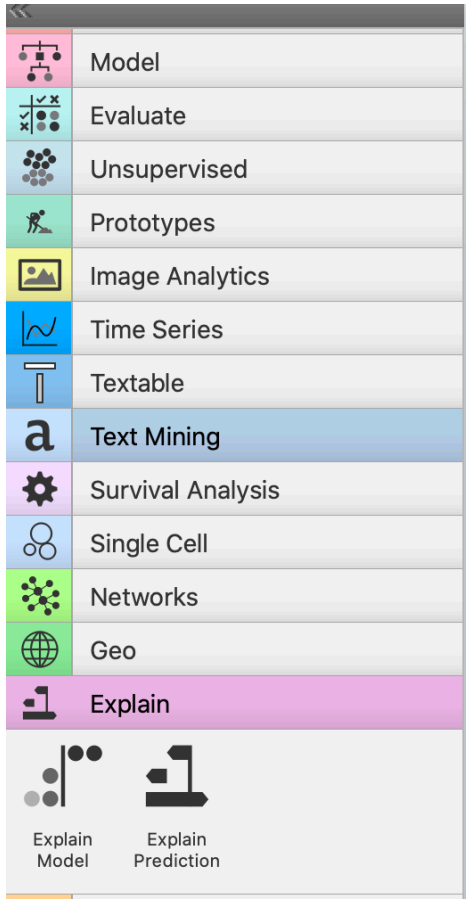




Explain models

VISUAL PROGRAMMING // **ORANGE**

Explain // Prediction

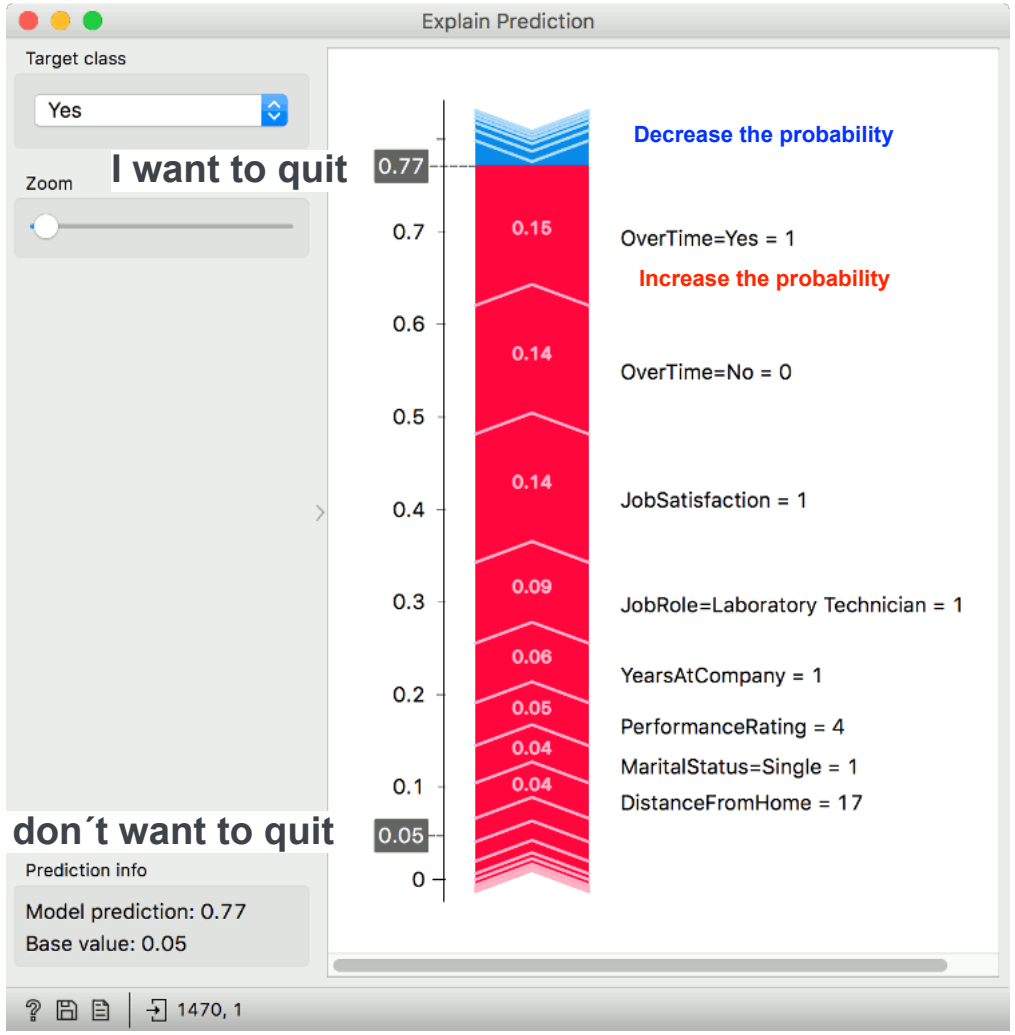


Explain Predictions.Variables in red increase the probability of the target value (conversely, blue decrease it). The size of the arrow corresponds to the SHAP value - in other words, the larger the arrow the larger the variable's contribution to the target value. The model also predicted that John will leave the job with 77 % probability.

As before, the most important variable for John is overtime. Him working overtime contributes a lot to the final prediction. Also, his job satisfaction is low (1 out of 5), making him likely to quit.
The results correspond very much to those of the model, but it might not always be the case. Some people might leave because they are very dissatisfied without working overtime. This would show in Explain Predictions. See how the results change for the other two employees, Rachel and Veronica. Or make up your own employee with Excel and see what would the prediction be.

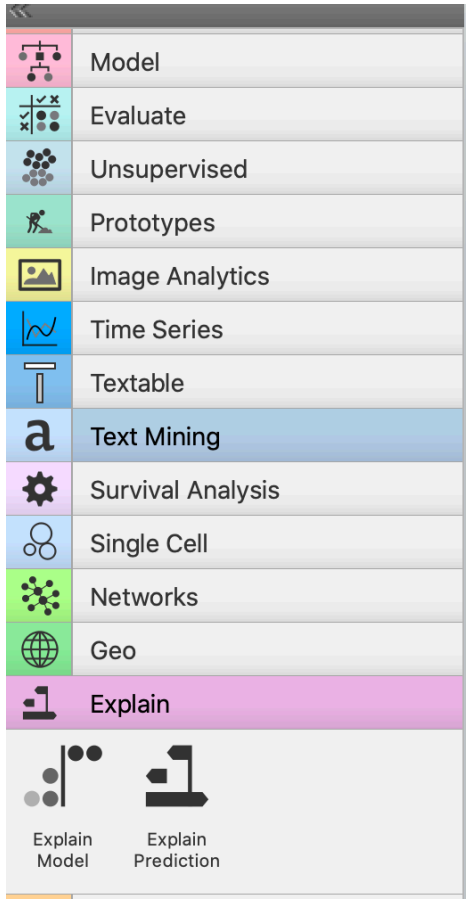
TARGET VARIABLE: YES

RowId0000234: Jhon



VISUAL PROGRAMMING // **ORANGE**

Explain // Prediction

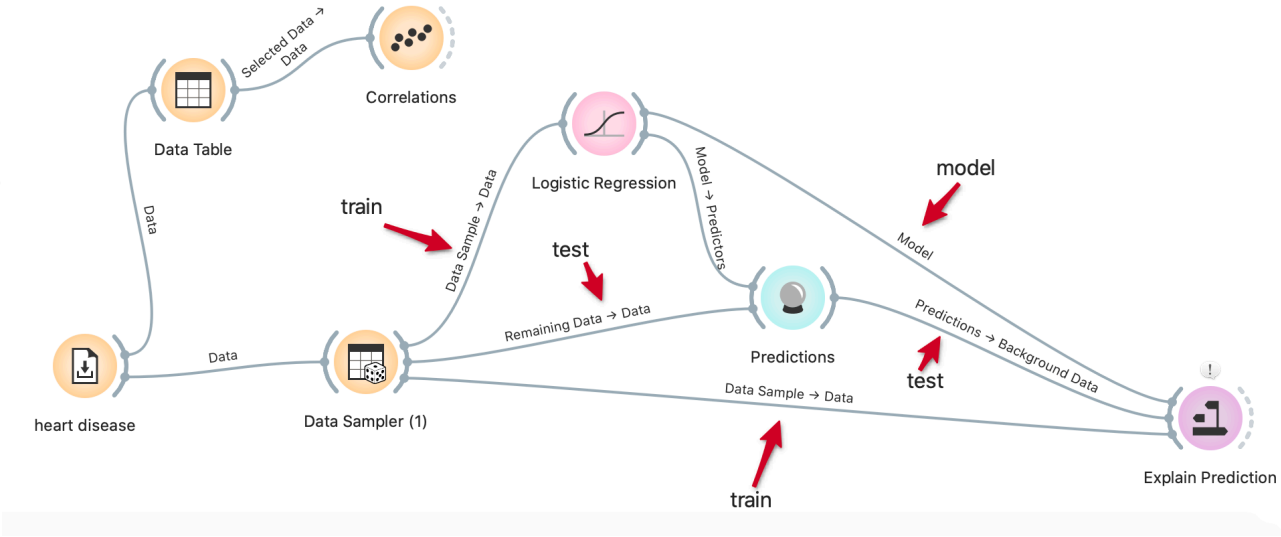


Inputs

- **Model**: a model whose predictions are explained by the widget
- **train data**: data needed to compute explanations
- **Test data**: Single data instance whose prediction is explained by the widget

Outputs

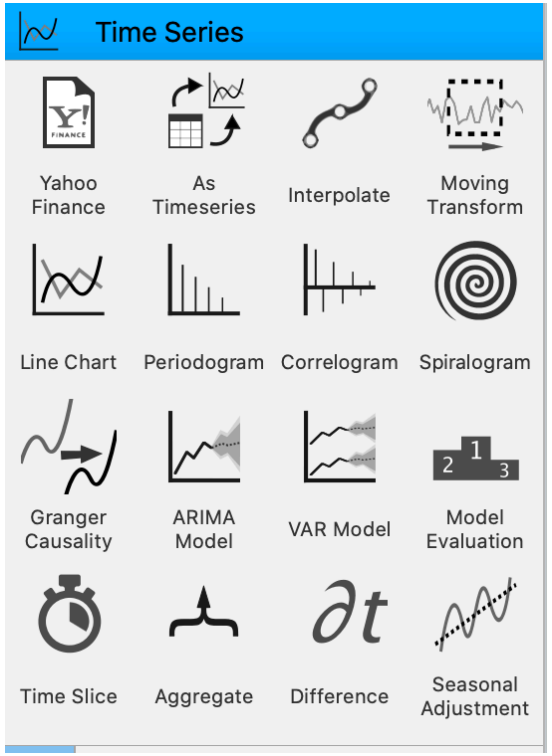
- Scores: **The SHAP value of each features value.** Features that contribute more to prediction have higher score deviation from the 0.

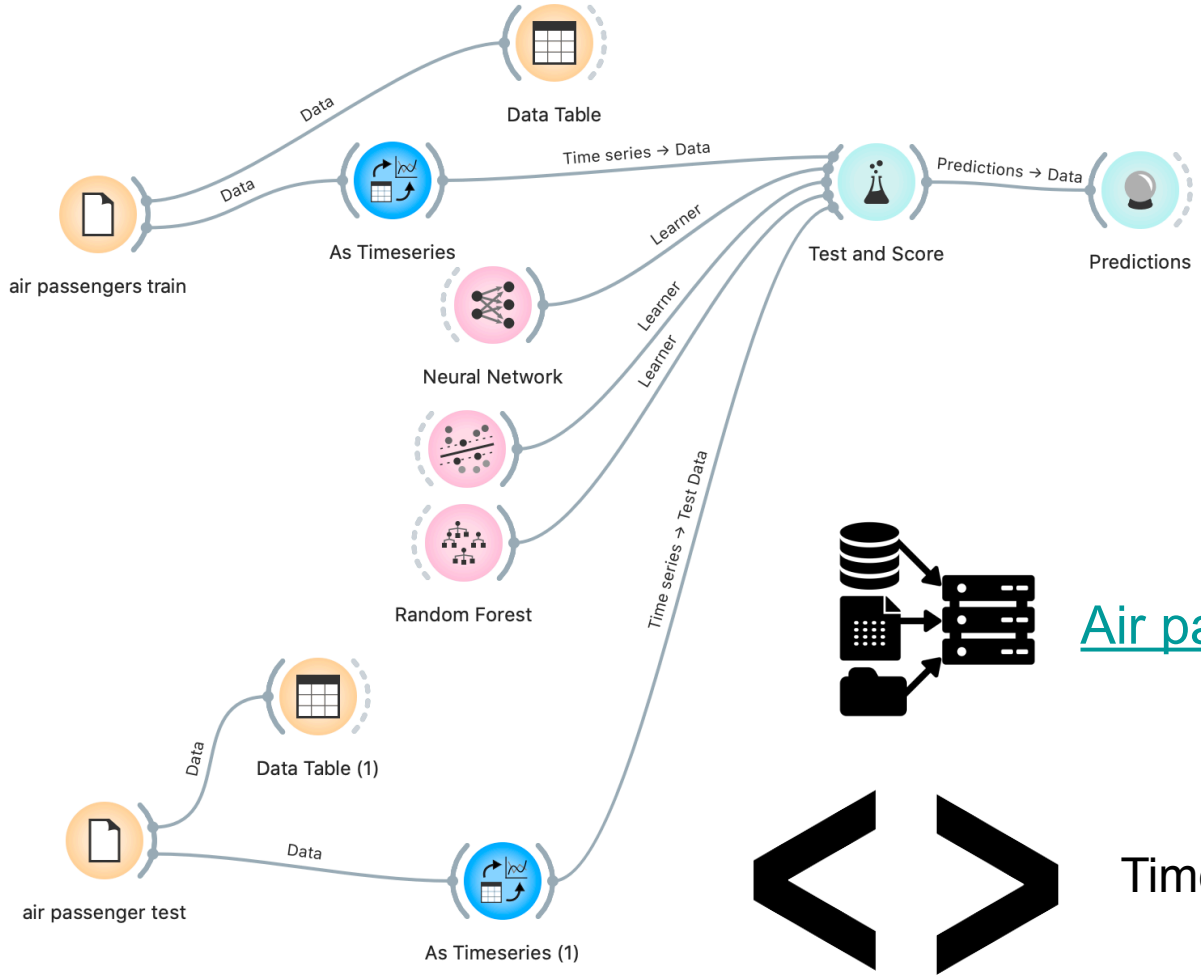


Explain predictions

VISUAL PROGRAMMING // **ORANGE**

Time series // NN MLP



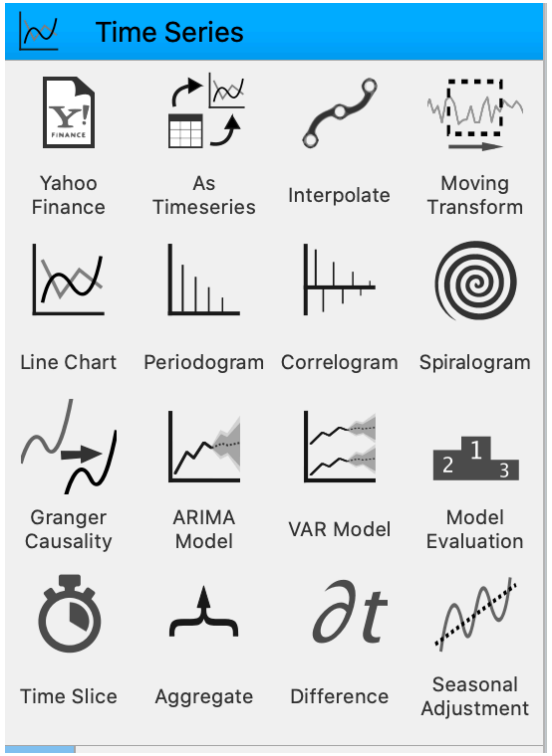


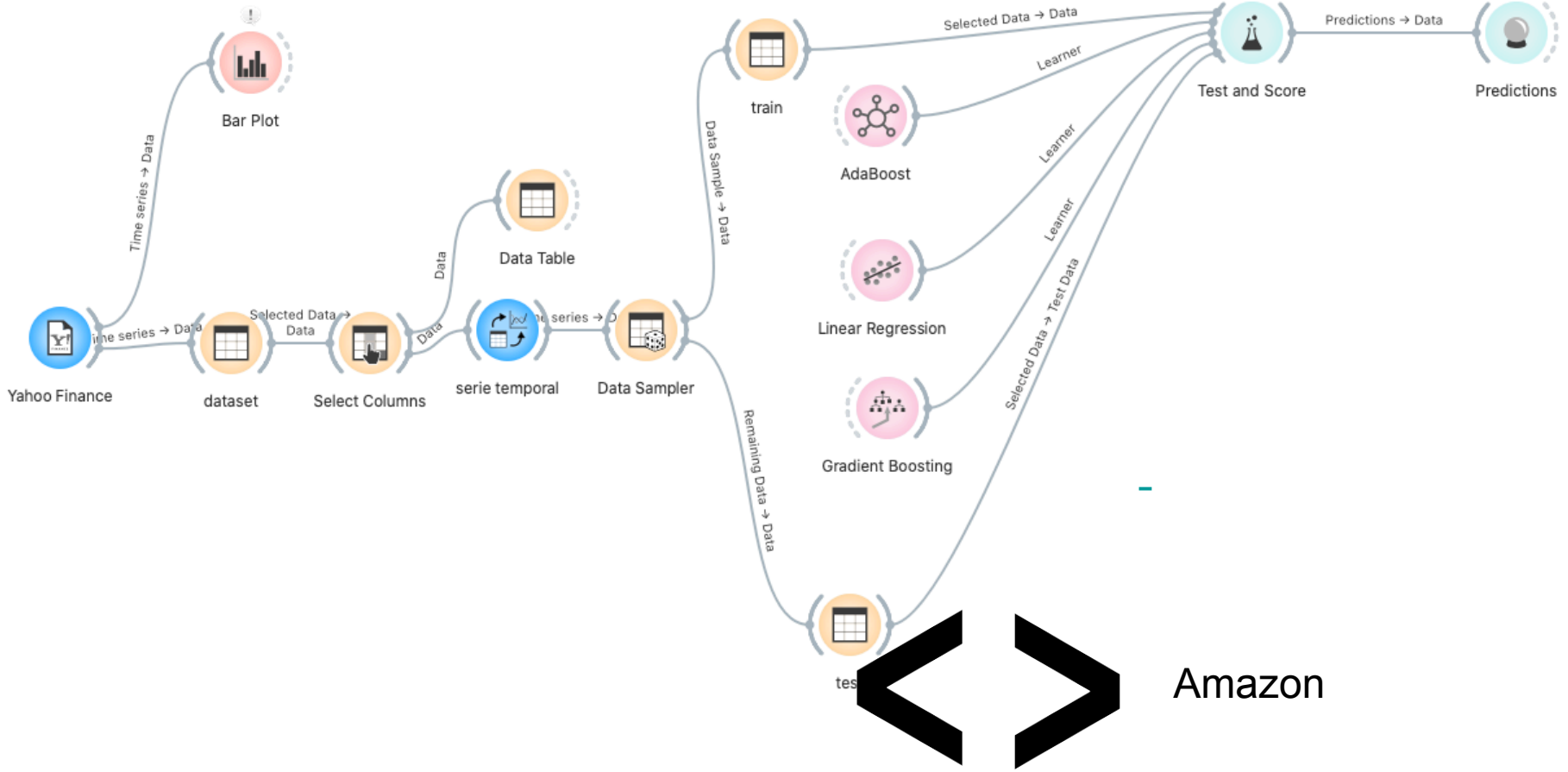
[Air passengers]

Time series NN

VISUAL PROGRAMMING // **ORANGE**

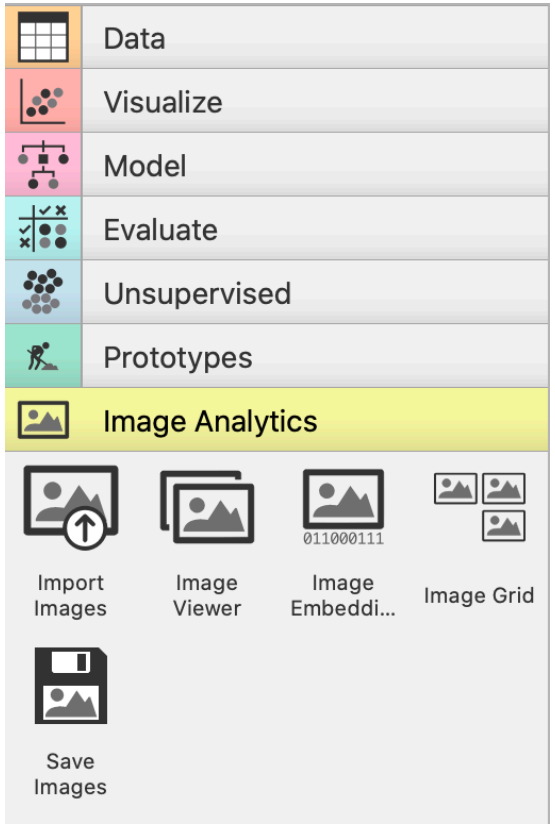
Time series // Stock market





Amazon

Image Analytics // Image Embedding



Import Images. Images can be imported with [Import Images](importimages.md) widget or as paths to images in a spreadsheet. In this case the column with images paths needs a three-row header with *type=image* label in the third row.

Image Embedding reads images and uploads them to a remote server or evaluate them locally. Deep learning models are used to calculate a feature vector for each image. It returns an enhanced data table with additional columns (image descriptors).

Image Embedding offers **several embedders**, each trained for a specific task.

- SqueezeNet: [Small and fast](https://arxiv.org/abs/1602.07360) model for image recognition trained on ImageNet.
- Inception v3: [Google's Inception v3](https://arxiv.org/abs/1512.00567) model trained on ImageNet.
- VGG-16: [16-layer image recognition model](https://arxiv.org/abs/1409.1556) trained on ImageNet.
- VGG-19: [19-layer image recognition model](https://arxiv.org/abs/1409.1556) trained on ImageNet.
- Painters: A model trained to [predict painters from artwork images](_).
- DeepLoc: A model trained to analyze [yeast cell images](https://www.ncbi.nlm.nih.gov/pubmed/29036616).

VISUAL PROGRAMMING // **ORANGE**

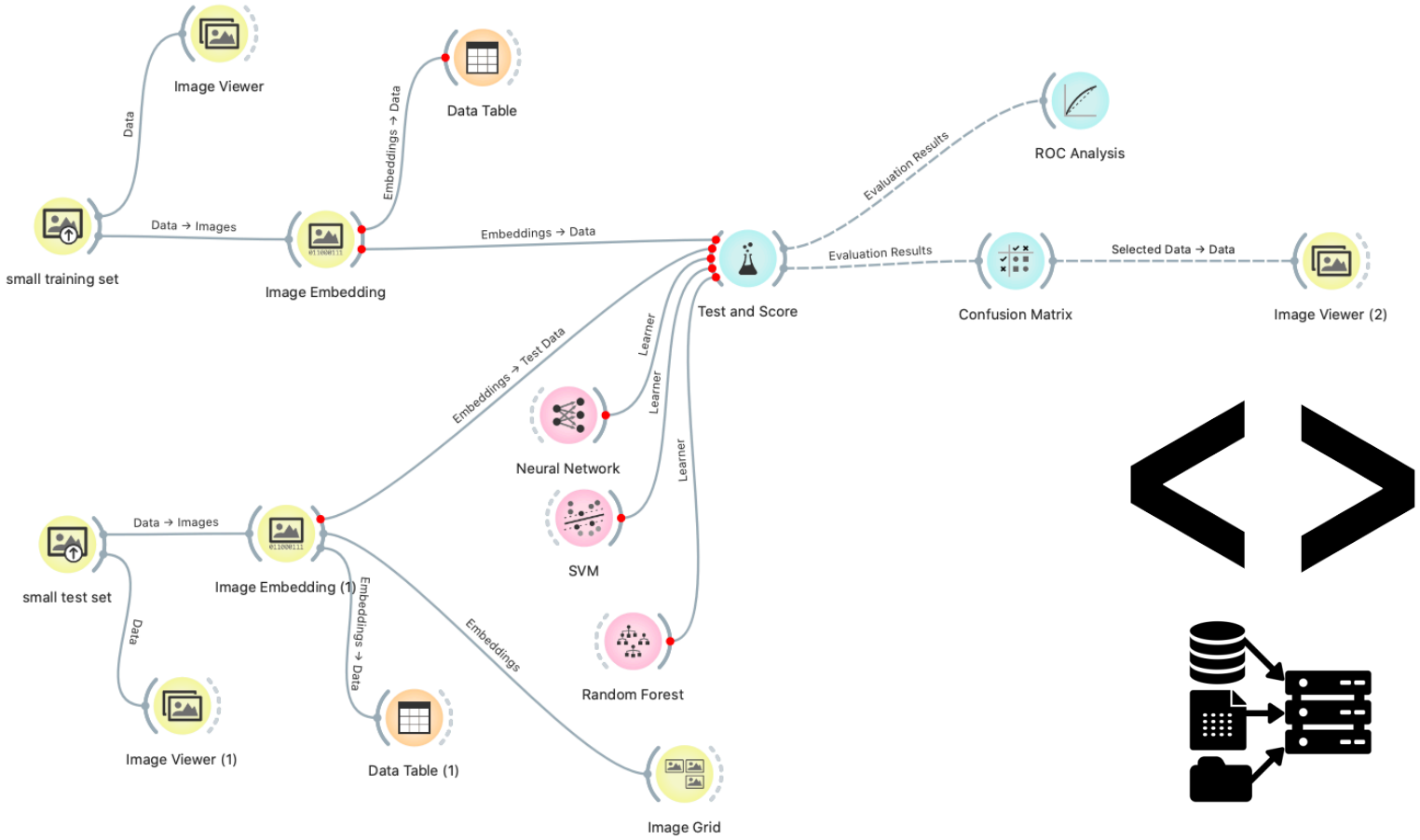


Image classification workflow

[Animals Image classification data]

VISUAL PROGRAMMING // **ORANGE**

Text Mining // Preprocess text

Preprocess Text splits your text into smaller units (tokens), filters them, runs normalization (stemming, lemmatization), creates n-grams and tags tokens with part-of-speech labels. Steps in the analysis are applied sequentially and can be reordered. Click and drag the preprocessor to change the order.

Transformation transforms input data. It applies lowercase transformation by default

Tokenization is the method of breaking the text into smaller components (words, sentences, bigrams).

This example. → (This), (example), (.)

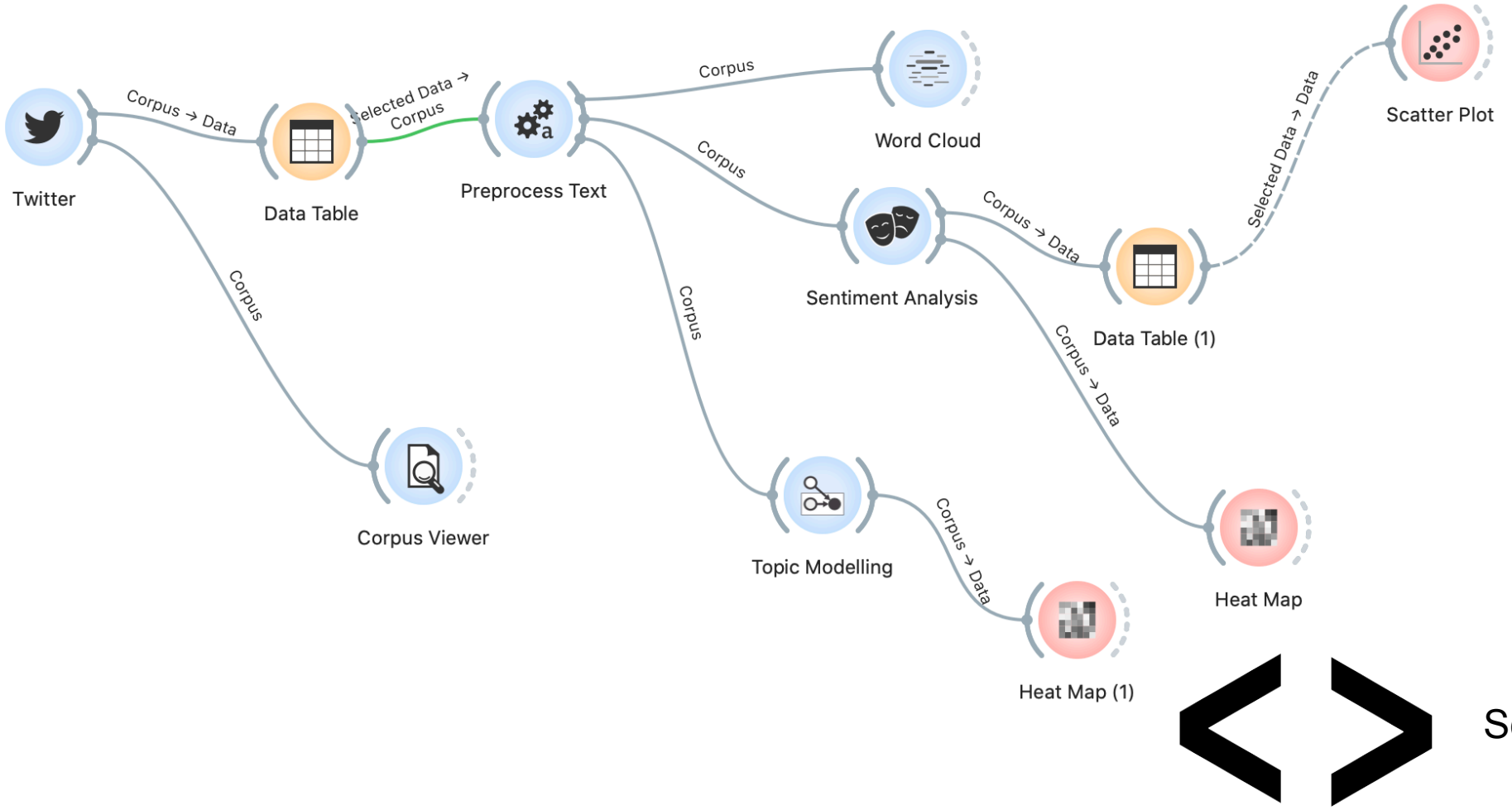
Normalization applies stemming and lemmatization to words.

(I've always loved cats. → I have alway love cat.)

Filtering removes or keeps a selection of words.

Stopwords removes stopwords from text (e.g. removes 'and', 'or', 'in'…)

VISUAL PROGRAMMING // **ORANGE**

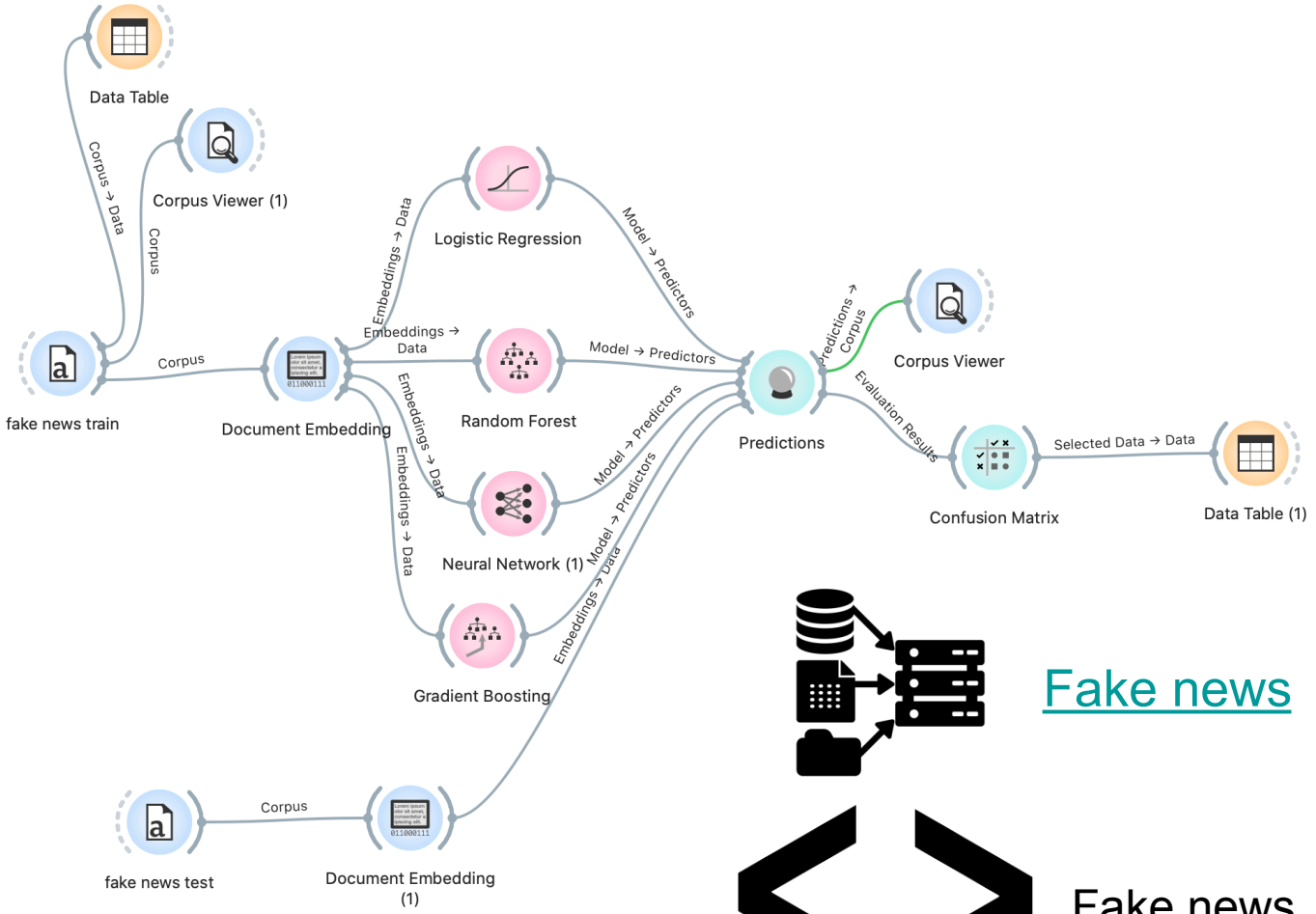


Sentiment analysis twitter

VISUAL PROGRAMMING // **ORANGE**

Document // embeddings

Orange now offers document embedders through Document Embedding widget. We decided to use fastText pretrained embedders, which support 157 languages. Orange's Document Embedding widget currently supports 31 most common languages.



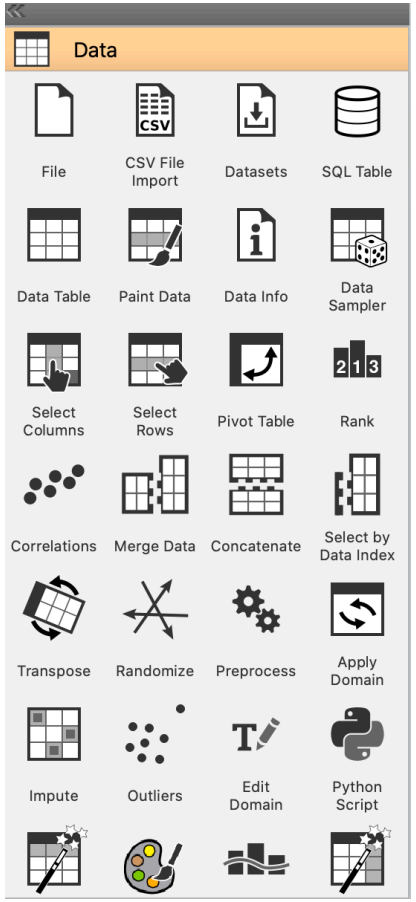
Fake news

Fake news

VISUAL PROGRAMMING // **ORANGE**

```
import numpy as np
out_data = in_data.copy()
#copy, otherwise input data will be overwritten
np.round(out_data.X, 0, out_data.X)
```

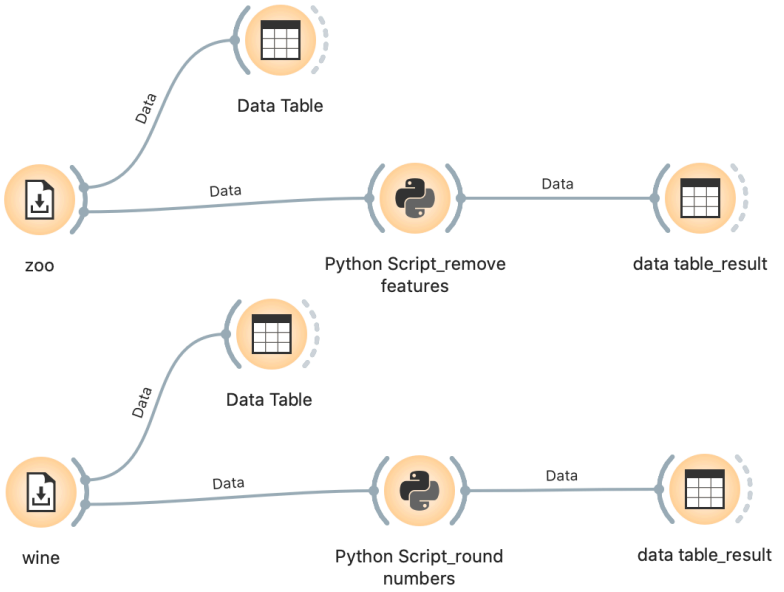
Data // Python script



After the script is executed variables from the script's local namespace are extracted and used as outputs of the widget. **The widget can be further connected to other widgets for visualizing the output.**

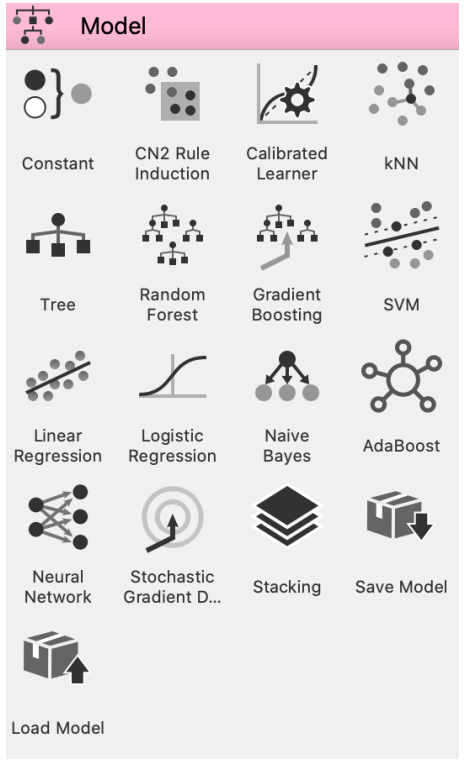
For instance the following **connectors** script would simply pass on all signals it receives:

```
out_data = in_data
out_distance = in_distance
out_learner = in_learner
out_classifier =
in_classifier
out_object = in_object
```

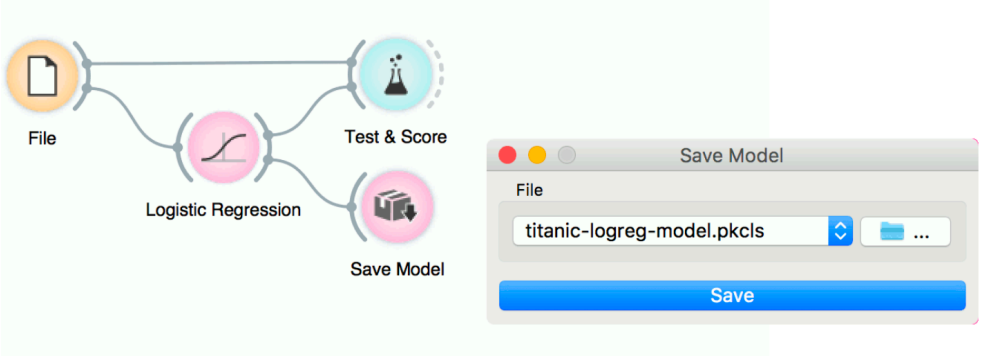


Python

Model // Save model



Can I export Orange workflow as a Python script?
Unfortunately, no. We've debated this long and hard. A (limited) functionality of this type would probably be possible, but would be a very big project to do well and would cost more than the expected benefits.



Use the Save Model widget to save the model into a pickle file, which you can then load into Python and use it to classify new data.
Assuming that the model is save to "my_model.pkcls" and your (new) data is in "my_data.tab", do this in Python:

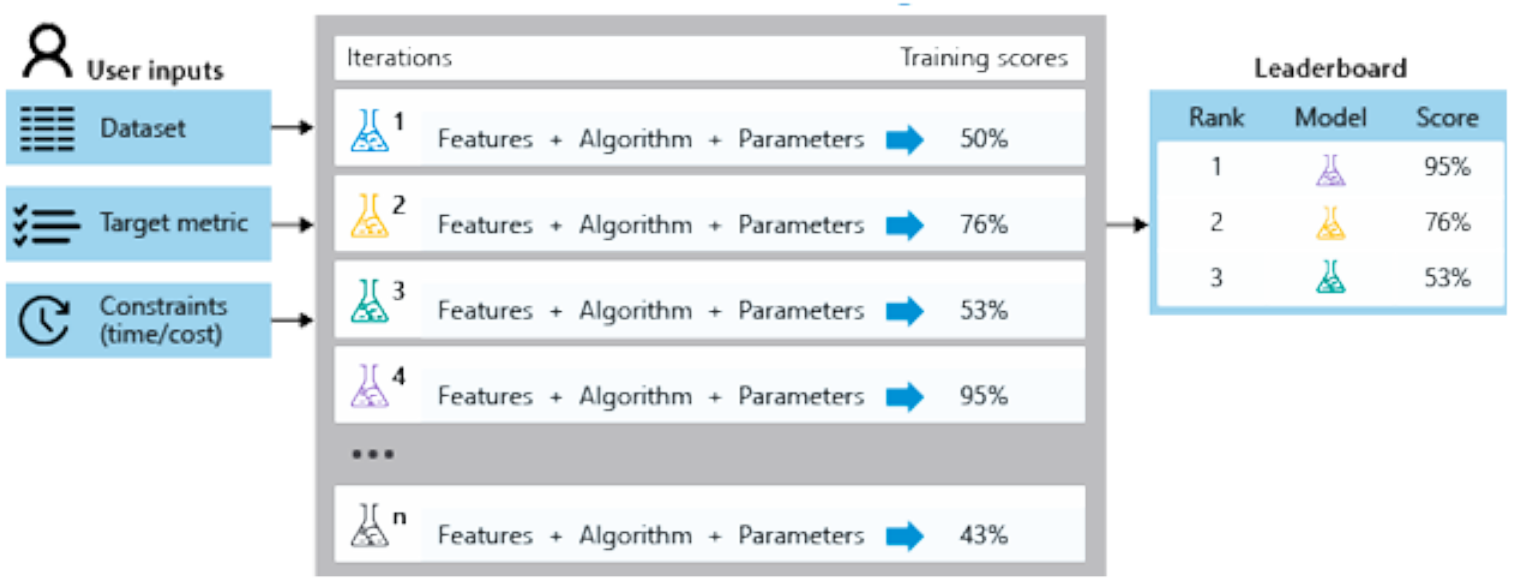
```
import Orange
import pickle

model = pickle.load(open("my_model.pkcls", "rb"))
data = Orange.data.Table("my_data.tab")
print(model(data))
```

Auto Machine Learning

AUTO ML //

- It improves the efficiency **by automating the most repetitive tasks.** This allows the data scientists to devote more time on the problems rather than on the models.



Menu Search + Add New Model Filter Models Export Metric LogLoss

Model Name & Description	Feature List & Sample Size	Validation	Holdout
eXtreme Gradient Boosted Trees Classifier with Early Stopping Single Column Converter | SqueezeNet Image Pretrained Featurizer | eXtreme Gradient Boosted Trees Classifier with Early Stopping M15 BP2 40.68% RECOMMENDED FOR DEPLOYMENT	Informative Features 85.0 %	0.1334	
eXtreme Gradient Boosted Trees Classifier with Early Stopping Single Column Converter | SqueezeNet Image Pretrained Featurizer | eXtreme Gradient Boosted Trees Classifier with Early Stopping M9 BP2 MOST ACCURATE	Informative Features 40.68 %	0.1322	
AVG Blender M19 M11+9+10	Informative Features 40.68 %	0.1476	
TensorFlow Deep Learning Classifier Single Column Converter | SqueezeNet Image Pretrained Featurizer | Standardize | TensorFlow Deep Learning Classifier | Calibrate predictions: Platt M10 BP3	Informative Features 40.68 %	0.2044	
Elastic-Net Classifier (L2 / Binomial Deviance) SqueezeNet Image Pretrained Featurizer | Elastic-Net Classifier (L2 / Binomial Deviance) M11 BP1	Informative Features 40.68 %	0.2084	

WORKERS

Using 0 of 10 total workers across all projects 10

STATUS

Autopilot has finished

ACTIONS

Rerun Autopilot

Unlock project Holdout for all models

AUTO ML //

*“One of the holy grails of machine learning is **to automate more and more of the feature engineering process.**”* — Pedro Domingos

AUTO ML //

AutoML refers to automated machine learning. It explains how the end to end process of machine learning can be automated at the organizational and educational level. The machine learning model includes basically the following steps :

1 // **Data reading and merging** and making it ready to use.
2 // Data preprocessing which refers data **cleaning and data wrangling.**
3 // **Optimization** where the feature and model selection process is done.
4 // Applying it to the application **to predict** the accurate values.

AUTO ML //

Initially all these steps were done manually. But now with the advent of the AutoML these steps can be automated. AutoML currently falls into three categories:

a // AutoML for automated **parameter tuning** (a relatively basic type)

b // AutoML for non-deep learning, for example, AutoSKlearn. This type is mainly applied in data pre-processing, automated feature analysis, automated feature detection, automated feature selection, and **automated model selection**.

c // AutoML for **deep learning/neural networks**, including NAS (Neural architecture search) and ENAS (Efficient Neural architecture search) as well as Auto-Keras for frameworks.

Why AutoML is raising ?

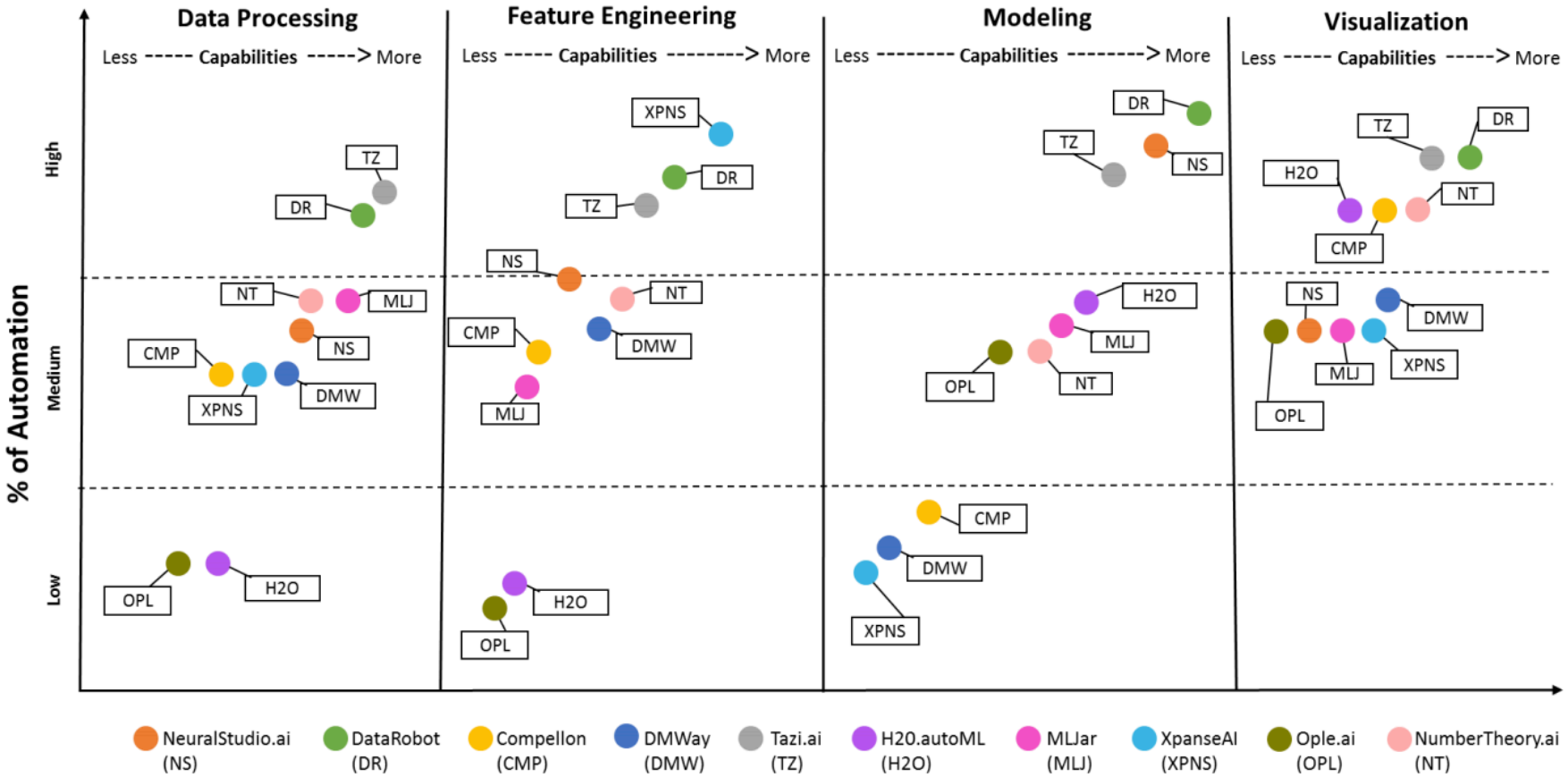
AutoML tends to automate as many steps as possible in ML pipelines and retain good model performance with **minimum manpower**.
It also provides methods and processes to:

- make machine learning **more accessible**
- **improve efficiency** of machine learning systems
- **accelerate** research and AI application development
- It improves the efficiency by automating the most **repetitive tasks**. This allows the data scientists to devote more time on the problems rather than on the models.
- Automated ML pipelines also help **avoid potential errors caused by manual work.**
- AutoML is a big step toward the **democratization** of machine learning and allows everyone to use ML features.

AUTO ML // players

Company	Total Funding Amount	Founded Year	Number of Employees
DataRobot	$430.6M	2012	1001-5000
H2O.ai	$151.1M	2012	11-50
Dataiku	$146.8M	2013	251-500
dotData	$43M	2018	51-100
Compellon (Acquired by Hoist Finance)	$11.6M	2010	11-50
Coldlight Solutions (Acquired by PTC)	$11M	2007	1-10
PurePredictive	$10.2M	2011	11-50
Ople	$10M	2017	11-50
Predikto (Acquired by United Technologies)	$7.6M	2013	11-50
VEDA Data Solutions	$7.2M	2015	1-10
Snark AI	$1.7M	2018	1-10
MyDataModels	$1M	2018	11-50
Tazi.ai	$1M	2015	11-50
DMWay	$1M	2013	1-10

AUTO ML // players



Auto ML.org

AUTO ML // players

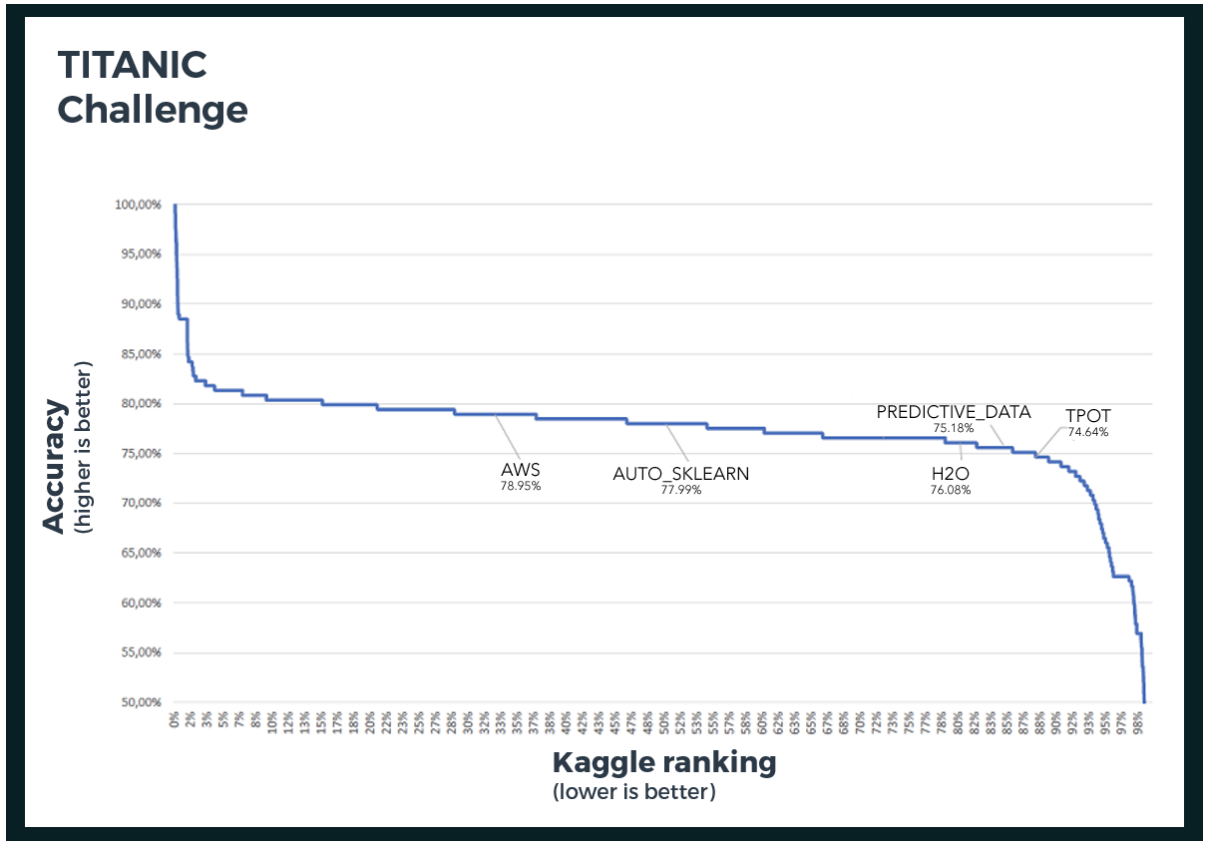


Figure 1: Magic Quadrant for Data Science and Machine Learning Platforms

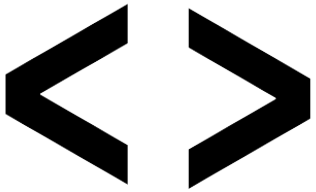
Source: Gartner (March 2021)

AUTO ML // Code

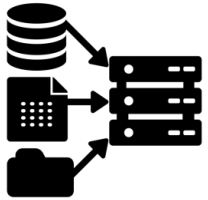
AutoML in code:

1 // **Tpot**
2 // **H2o**
3 //Auto **Keras** (Deep Learning)
4 // Auto **Sklearn**
5// **Auto Pytorch**
6// **HPOBench** (Hyperparameter optimization)

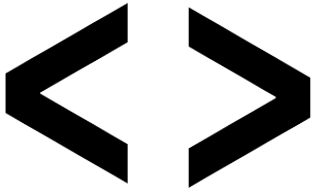
AUTO ML // Auto-ML



Auto-H2o.ipynb



Wine



Auto-Tpot & Auto-keras.ipynb

Hotel Workshop

WORKSHOP // **ALGARVE RESORT**

WORKSHOP // **ALGARVE RESORT**

Algarve dataset comprehend **bookings** due to arrive between the **1st of July of 2015 and the 31st of August 2017**, including bookings **that effectively arrived and bookings that were canceled**. Since this is hotel real data, all data elements pertaining hotel or costumer identification were deleted.

Table 2
H1 dataset summary statistics – Date variables.

Variable	Min	Max	Median	Unique
ReservationStatusDate	2014-11-18	2017-09-14	2016-07-31	913

Table 3
H1 dataset summary statistics – Categorical variables.

Variable	Unique	Top counts
Agent	186	240: 13 095, NULL: 8 209, 250: 2 869, 241: 1 721
ArrivalDateMonth	12	Aug: 4 894, Jul: 4 573, Apr: 3 609, May: 3 559
AssignedRoomType	11	A: 17 046, D: 10 339, E: 5 638, C: 2 214
Company	236	NULL: 36 952, 223: 784, 281: 138, 154: 133
Country	125	PRT: 17 630, GBR: 6 814, ESP: 3 957, IRL: 2 166
CustomerType	4	Tra.: 30 209, Tra.-Party: 7 791, Con.: 1 776, Gro.:284
DepositType	3	No Dep.: 38 199, Non-Refund.: 1 719, Ref.: 142
DistributionChannel	4	TA/TO: 28 295, Dir.: 7 865, Cor.: 3 269, Und.: 1
IsCanceled	2	0: 28 938, 1: 11 122
IsRepeatedGuest	2	0: 38 282, 1: 1 778
MarketSegment	6	Onl.: 17 729, Off.: 7472, Dir.: 6 513, Gro.: 5 836
Meal	5	BB: 30 005, HB: 8 046, Und.: 1 169, FB: 754
ReservationStatus	3	C.Out: 28 938, Can.: 10 831, No-Show: 291
ReservedRoomType	10	A: 23 399, D: 7 433, E: 4 892, G: 1610

Table 4
H1 dataset summary statistics – Integer and numeric variables.

Variable	Mean	SD	P0	P25	Median	P75	P100
ADR	94.95	61.44	-6.38	50	75	125	508
Adults	1.87	0.7	0	2	2	2	55
ArrivalDateOfMonth	15.82	8.88	1	8	16	24	31
ArrivalDateWeekNumber	27.14	14.01	1	16	28	38	53
ArrivalDateYear	2016.12	0.72	2015	2016	2016	2017	2017
Babies	0.014	0.12	0	0	0	0	2
BookingChanges	0.29	0.73	0	0	0	0	17
Children	0.13	0.45	0	0	0	0	10
DaysInWaitingList	0.53	7.43	0	0	0	0	185
LeadTime	92.68	97.29	0	10	57	155	737
PreviousBookingsNotCanceled	0.15	1	0	0	0	0	30
PreviousCancellations	0.1	1.34	0	0	0	0	26
RequiredCarParkingSpaces	0.14	0.35	0	0	0	0	8
StaysInWeekendNights	1.19	1.15	0	0	1	2	19
StaysInWeekNights	3.13	2.46	0	1	3	5	50
TotalOfSpecialRequests	0.62	0.81	0	0	0	1	5

Table 5
H2 dataset summary statistics – Date variables.

Variable	Min	Max	Median	Unique
ReservationStatusDate	2014-10-17	2017-09-07	2016-08-10	864

Table 6
H2 dataset summary statistics – Categorical variables.

Variable	Unique	Top counts
Agent	224	9: 31 955, NULL: 8 131, 1: 7 137, 14: 3 640
ArrivalDateMonth	12	Aug: 8 983, May: 8 232, Jul: 8 088, Jun: 7 894
AssignedRoomType	9	A: 57 007, D: 14 983, E: 2 168, F: 2 018
Company	208	NULL: 75 641, 40: 924, 67: 267, 45: 250
Country	166	PRT: 30 960, FRA: 8 804, DEU: 6 084, GBR: 5315
CustomerType	4	Tra.:59 404, Tra.-P.: 17 333, Con.: 2 300, Gro.:293
DepositType	3	No Dep.: 66 442, Non-Refund.: 12 868, Ref.: 20
DistributionChannel	5	TA/TO: 68 945, Dir.: 6 780, Cor.: 3 408, GDS: 193
IsCanceled	2	0: 46 228, 1: 33 102
IsRepeatedGuest	2	0: 77 298, 1: 2 032
MarketSegment	8	Onl.: 38 748, Off.: 16 747, Gro.: 13 975, Dir.: 6 093
Meal	4	BB: 62 305, SC: 10 564, HB: 6 417, FB: 44
ReservationStatus	3	C.Out: 46 228, Can.: 32 186, No-Show: 916
ReservedRoomType	8	A: 62 595, D: 11768, F: 1 791, E: 1 553

Table 7
H2 dataset summary statistics – Integer and numeric variables.

Variable	Mean	SD	P0	P25	Median	P75	P100
ADR	105.3	43.6	0	79.2	99.9	126	5400
Adults	1.85	0.51	0	2	2	2	4
ArrivalDateOfMonth	15.79	8.73	1	8	16	23	31
ArrivalDateWeekNumber	27.18	13.4	1	17	27	38	53
ArrivalDateYear	2016.17	0.7	2015	2016	2016	2017	2017
Babies	0.0049	0.084	0	0	0	0	10
BookingChanges	0.19	0.61	0	0	0	0	21
Children	0.091	0.37	0	0	0	0	3
DaysInWaitingList	3.23	20.87	0	0	0	0	391
LeadTime	109.74	110.95	0	23	74	163	629
PreviousBookingsNotCanceled	0.13	1.69	0	0	0	0	72
PreviousCancellations	0.08	0.42	0	0	0	0	32
RequiredCarParkingSpaces	0.024	0.15	0	0	0	0	3
StaysInWeekendNights	0.8	0.89	0	0	1	2	16
StaysInWeekNights	2.18	1.46	0	1	2	3	41
TotalOfSpecialRequests	0.55	0.78	0	0	0	1	5

WORKSHOP // **ALGARVE RESORT**

Variable	Type	Description	Source/Engineering
ADR	Numeric	Average Daily Rate as defined by [5]	BO, BL and TR / Calculated by dividing the sum of all lodging transactions by the total number of staying nights
Adults	Integer	Number of adults	BO and BL
Agent	Categorical	ID of the travel agency that made the booking[a]	BO and BL
ArrivalDateDayOfMonth	Integer	Day of the month of the arrival date	BO and BL
ArrivalDateMonth	Categorical	Month of arrival date with 12 categories: "January" to "December"	BO and BL
ArrivalDateWeekNumber	Integer	Week number of the arrival date	BO and BL
ArrivalDateYear	Integer	Year of arrival date	BO and BL
AssignedRoomType	Categorical	Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons	BO and BL
Babies	Integer	Number of babies	BO and BL
BookingChanges	Integer	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation	BO and BL/Calculated by adding the number of unique iterations that change some of the booking attributes, namely: persons, arrival date, nights, reserved room type or meal
Children	Integer	Number of children	BO and BL/Sum of both payable and non-payable children
Company	Categorical	ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons	BO and BL.
Country	Categorical	Country of origin. Categories are represented in the ISO 3155–3:2013 format [6]	BO, BL and NT
CustomerType	Categorical	Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking	BO and BL
DaysInWaitingList	Integer	Number of days the booking was in the waiting list before it was confirmed to the customer	BO/Calculated by subtracting the date the booking was confirmed to the customer from the date the booking entered on the PMS
DepositType	Categorical	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay.	BO and TR/Value calculated based on the payments identified for the booking in the transaction (TR) table before the booking's arrival or cancellation date. In case no payments were found the value is "No Deposit". If the payment was equal or exceeded the total cost of stay, the value is set as "Non Refund". Otherwise the value is set as "Refundable"

WORKSHOP // **ALGARVE RESORT**

Variable	Type	Description	Source/Engineering
DistributionChannel	Categorical	Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"	BO, BL and DC
IsCanceled	Categorical	Value indicating if the booking was canceled (1) or not (0)	BO
IsRepeatedGuest	Categorical	Value indicating if the booking name was from a repeated guest (1) or not (0)	BO, BL and C/ Variable created by verifying if a profile was associated with the booking customer. If so, and if the customer profile creation date was prior to the creation date for the booking on the PMS database it was assumed the booking was from a repeated guest
LeadTime	Integer	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date	BO and BL/ Subtraction of the entering date from the arrival date
MarketSegment	Categorical	Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"	BO, BL and MS
Meal	Categorical	Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)	BO, BL and ML
PreviousBookingsNotCanceled	Integer	Number of previous bookings not cancelled by the customer prior to the current booking	BO and BL / In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and not canceled.
PreviousCancellations	Integer	Number of previous bookings that were cancelled by the customer prior to the current booking	BO and BL/ In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and canceled.
RequiredCardParkingSpaces	Integer	Number of car parking spaces required by the customer	BO and BL
ReservationStatus	Categorical	Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed; No-Show – customer did not check-in and did inform the hotel of the reason why	BO

Variable	Type	Description	Source/Engineering
ReservationStatusDate	Date	Date at which the last status was set. This variable can be used in conjunction with the *ReservationStatus* to understand when was the booking canceled or when did the customer checked-out of the hotel	BO
ReservedRoomType	Categorical	Code of room type reserved. Code is presented instead of designation for anonymity reasons	BO and BL
StaysInWeekendNights	Integer	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel	BO and BL/ Calculated by counting the number of weekend nights from the total number of nights
StaysInWeekNights	Integer	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel	BO and BL/Calculated by counting the number of week nights from the total number of nights
TotalOfSpecialRequests	Integer	Number of special requests made by the customer (e.g. twin bed or high floor)	BO and BL/Sum of all special requests

[a] ID is presented instead of designation for anonymity reasons.

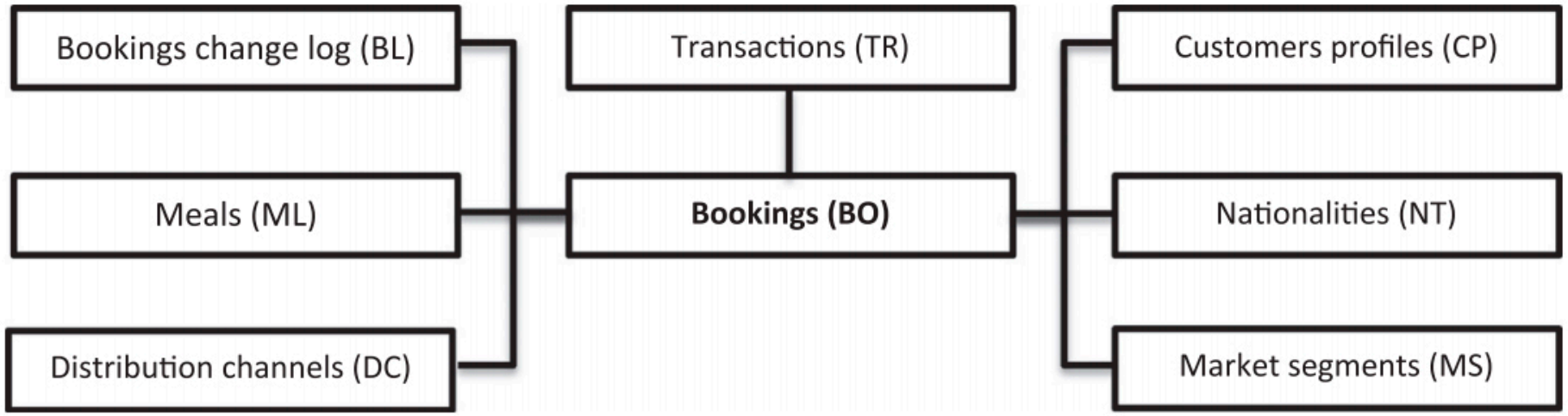


Fig. 1. Diagram of PMS database tables where variables where extracted from.

WORKSHOP // **XTRA INFO**

About Carnival / Shrove Tuesday

Read more about Carnival / Shrove Tuesday.

Carnival / Shrove Tuesday Observances

Showing: 2015–2025

Year	Weekday	Date	Name	Holiday Type
2015	mar	17 de feb	Carnival / Shrove Tuesday	Optional Holiday
2016	mar	9 de feb	Carnival / Shrove Tuesday	Optional Holiday
2017	mar	28 de feb	Carnival / Shrove Tuesday	Optional Holiday
2018	mar	13 de feb	Carnival / Shrove Tuesday	Optional Holiday
2019	mar	5 de mar	Carnival / Shrove Tuesday	Optional Holiday
2020	mar	25 de feb	Carnival / Shrove Tuesday	Optional Holiday
2021	mar	16 de feb	Carnival / Shrove Tuesday	Optional Holiday
2022	mar	1 de mar	Carnival / Shrove Tuesday	Optional Holiday
2023	mar	21 de feb	Carnival / Shrove Tuesday	Optional Holiday
2024	mar	13 de feb	Carnival / Shrove Tuesday	Optional Holiday
2025	mar	4 de mar	Carnival / Shrove Tuesday	Optional Holiday

We diligently research and continuously update our holiday dates and information. If you find a mistake, please let us know.

February: Carnival

You'll find ***Carnaval* parades all over Portugal**, with **Lisbon** and the towns of **the Algarve** having particularly spectacular celebrations. While it may seem to be all Rio-style feathers, spandex, and sequins, *Carnaval* festivals in Portugal date back centuries to when people held huge feasts to eat up all the meat, which was forbidden during Lent. Traditionally, *Carnaval* begins on the last Friday before Lent and ends on Shrove Tuesday.

WORKSHOP // **XTRA INFO**

Portuguese public holidays during 2022

- **1 January (Saturday):** New Year's Day *(Ano Novo)*
- **15 April (Friday):** Good Friday *(Sexta-feira Santa)*
- **17 April (Sunday):** Easter Sunday
- **25 April (Monday):** Freedom Day *(Dia da Liberdade)*
- **1 May (Sunday):** Labor Day *(Dia do Trabalhador)*
- **10 June (Friday):** Portugal National Day (*Dia de Portugal*)
- **16 June (Thursday):** Corpus Christi *(Corpo de Deus)*
- **15 August (Monday):** Assumption of Mary *(Assunção de Nossa Senhora)*
- **5 October (Wednesday):** Republic Day *(Implantação da República)*
- **1 November (Tuesday):** All Saints' Day *(Todos os Santos)*
- **1 December (Thursday):** Restoration of Independence Day (*Restauração da Independência*)
- **8 December (Thursday):** Immaculate Conception *(Imaculada Conceição)*
- **25 December (Sunday):** Christmas Day *(Natal)*

Algarve Monthly temperatures 2015 - 2019

January | February | March | April | May | June | July | August | September | October | November | December

	January	February	March	April	May	June	July	August	September	October	November	December
Daytime Temperature	20℃	19℃	20℃	23℃	26℃	32℃	34℃	33℃	31℃	26℃	23℃	22℃
Night-time Temperature	9℃	8℃	9℃	12℃	14℃	18℃	19℃	19℃	17℃	15℃	11℃	10℃
Rainy days	3	5	11	2	6	0	0	0	0	12	7	7
Snow days	0	0	0	0	0	0	0	0	0	0	0	0

WORKSHOP // **PRESENTACIÓN**

1 // Descarga el dataset desde el siguiente enlace: https://drive.google.com/drive/folders/122xbqsjVLXd_2IFjUv_OUElyqhp6zjM-?usp=sharing

2 //**Analiza los datos y visualiza** el dataset H1.csv

3 // Aplica el proceso de *ETL*.

4 // Realiza una **segmentación de clientes**. Justifica tus decisiones respecto al número de *clústers*.

5 // **Elabora un algoritmo clasificador** para predecir si un cliente va a cancelar o no su reserva. *(Is canceled)*

5.1// Muestra la **feature importance**

6 // Desarrolla un algoritmo predictivo para saber cuánto va a consumir cada cliente *(ADR) (opcional)*

7 // **Aporta métricas de evaluación** de los algoritmos descartados y seleccionados.

8 // **Justifica tus decisiones** y **presenta en clase** por grupos de Data Project.