
Prevision.io Documentation

Gerome Pistre

juil. 02, 2021

| | | |
|----------|---|----------|
| 1 | Présentation | 1 |
| 1.1 | Introduction | 1 |
| 1.1.1 | value proposition | 1 |
| 1.1.2 | requirements | 1 |
| 1.1.3 | conditions | 1 |
| 1.1.4 | contacts | 1 |
| 1.2 | Getting started | 2 |
| 1.2.1 | Account creation | 2 |
| 1.2.2 | Connection | 2 |
| 1.2.3 | Cloud & freetrial limitations | 2 |
| 2 | Studio | 5 |
| 2.1 | Navigation | 5 |
| 2.2 | Concepts | 8 |
| 2.2.1 | Projects | 8 |
| 2.2.2 | Data | 11 |
| 2.2.3 | Use cases | 22 |
| 2.2.4 | Pipelines | 52 |
| 2.2.5 | Contributors | 53 |
| 2.2.6 | Notebooks | 54 |
| 2.2.7 | Help | 55 |
| 2.2.8 | User | 55 |

1.1 Introduction

1.1.1 value proposition

AI in the enterprise promises leaps in efficiency, business innovation and customer-facing performance. However, to enable greater adoption, democratization and acceptance of AI, organizations must overcome not only talent gaps, but deployment, usability and governance gaps too. The real enabler of enterprise AI is the removal of friction between Business Departments, Data Science, and IT users. Data science is following in the path of software development : Integrated environments, agile, iterative methods, and move to modularity and no-code approaches that empower both expert and citizen developers and data scientists alike. The winners in the Enterprise AI revolution are those who can achieve faster, more agile, more integrated, more collaborative production cycles across DataOps, ML Ops and DevOps areas. Prevision is an end-to-end enterprise AI platform, specifically designed to enable business users, data scientists, and developers to deliver AI projects that deliver ROI, faster. It streamlines the creation, deployment and management of AI-powered business applications across their full lifecycle.

1.1.2 requirements

Prevision.IA is a SAAS platform optimized for Firefox and Chrome navigators. The cloud version can be accessed online at (<https://cloud.prevision.io>), or it can be deployed on-premise or in your private cloud. Please contact us at (support@prevision.io) if you have any questions regarding our deployment capabilities.

1.1.3 conditions

Please read the general terms and conditions available on following link : <https://cloud.prevision.io/terms> :

1.1.4 contacts

If you have any questions about using Prevision.IO platform please contact us using the chat button on the Prevision.IO store interface or by email at the following contact address :

support@prevision.io

1.2 Getting started

1.2.1 Account creation

By clicking to the following address, you will land on the connection page which allows you to create an account or register if you already have a Prevision.IO account.

<https://cloud.prevision.io>

In order to create a new account, just click on the sign up button next to the log in button. You will access the following account creation form.

Once you have fulfilled the needed information, you will have a 30 days free but limited access to the Prevision.IO platform. In order to upgrade your free trial account into a full access one, please contact us on following email (support@prevision.io)

1.2.2 Connection

Once your account has been created, you will be able to access the prevision's Studio and Store and start creating models and deploying them.

Please note that SSO using your google/linkedin/GitHub account is available.

1.2.3 Cloud & freetrial limitations

If you are using our cloud platform (<https://cloud.prevision.io>) using a free trial account, some limitations are setted up. Here a quick view of limitations for free testing accounts :

Tableau 1 – Freetrial Limitation

| Entity | Action | Limitation |
|------------------|---|--|
| PROJECT | Create Project | Free trial users can create 2 Limited Projects |
| DATASETS | Add dataset from file / datasource | 10 Datasets max + 1GB per dataset |
| IMAGE FOLDER | Add / Update / Delete image Folder in project | 1 Image Folder |
| DATA SOURCE | Add / Update / Delete datasource in project | 1 Datasource max |
| CONNECTOR | Add / Update / Delete connector in project | 1 Connector max |
| USECASE | Add / Update / Delete usecase in project | 5 Use Cases max |
| USECASE VER-SION | Add / Update usecase version in project | 3 Concurrent usecase versions |
| PREDICTION | Add / Update prediction in usecase version | 2 Concurrent predictions |
| STORE APP | Deploy Apps | 5 Concurrent deployed apps |

[English](#)

LOG IN. SIGN UP.

Register for a free 30-days trial period

First name

Last name

Country

Job title

Enterprise

Sector

Email

Password

Confirm password

☐ Je ne suis pas un robot



reCAPTCHA
Confidentialité - Conditions

Fig. 1 – image alt text

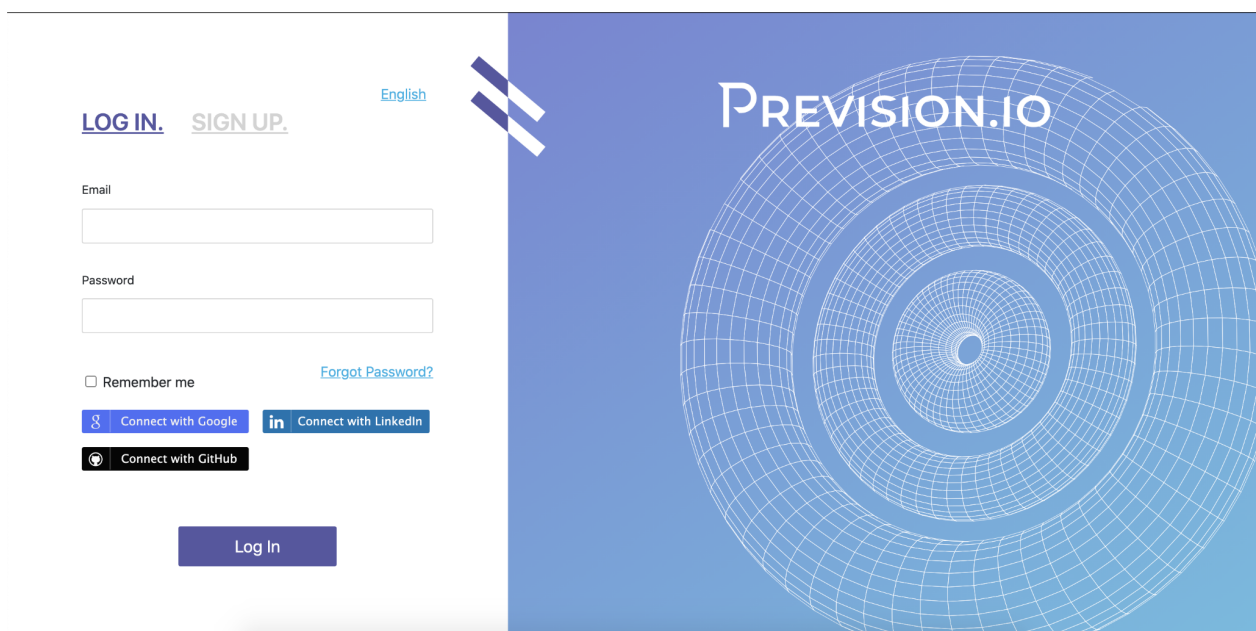


Fig. 2 – image alt text

2.1 Navigation

Several levels of navigation are displayed on the interface in order for you to navigate into the different projects and resources created on the platform.

The first level of navigation is the left navigation bar. You will find the following menu :

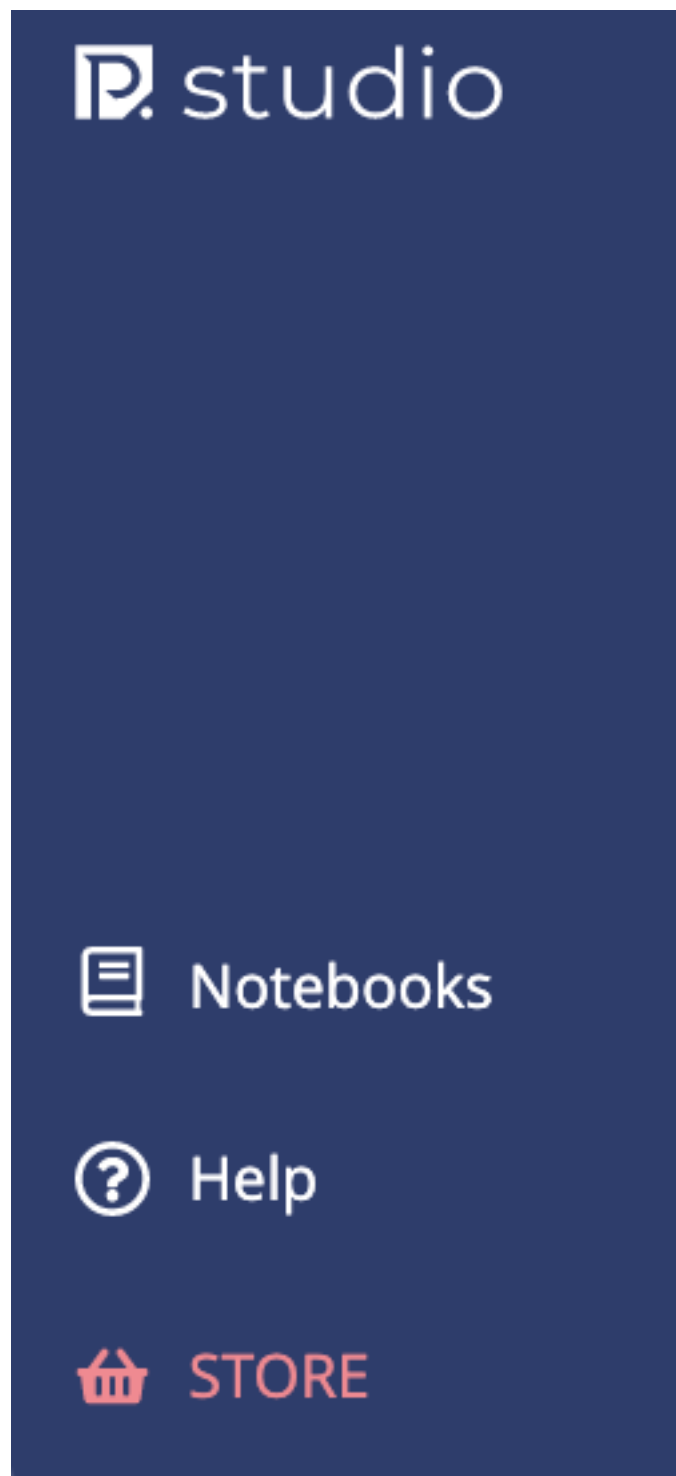
- Studio : list of all available projects
- Notebooks : access to the notebooks interface
- Help : documentation and examples of prevision.IO possibilities
- Store : opening a new tab to the prevision's store

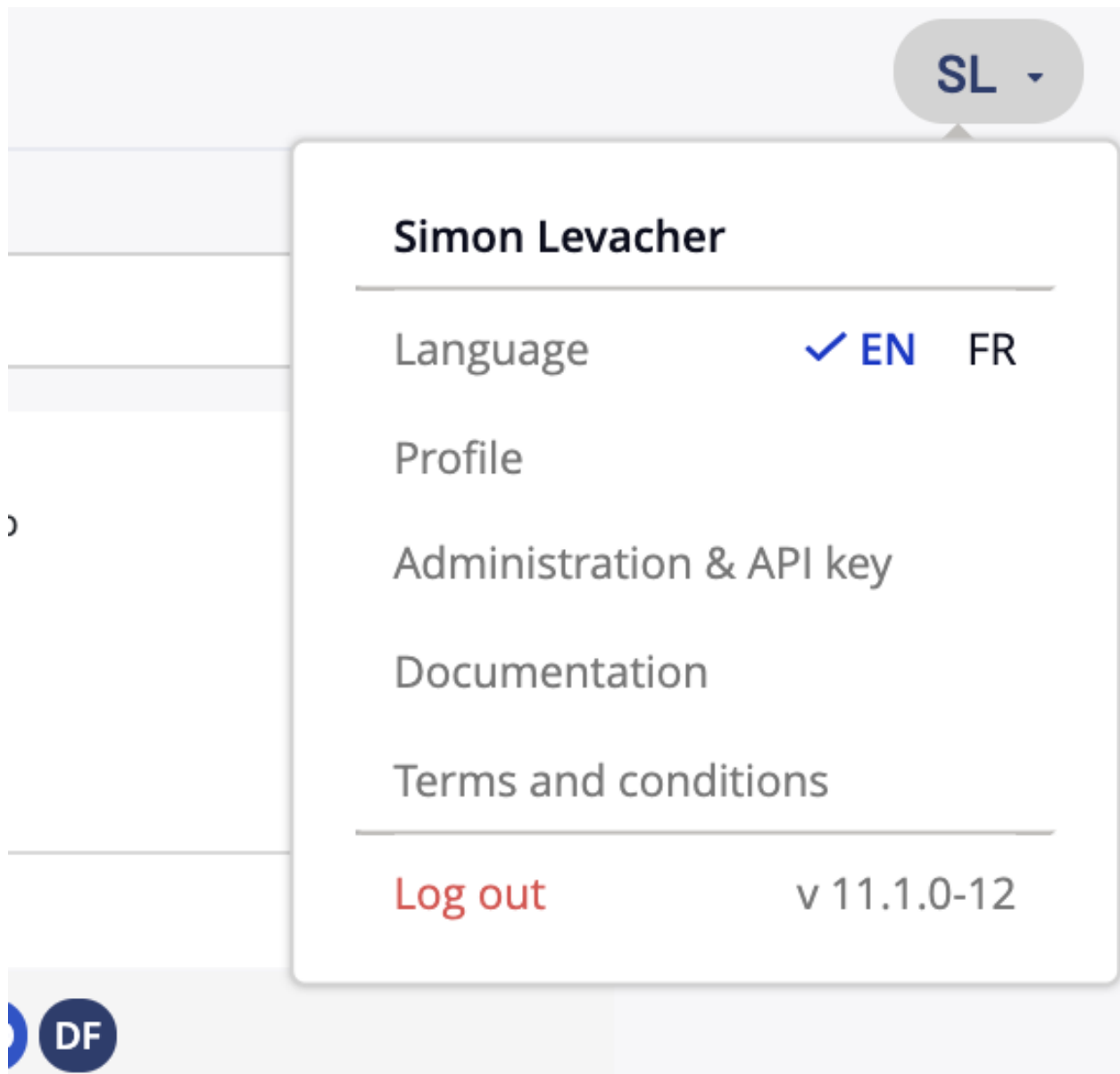
Note : Tips : you can collapse/expand the left menu by clicking on the arrow button next to prevision studio logo on top of the menu

This menu is evolutive, once you will get into a project, resources available for this project will be reachable thanks to this main navigation menu.

By clicking on the top right user icon, a sub-navigation will appear

- Language : you can switch between french and english version
- Profile : access to your personal profile information
- Administration & API key : access the API key configuration and the admin screen (if you have the admin rights)
- Documentation : access to platform documentation
- Terms and conditions : display terms and conditions of Prevision.IO platform
- Log out : logging out the platform





2.2 Concepts

2.2.1 Projects

Introduction

In Prevision.IO studio, ressources, such as datasets or models, are scoped by project in order to structure your work and collaborate easily with people inside a project.

Create a new project

In order to create a new project, you have to click on the “new project” button on top right of the “my projects” view. You will access to the following interface :

In order to create your project you have to fulfill at least a color and a project name. You can also add a description of your project.

Please note that you can at any moment, if admin role into a project has been setted up for your account, change these information by clicking on “settings” into the project menu.

All your projects will be displayed in the “my projects” view. Two different displays are list view and cards view and you can switch between one view and another by clicking on the view button you preferre next to the search bar.

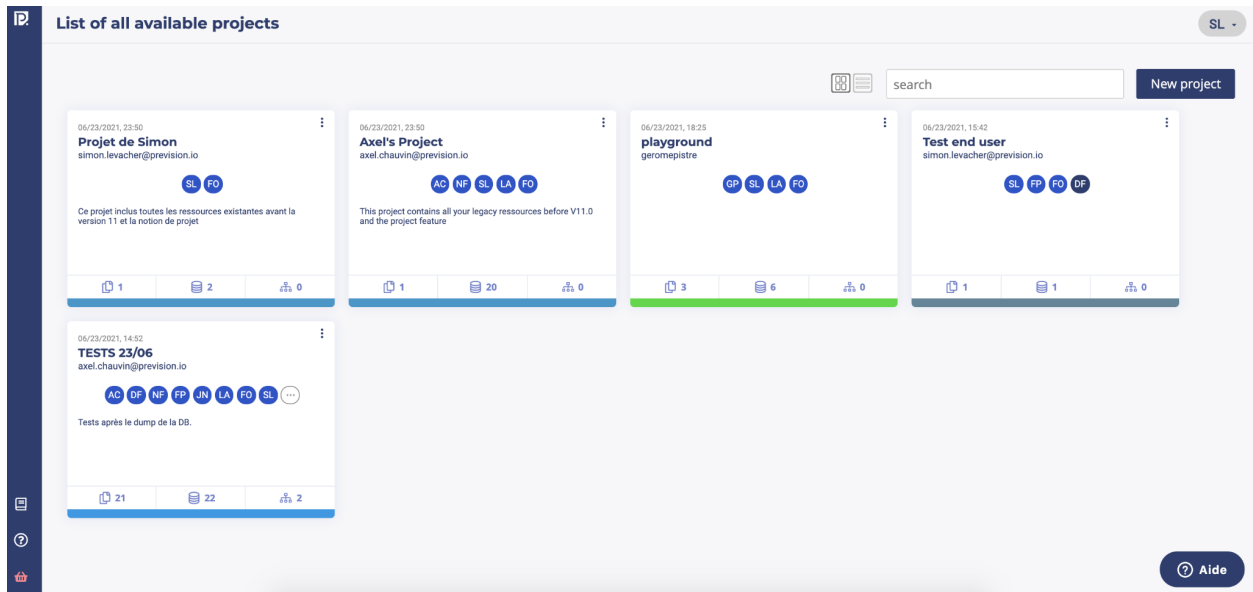
View cards

The view card is displaying all your projects as cards.

You will find the following information on the cards :

- project name
- created by and creating at
- description (if available)
- number of datasets/pipelines/use cases
- list of collaborators into the project and their associated role into this project

If your role has been setted up as admin into a project, an action button on top right of each card will be available. By clicking on this button you will be able to edit the project information and delete the project. Please note that deleting a project will also delete all sub project’s items, such as pipelines or datasets, created on the project.



Note : Tips : you can filter projects by their names using the search bar on top right of the projects view

View list

In this view you will find the same information and actions than the card view at the exception of project description.

| NAME | CREATED BY | CREATED AT | USECASES | DATASETS | PIPELINES | COLLABORATORS |
|---------------------------------|----------------------|-------------------|----------|----------|-----------|-------------------------|
| Projet de Simon | Simon Levacher | 06/23/2021, 23:50 | 1 | 1 | 0 | SL FO |
| Axel's Project | Axel Chauvin QA test | 06/23/2021, 23:50 | 1 | 19 | 0 | AC NF SL LA FO |
| playground | gerome pistre | 06/23/2021, 18:25 | 3 | 6 | 0 | GP SL LA FO |
| Test end user | Simon Levacher | 06/23/2021, 15:42 | 1 | 1 | 0 | SL FP FO DF |
| TESTS 23/06 | Axel Chauvin QA test | 06/23/2021, 14:52 | 19 | 18 | 0 | AC DF NF FP JN LA FO SL |

rows per page: 10

prev 1 - 5 of 5 next

Project navigation

Into a project you can load data, create a pipeline in order to automate some task or data transformations and train models into use cases menu. Once you enter a project by clicking on its card or on the list, the project menu will be

loaded on the left navigation bar.

Axel's Project

Datasets

| NAME | CREATED AT | ROWS | COLUMNS | SIZE | DATA SOURCE | STATUS |
|--------------------------|-------------------|---------|---------|----------|-------------|------------------|
| regression_house_80_lite | 06/23/2021, 17:14 | 399 | 21 | 39.67 KB | | Ready to be used |
| french_tweets_lite | 06/23/2021, 15:21 | 404,769 | 2 | 39.42 MB | | Ready to be used |
| series-trafics | 06/23/2021, 15:20 | 99,999 | 3 | 3.29 MB | | Ready to be used |

See all Datasets

Pipelines

No data for table

See all Pipelines Create Pipeline

Usecases

| NAME | VERSION | CREATED AT | DATA TYPE | TRAINING TYPE | SCORE | MODELS | PREDICTIONS | STATUS |
|---------------------|---------|-------------------|-----------|----------------|-------|--------|-------------|------------------|
| classif_edf_80_lite | 1 | 06/24/2021, 10:28 | Tabular | Classification | - | 0 | 0 | Ready to be used |

See all Usecases Create Usecase

Aide

- Home : you will find here last items from datasets, pipelines and use cases created into the project.
- Usecases : you will find all your use cases trained into the selected project
- Data : you will find here your datasets and the connectors setted up on the project
- Pipelines : you will be able thanks to pipeline to augment your data and automate some actions
- Deployment : deploy and monitor deployed models and applications
- Collaborators : list of all project collaborators and their associated project role
- Settings : if your project role is admin, this menu is available and allows you to edit the project informations

collaboration into a project

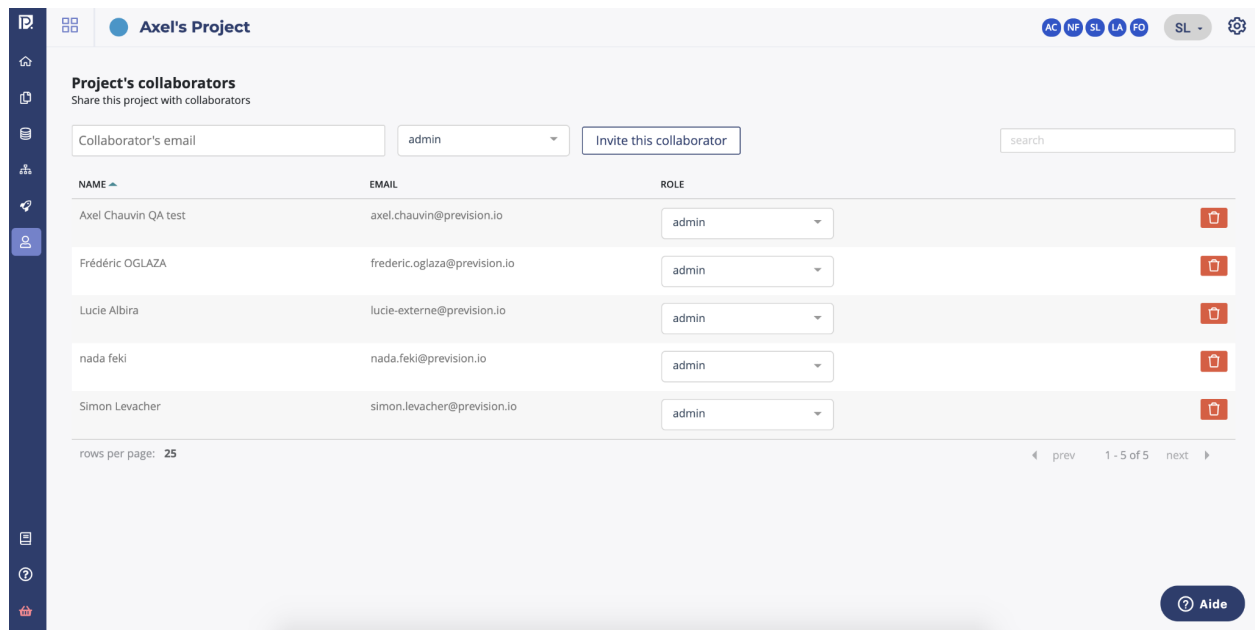
Prevision.IO studio is built in order for our users to collaborate within the projects. To do that, into the “collaborators” menu of a selected project you can manage, if your role is admin on the project, the collaborators and the right inside the project.

- Add a user : by enter the email address of a Prevision.IO platform registered user you can add a collaborator
- role : if your role level is admin into the project you will be able to edit user roles
- by clicking the delete button on the right side of a user, you can disable the access to the selected user to the project

Project roles

Into a project there are 3 levels of roles :

- End-user : in this project, the user can only access to the list of deployed models and applications and make predictions
- Viewer : you can navigate into all ressources of a project and visualize information but you can’t download or create ressources
- Contributor : Viewer rights + you can create and manage resources of the project
- Admin : Contributor rights + you can manage project users and project settings



edit a project

You can change the following parameters of your project by clicking on settings on the main navigation of your project or, on the list/card view of your project by clicking on the action button.

- Name of the project
- Description of the project
- Color displayed on the card of the project

delete a project

If a project is no longer useful, you can delete it by clicking on the action button on the card/list projects view.

Warning : all ressources created into the project will be deleted with the suppression of the project with no possibility of back-up. If deleted, a project and its resources are no longer available for you but also for all users previously added to this project.

Project home

By entering a project, you will first be redirected to the project homepage. The following sections, including the 3 latest entries for each section, are displayed :

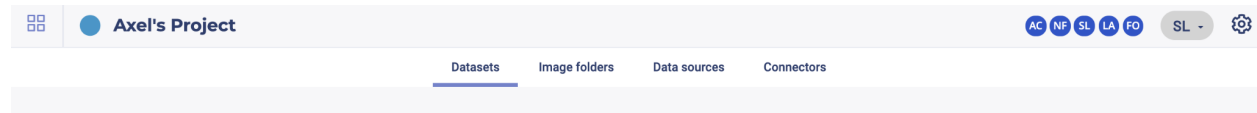
- Datasets : last uploaded dataset
- Pipelines : last pipeline templates
- Usecases : last usecases

Under each section you will also find a link to the dedicated page, also available through the left project main menu, and, for pipelines and usecase, a shortcut to create new ones.

2.2.2 Data

Navigation

In a project, by clicking on data on the main left navigation, you will enter the data page where all actions and files relative to data are localized. The following 2nd level navigation of the data interface will appear on top of the page.

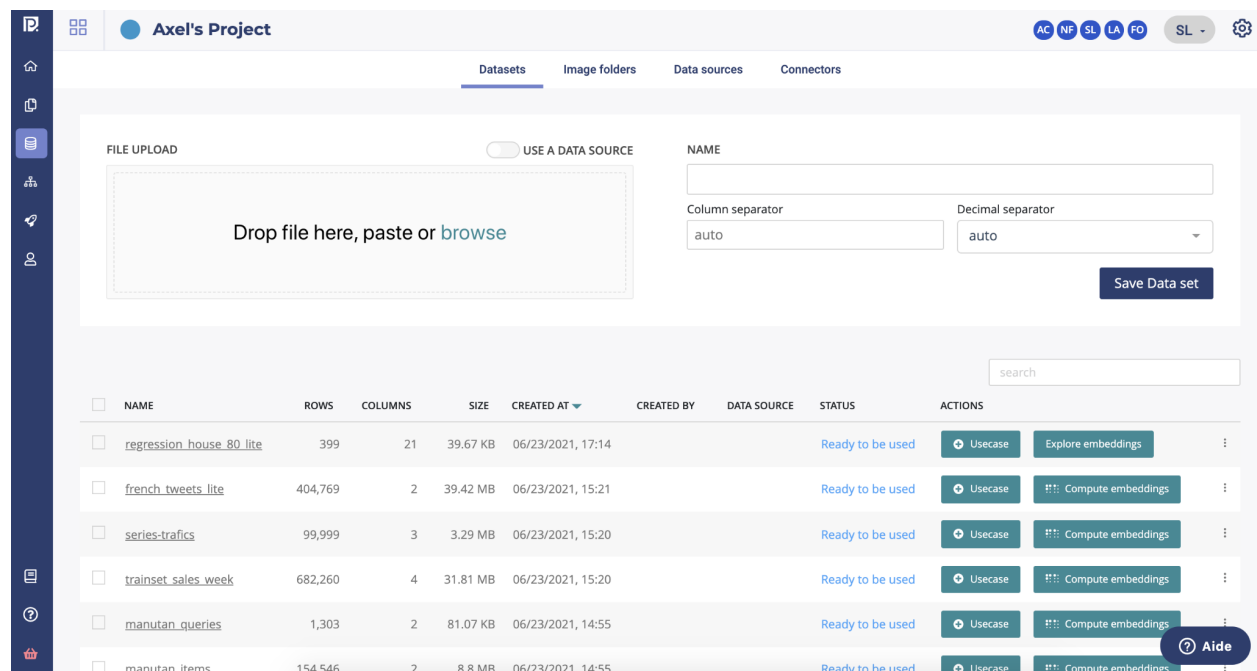


- Datasets : you can upload .zip or .csv files in this page
- Image folders : you can load .zip files containing images in order to do specific image use cases
- Data sources : by using connectors you can setup data sources in order to generate datasets from database or file server
- Connectors : you can manage connectors to your external databases or file servers in order to apply them on data sources

datasets

Upload a dataset

By clicking on the Datasets button of the dedicated data page menu, you will land on the dataset page.



This page allows you to consult all the project datasets uploaded into the application and import new ones by using one of the following methods :

- either from files (CSV or ZIP)
- either from a Data Source at a given time (snapshot)

In order to upload a dataset from a file, you can drag & drop in the dedicated area of your file or click on “Browse” in order to open your computer file explorer. Once your file is selected, you can start the upload by clicking on the “save Data set” button on the right side of the file upload area.

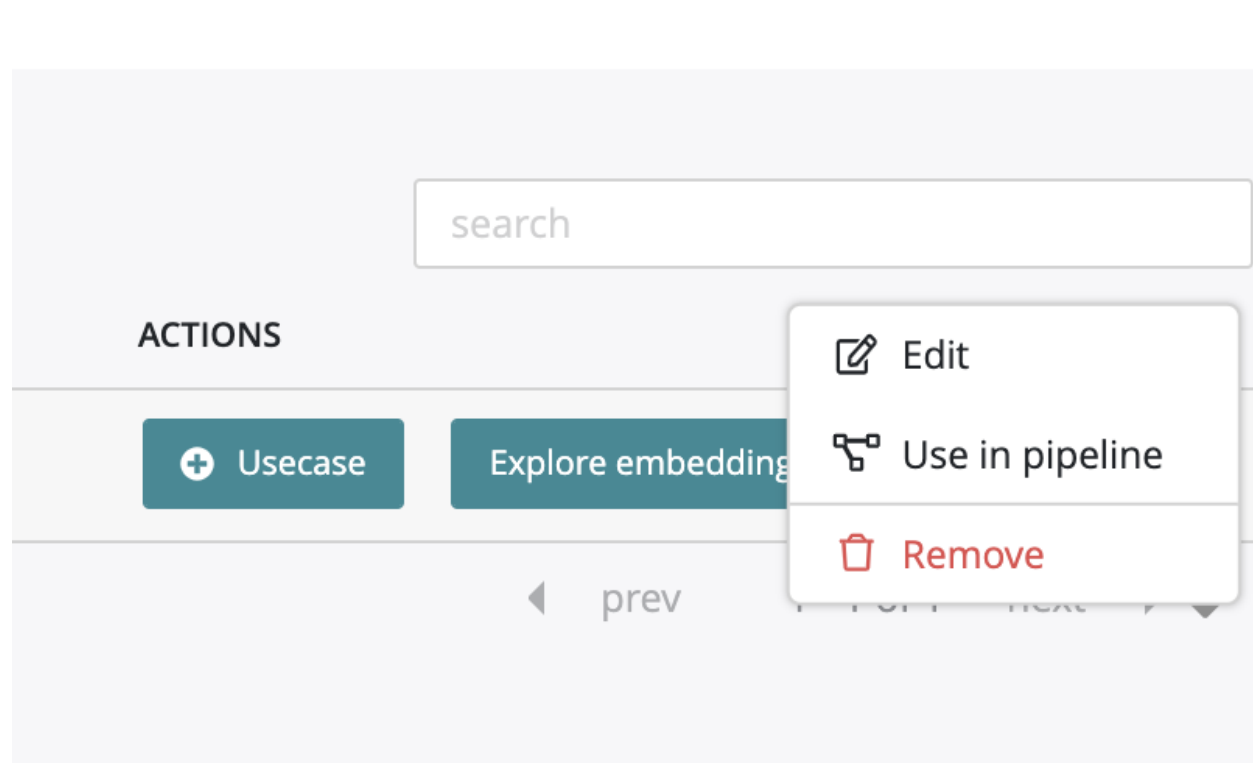
In order to create a dataset from a datasource, you have to use the toggle button “use a data source” and then select a data source from the dropdown list.

When the upload of the dataset is done, the platform will automatically compute information regarding your dataset in order to maximize the automation of machine learning computes. You can follow thanks to the status column on the list the progress of these operations.

dataset statistics pending : pre-computing of dataset information ready to be used : the dataset is ready to be used in the platform
 dataset statistic failed : the dataset can't be used in a train, you have to re upload the file
 drift pre-computing failed : you can train with this dataset but once a model deployed, the drift will be not available

From the dataset table, several actions are possible

Actions



By clicking on the usecase button, you can start the configuration of a training based on the selected dataset. By clicking on the start embedding button, you can launch the dataset analysis computing. Once it is done, the icon in the list will change for “explore embedding”. By clicking on it you will access to the dataset analysis dedicated page. By clicking on the action button on the right side of the table, you will be able to :

- edit the name of the dataset
- use this dataset into a pipeline
- delete the dataset

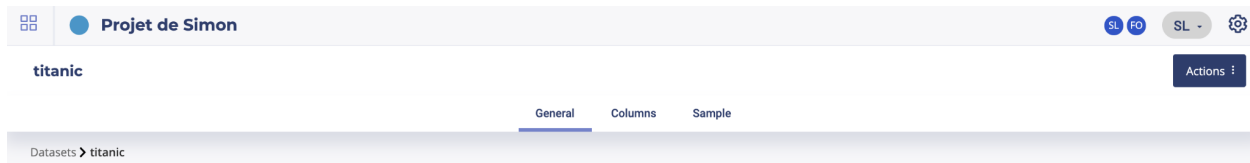
Please note that this action button will be also available into a dataset page.

Datasets informations

Navigation

Once your dataset is uploaded and computed into the platform, you will be able to access information about it by clicking on your dataset on the list.

Three menu regarding the dataset are available allowing you to understand better your data

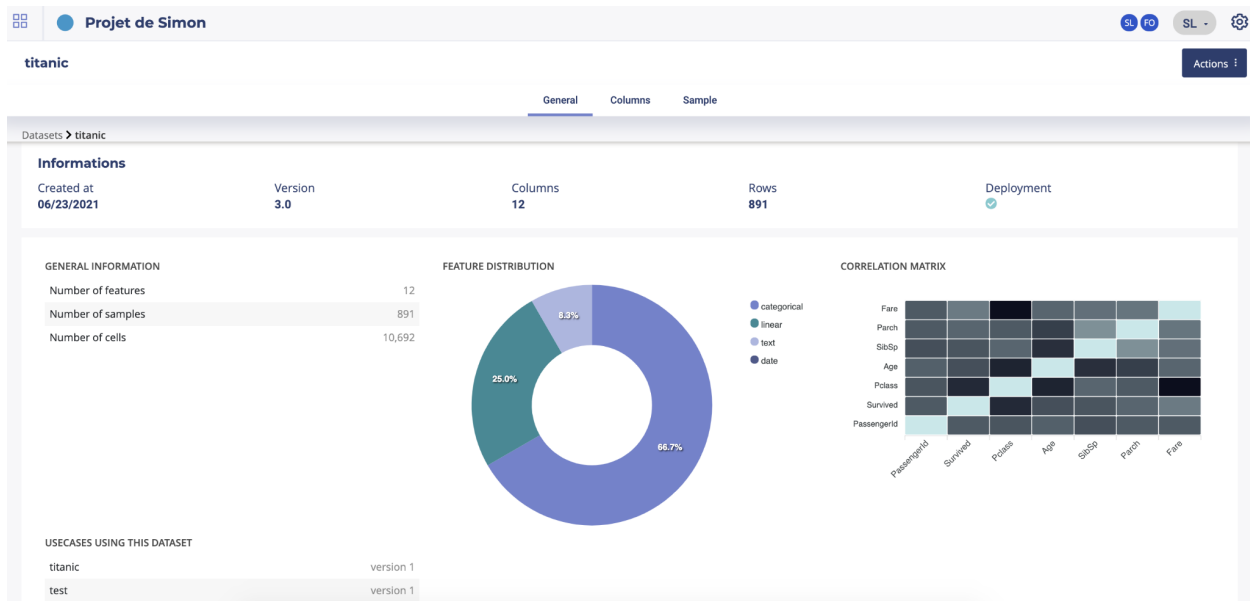


- General : general information about your dataset
- Columns : information about features of your dataset
- Sample : a sample visualisation of your dataset

General

On the general screen of a dataset you will find generic information about your dataset such as the number of columns, number of samples and number of cells or the usecases using this dataset. Two graph are also displayed showing : * the feature distribution regarding the feature type (linear, categorical, date or text). This distribution is automatically calculated when uploading a dataset into the platform * correlation matrix showing the correlation coefficients between variables. Each cell in the table shows the correlation between two variables

You will also have on the bottom of the screen the list of usecases trained from this dataset.



Columns

By clicking on the column button on the top menu you will find a listing of all the dataset features, their role (categorical, linear, text or date) and the percentage of missing value for each feature

Sample

By clicking on the sample button, a sample of 10 rows of your dataset will be displayed.

Dataset analysis

Introduction

The Data Explorer is a specific module that aims to detect similarities between samples of your dataset. It uses a combination of Dimension reduction algorithms for representing your dataset into a vector space, sometimes called embedding By using it, you're being able to :

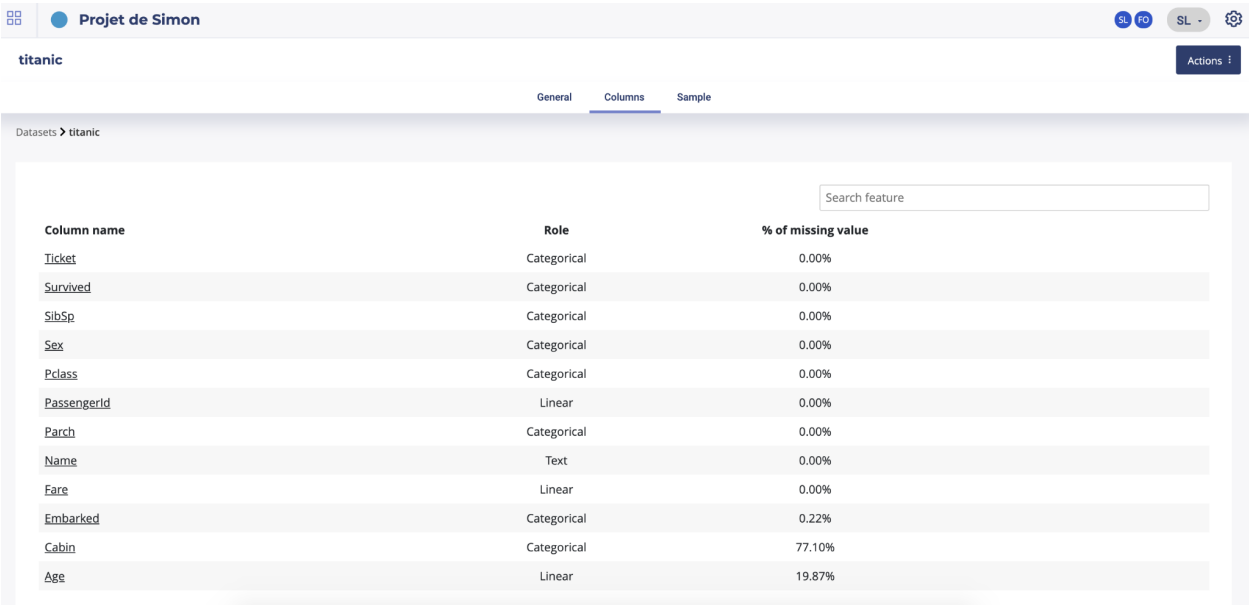
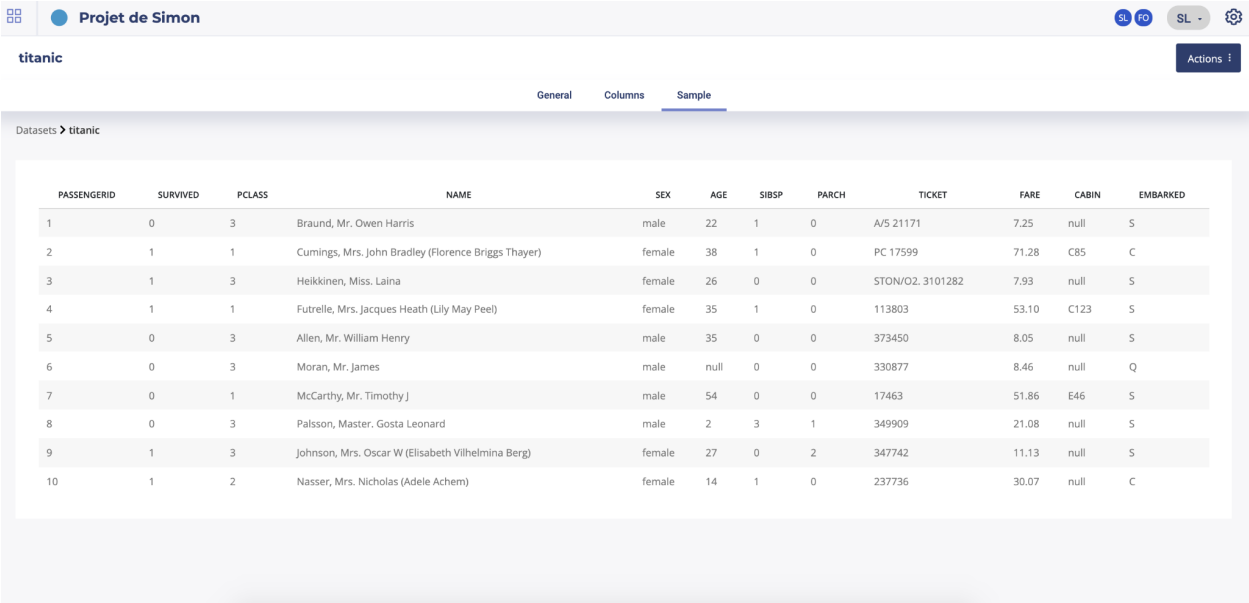


Fig. 1 – image alt text

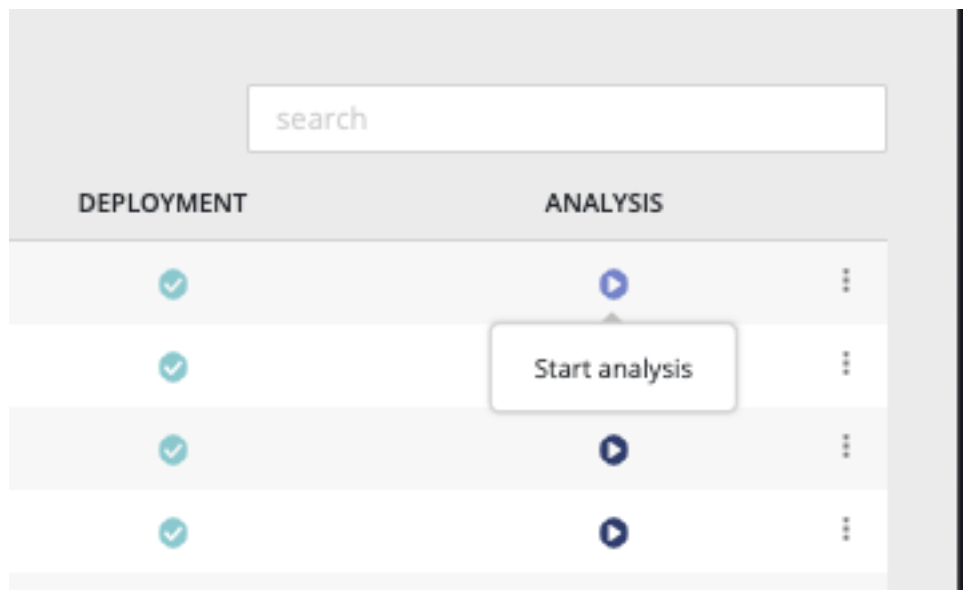
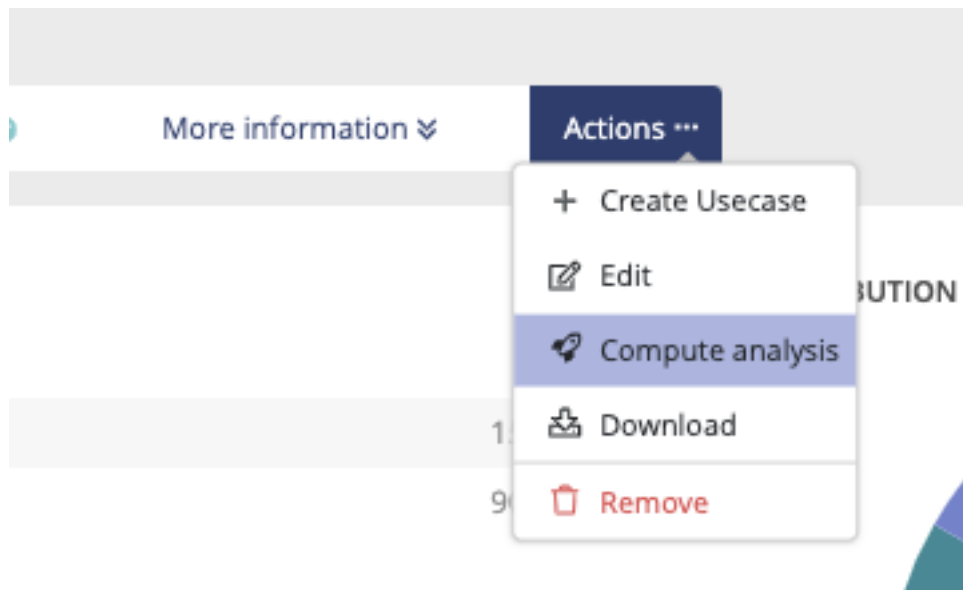


Visually observe cluster see which samples are the most similar to a selected one, for example a Customer in his buying habits See in which population a given feature, like expenses, is present or higher Have a global view of your data

The Data explorer is often used as a pre-analysis of datasets, as it uses an unsupervised algorithm, but it can be used as a standalone feature. Once the embedding has been generated you can request them by API or download them for use in a third party tool like Excel.

Start a dataset analysis

Two possibilities in order to launch the explorer. By clicking on the “start embedding” on the dataset list or, after opening a dataset page, by clicking the actions button on top of the screen and clicking on “start embedding” .



Compute embedding will take more or less time regarding the size of your dataset. Once the computing is done, you will see on the list an eye icon and, on the dataset page on the actions button, the compute analysis button will

be replaced by an “explore embedding” button. By clicking on one of these buttons, you will enter into the dataset analysis interface.

The explorer

The Data Explorer is now accessible and will give you a visual representation in 2 or 3 dimensions of the selected Data Set. This representation is a dimension reduction constrained to 2 or 3 dimensions, applied on the embedded vectors, that may be of a higher dimension. There are five important sections in the data explorer.

— Graphical projection

The main screen is a visual representation of the dataset. Each point is a sample of your dataset (up to 5000). You can pan and zoom and if you click on a point, or use the selecting box tool, some more info is displayed.

In this representation, points are grouped by similarities as much as possible, meaning that if two points are near in this space, the samples share some important similarities.

The nature of the displayed information are selected on the section (3)

— Search and similarities

The second section is a dynamic list of similar sample.

You can search any sample from any feature. For example if your dataset has an index with name, you can search a sample by using its name but you can too search all the sample that have « RPG » as type or « 5 » for size.

Once a sample is selected, it and a list of similar are highlighted in the main section. They can be further isolated by clicking on the « isolate N points » button on top of the section.

The number of similar samples to display can be choosen with the « neighbors » slider

— Labels

Section 3 main purpose is to set labels displayed in section 1. Here you can set :

the label displayed above each point

the feature use for coloring each point :

— Segmentation and clustering

Section 4 is all about Segmentation and clustering your samples.

Here you can choose an algorithm and tune its parameter to display the most similar point together. Thus, you can start to observe sample clusters, or segments of data that represent big groups that share important similarities.

Yet, as we try to project a lot of dimensions in a smaller space (3D or 2D), note that this algorithm is just for displaying and shaping human decision. A lot of the process is a little bit subjective and further conclusion should be driven by a supervised algorithm.

Anyway, here you can choose between 3 algorithms :

- PCA : the quickest and simplest algorithm. Clicking on the PCA tab immediately led to a 3D representation of your samples. Yet, this is a very simple algorithm that only shows sample variability along 3 axes. You can find more information about [PCA on Wikipedia](#)
- t-SNE : once you click on the t-SNE tab, a process of convergence is launched. t-SNE is a very time consuming algorithm but that can lead to very accurate segmentation. You can change its parameters and click on the « Stop » button then « Re-run » it. But in most cases it's better to already know this algorithm to use it. You can find more information about [t-SNE on Wikipedia](#)
- UMAP : UMAP is a good alternative to t-SNE and PCA. Quicker than t-SNE , it offers better results than PCA. The only parameters is « Neighbors », that change the size of clusters. The more neighbors you ask for, the bigger the cluster. You can find more information about [UMAP on Wikipedia](#).

We recommend using UMAP in most cases.

— API informations

Search by

neighbors 10

distance ☒ COSINE ☐ EUCLIDEAN

Nearest points in the original space:

| | |
|-----------------------|-------|
| Aligre | 0.002 |
| Ile Saint Louis | 0.003 |
| porte de clignancourt | 0.004 |
| Austerlitz | 0.005 |
| Port Royal | 0.007 |
| max Dormoy | 0.009 |
| XV | 0.009 |
| Nation | 0.014 |
| Saint Sulpice | 0.014 |
| nation | 0.018 |

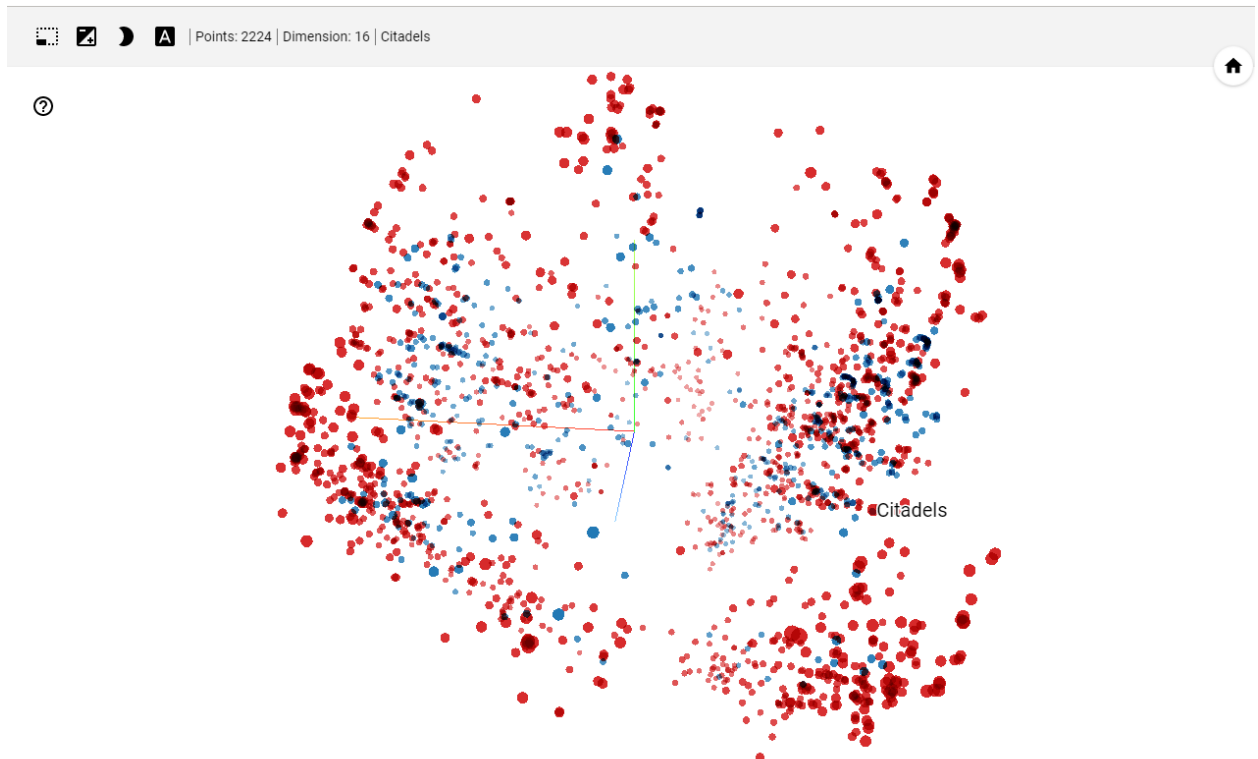
Fig. 2 – image alt text

neighbors 28

Fig. 3 – image alt text

1 tensor found
steam_train ▼

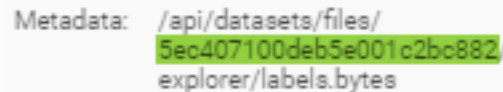
Label by ID ▼ Color by GenrelsStrategy ▼



The 5th part is only about API information.

When launching a dataset Analysis, the platform builds an embedding of the dataset, namely, it projects each sample of the dataset to a vector. This embedding is attached to the dataset and can be retrieved with the dataset ID. Then you can use it for running any mathematical operation, in most cases a distance, that can be run on vectors.

Section 5 of the tools gives you the Id of your dataset :



```
Metadata: /api/datasets/files/  
5ec407100deb5e001c2bc882/  
explorer/labels.bytes
```

Fig. 4 – image alt text

With it you can access several URL :

- GET <https://.prevision.io/api/datasets/files/download> : get the original dataset
- GET <https://.prevision.io/api/datasets/files/> : JSON info about your dataset
- GET <https://.prevision.io/api/datasets/files/explorer> : JSON info about the embeddding
- GET <https://.prevision.io/api/datasets/files/explorer/tensors.bytes> : numpy files of embeddings
- GET <https://.prevision.io/api/datasets/files/explorer/labels.bytes> : tsv files of labels

The embedding files (tensor.bytes) is a numpy float 32 file whom shape is in the json file if explorer URL. You can read it with the following python code for example

```
req = Request('https://<YOUR_DOMAIN>.prevision.io/ext/v1/datasets/files/<DATASET_ID>/  
→explorer/tensors.bytes')  
req.add_header('Authorization', <YOUR_TOKEN> ) #get YOUR_TOKEN in the admin page  
content = urlopen(req).read()  
vec = np.frombuffer(BytesIO(content).read(), dtype="float32").reshape(u,v) # u,v is_  
→the shape gotten in /ext/v1/datasets/files/<DATASET_ID>/explorer  
print(vec.shape)
```

Please note that you can use SDK's functions in order to simplify this process.

image folders

In order to train image use cases you will have to upload images using a zip file. By drag & drop a zip file in the dedicated area you will be able to load your image folder. All images folders uploaded into your project will appear in the list under the drag & drop section.

It is recommended to use an images dataset whose total volume does not exceed 4 GB. We invite you to contact us if you want to use larger datasets.

By clicking the action button on the list you will be able to :

- edit the name of your image folder
- delete your image folder

Connectors

In the Prevision.IO platform you can set connectors in order to connect the application directly to your data spaces and generate datasets. Several connector types are available :

- SQL databases
- HIVE databases
- FTP server
- Amazon S3 datastore

TESTS 23/06

AC GP NP PP JN LA FO SL ... SL -

Datasets Image folders Data sources Connectors

IMAGES ZIP

Drop file here, paste or [browse](#)

NAME

Upload image folder

| NAME | SIZE | STATUS | CREATED AT |
|--------------------|-----------|--------|-------------------|
| aquarium_valid | 13.45 MB | OK | 06/25/2021, 16:12 |
| aquarium_train | 46.91 MB | OK | 06/25/2021, 16:12 |
| aquarium_test | 6.01 MB | OK | 06/25/2021, 16:12 |
| Mask_Wearing_valid | 748.95 KB | OK | 06/25/2021, 14:59 |
| Mask_Wearing_train | 2.51 MB | OK | 06/25/2021, 14:59 |
| Mask_Wearing_test | 325.88 KB | OK | 06/25/2021, 14:59 |
| fourmes_test | 1.12 MB | OK | 06/24/2021, 15:14 |
| fourmes_train | 30.02 MB | OK | 06/24/2021, 15:14 |

rows per page: 25

search

prev 1 - 8 of 8 next

— GCP

By clicking on the “new connector” button, you will be able to create and configure a new connector. You will need to provide information depending on connector’s type in order for the platform to be able to connect to your database/file server.

Note : TIPS : you can test your connector when configured by clicking the “test connector” button.

Once connectors are added, you will find under the new connector configuration area the list of all your connectors. You can, by clicking on the action button :

- test the connector
- edit the connector
- delete the connector

Once at least one connector is well configured, you will be able to use the data sources menu in order to create CSV from your database or file server.

Data sources

In order to create datasets, you first need to configure a data source using connector information. To do this, click on the data sources menu and select the configured connector in the dropdown list. Depending on the connector type, you will have to configure the data source differently.

When your data source is ready, in order to generate a dataset from it, you have to go to the datasets page, enable the toggle button “use a data source” and select your data source from the dropdown list.

SQL data sources

Once a SQL connector is selected, you will have to choose first the database you want to use. Then, two different methods are available in order to configure your data source :

- by selecting a table
- by clicking the “select by query” button and entering a valid SQL query

HIVE data sources

FTP data sources

SFTP data sources

Amazon S3 datastore

GCP data sources

2.2.3 Use cases

Introduction

Once in a project, you can go to the “use case” page using lateral navigation and start creating new use cases or explore already existing ones.

Regarding the problematic and the data type you have, several training possibilities are available in the platform :

Tableau 1 – Type of Training

| Training type / Data type | Tabular | Time-series | Image | Definition | Exemple |
|---------------------------|---------|-------------|-------|---|---|
| Regression | Yes | Yes | Yes | Prediction of a quantitative feature | 2.39 / 3.98 / 18.39 |
| Classification | Yes | No | Yes | Prediction of a binary quantitative feature | « Yes » / « No » |
| Multi Classification | Yes | No | Yes | Prediction of a qualitative feature whose cardinality is > 2 | « Victory » / « Defeat » / « Tie game » |
| Object Detection | No | No | Yes | Detection from 1 to n objects per image + location | Is there a car in this image ? If so, where ? |
| Text Similarity | Yes | No | No | Estimate the similarity degree between two text. Find texts that are similar in context and meaning with your queries | « a tool for screws » should lead to a a screw-driver description |

Then, for each data type, you will have to choose between several usecase types demanding a specific configuration for each.

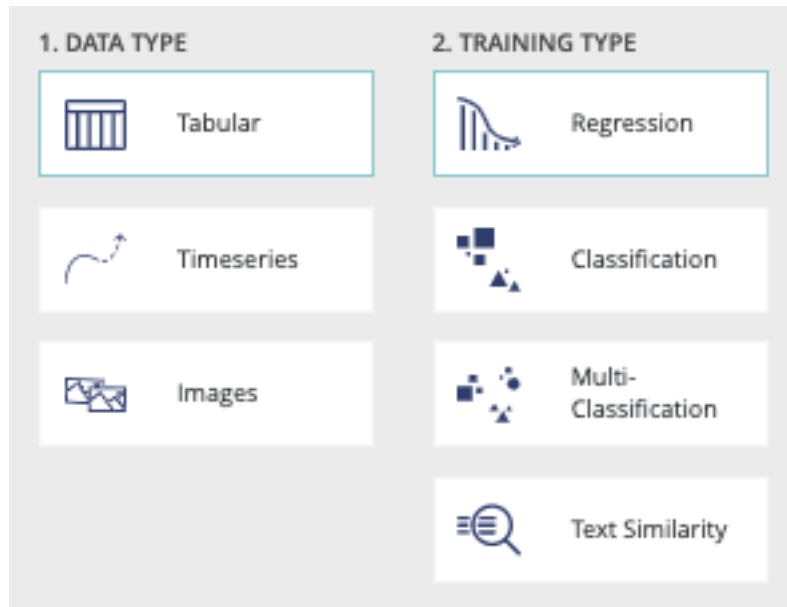
Create a new usecase

In order to create a new usecase using the interface, three possibilities are available :

- In the usecase menu by clicking on the “new usecase” button top right of the screen
- By clicking the actions button of the dataset list and clicking on the “create usecase” button
- On a dataset page by clicking on the “actions” button and select “create usecase” on the menu

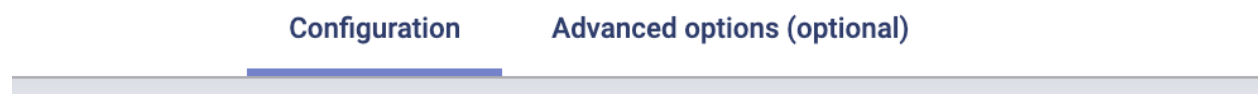
Then you will land on the new usecase page and will have to choose the datatype and the training type regarding your problem.

As training types requires specific configuration, all information needed to start the training of a usecase will be explain on each training type dedicated chapters



Usecase configuration navigation

Once the training options are selected, by clicking on the “next step” button you will be able to navigate and parameter your usecase. The following image shows the navigation menu regarding the use case parameters.



- Configuration : Includes all required parameters such as the file target
- Advanced options : Includes all advanced options such as model selection and feature engineering

Due to specific choices, options will be detailed on specific training sections.

Data Type : Tabular

Introduction

Following tables show you all type of training options for a tabular data type usecases :

Tableau 2 – Type of Training. Tabular Data

| Algorithm | Training : simple | Training : Normal | Training : Advanced | Regression | Classification | Multi-classification | Blend |
|-------------------|-------------------|-------------------|---------------------|------------|----------------|----------------------|-------|
| Logistic Model | Yes | No | No | No | Yes | Yes | Yes |
| Linear Regression | Yes | No | No | Yes | No | No | Yes |
| Decision Tree | Yes | No | No | Yes | Yes | Yes | No |
| XG Boost | No | Yes | Yes | Yes | Yes | Yes | Yes |
| Linear Model | No | Yes | Yes | Yes | Yes | Yes | Yes |
| Extra Trees | No | Yes | Yes | Yes | Yes | Yes | Yes |
| Random Forest | No | Yes | Yes | Yes | Yes | Yes | Yes |
| LightGBM | No | Yes | Yes | Yes | Yes | Yes | Yes |
| Neural Network | No | Yes | Yes | Yes | Yes | Yes | Yes |
| CAT Boost | No | Yes | Yes | Yes | Yes | Yes | Yes |
| Naive Bayes | No | Yes | No | No | Yes | Yes | No |

Tableau 3 – Text Embedding

| Models | Tf-Idf | Transformers | Transformers Fine-tuned |
|-----------------|--------|--------------|-------------------------|
| Brute Force | Yes | Yes | Yes |
| Cluster Pruning | Yes | No | No |
| IVF OPQ | No | Yes | Yes |
| LSH | No | Yes | Yes |
| HKM | No | Yes | Yes |

Tabular : usecase information

After selection of data type tabular and training type regression, you will see the following form displayed on your screen.

3. NAME & DATASET SELECTION**Next step**

Usecase name

set usecase name

Usecase description (optional)

set usecase description

Dataset ?

select a Dataset

Holdout (optional)

select a Dataset

- Usecase name (mandatory) : usecase name displayed on usecase list
- Usecase description (optional) : helpful to describe differences between versions of a usecase
- Dataset (mandatory) : dropdown list of previously added dataset into the project available as training dataset

- Holdout dataset (optional) : dropdown list of previously added dataset into the project usable as holdout. Holdout dataset is used at the end of the training in order to generate predictions using created models and calculate a holdout score. If no holdout, the calculated score will be down by creating synthetic holdout from the training dataset

Once mandatory information has been fulfilled, the next step button is enabled and allows you to continue the configuration of your usecase.

Tabular : use case configuration

The screenshot displays the 'Configuration' tab in the Prevision.io interface. At the top, it shows 'TESTS 23/06' and a 'New usecase > Step 2' button. The main configuration area is divided into two sections: 'Fields configuration' and 'Columns configuration'. The 'Fields configuration' section includes dropdowns for 'Target column', 'ID column', 'Weight', and 'Fold'. The 'Columns configuration' section has a search bar labeled 'search features'. To the right, the 'Directed Acyclic Graph (DAG) Preview' shows a flowchart starting from 'Start' to 'Dataset', then branching into 'Dataset statistics', 'Features transformations (simple)', 'Features transformations (big)', and 'Features transformations'. These lead to various models: 'SIMPLE-OT', 'SIMPLE-LR', 'XGB-1', 'LR-1', 'Hyper parameters optimization LR', 'LR-2', and 'XGB-2'. A 'Create and train' button is visible in the top right corner.

The first step you need to fulfil in order to start the training of your usecase is information regarding the dataset configuration. Two panels are displayed on the left side of your screen. The top one, called “fields configuration” allows you to select :

- Target (mandatory) : the columns containing the truth you want later predict
- ID column (optional) : if your dataset contains one column referencing IDs not interesting for the training, you can automatically, by selecting it, ignore it during training.
- Weight (optional) : sometimes, a numerical does not contain an actual feature but rather an indication of how important each row is — if that is the case, you can pass the name of this column as weight_column (the higher the weight, the more important the row — by default, all rows are considered to be of equal importance); note that if this is provided, the optimised metric will become weighted
- Fold (optional) : if no “fold” column is provided, a random stratification will be used and will try to force the same distribution of the target between folds. If you want to perform a custom stratification to improve the quality of your validation (which can be misleading if some kind of group structure appears in your data which is not reproduced in the train / test split), you can pass a specific column name to use as reference. The stratification will then ensure that no two samples that have the same “fold id” in the “fold” column will be located in different folds.

The second bottom left panel allows you to ignore some columns of your dataset. To do it, just deselect unwanted features. Please note that you can search by features name the columns you want to unselect and use the “select/unselect all” checkbox to apply your choice to the selection.

The right panel is showing you a preview of the AutoML pipeline that will be performed by Prevision.IO in order to train your usecase.

Tabular : advanced options

By clicking on the advanced options button, you will be able to configure more in detail your training.

Advanced options : Training options

The screenshot shows a web interface titled "Training options". Below the title, there is a section labeled "Metric to use" with a dropdown menu currently displaying "Root mean squared error". Below this, there is a section labeled "Performances" with three radio button options: "QUICK" (which is selected and has a blue dot), "NORMAL", and "ADVANCED". Each option has a small question mark icon next to it.

First panel of your advanced option is the training options one allowing you to select the metric to use, depending on your training type, and the training performances you want. In order to know all metrics supported by prevision.IO please refer to the dedicated area of this documentation.

Three type of performances are

- QUICK : Training is done faster but performance may be slightly lower. Ideal in iterative phase.
- NORMAL : Intermediate value, suitable for most usecases on a later stage.
- ADVANCED : The training is done in an optimal way. Though the performance will be more stable, the calculations will take longer to process. This is ideal when the model is put into production and the performance is discriminating.

Advanced options : Feature engineering

Four kinds of feature engineering are supported by the platform. :

- Date features : dates are detected and operations such as information extraction (day, month, year, day of the week, etc.) and differences (if at least 2 dates are present) are automatically performed
- Textual features : * Statistical analysis using Term frequency–inverse document frequency (TF-IDF). Words are mapped to numerics generated using tf-idf metric. The platform has integrated fast algorithms making it possible to keep all uni-grams and bi-grams tf-idf encoding without having to apply dimension reducing. More information about TF-IDF on <https://en.wikipedia.org/wiki/Tf%E2%80%93idf> * Word embedding approach using Word2Vec/Glove. Words are projected a dense vector space, where semantic distance between words are : Prevision trains a word2vec algorithm on the actual input corpus, to generate their corresponding vectors. More information about Word embedding on https://en.wikipedia.org/wiki/Word_embedding * Sentence Embedding using Transformers approach. Prevision has integrated BERT-based transformers, as a pre-trained contextual model, that captures words relationships in a bidirectional way. BERT transformer makes it possible to generate more efficient vectors than word Embedding algorithms, it has a linguistic “representation” of its own. To make a text classification, we can use these vector representations as input to basic classifiers to make text classification. Bert (base/uncased) is used on english text and Multi Lingual (base/cased) is used on french text. More information about Transformers on [https://en.wikipedia.org/wiki/Transformer_\(machine_learning_model\)](https://en.wikipedia.org/wiki/Transformer_(machine_learning_model)). The Python Package used is Sentence Transformers (https://www.sbert.net/docs/pretrained_models.html)
- Categorical features : * Frequency encoding : modalities are converted to their respective frequencies * Target encoding : modalities are replaced by the average (TARGET, grouped by modality) for a regression and by the proportion of the modality for the target’s modalities in the context of a classification
- Advanced features : * Polynomial features : features based on products of existing features are created. This can greatly help linear models since they do not naturally take interactions into account but are less usefull on tree based models * PCA : main components of the PCA * K-means : Cluster number coming from a K-means methode are added as new features * Row statistics : features based on row by row counts are added as new features (number of 0, number of missing values, ...)

Feature engineering

☐ Date features ?

☒ Textual features ?

- ☐ Statistical analysis (TF-IDF)
- ☐ Word embedding (Word2Vec)
- ☒ Sentence embedding (Transformers)

☒ Categorical features ?

- ☒ Frequency encoding
- ☒ Target encoding

☒ Advanced features ?

- ☐ Polynomial features
- ☐ PCA
 - ☐ K-means
- ☒ Row statistics

Please note that if you don't have a feature of one of these feature types in your train dataset, the corresponding feature engineering toggle button will be disabled. Also please note that textual features pretreatments only concern advanced models and normal Naive Bayes model

Advanced options : Model selection

The model selection area allows you to select the type of model you want to train. In order to know precisely which models you can train for each training type, please refer to the model matrix at the beginning of the tabular datatype chapter.

Model selection

Simple models:

☒ Logistic model
 ☒ Decision Tree

| | NORMAL [?] | ADVANCED [?] |
|--|-------------------------------------|-------------------------------------|
| <input checked="" type="checkbox"/> XGBoost | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> |
| <input checked="" type="checkbox"/> Logistic model | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> |
| <input type="checkbox"/> Extra Trees | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> Random Forest | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> LightGBM | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> Neural Network | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> CATBoost | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> Naive Bayes | <input type="checkbox"/> | <input type="checkbox"/> |

☐ Blend

By enabling the “blend” toggle, the platform will create blended models based on the model selection you choose.

Advanced options : Feature selection

In this part of the screen you can choose to enable feature selection (off by default).

☒ Features Selection

features count

percentage ☒ unit

50%
10/ 19 selected features

maximum duration

60 minutes

This operation is important when you have a high number of features (a couple hundreds) and can be critical when the number of features is above 1000 since the full Data Set won't be able to hold in RAM.

You can choose to keep a percentage or a count of feature and you can give a time budget to Prevision.io's to perform the search of optimal features given the TARGET and all other parameters. In this time, Prevision.io will subset the feature of the Data Set then start the classical process.

The variable selection strategy in Prevision.io is hybrid, depends on the characteristics of the dataset and the time available.

1. It is hybrid because it combines both so-called filtering methods, encapsulation methods and integrated methods. The filtering methods perform the selection of entities independently of the construction of the classification model. Encapsulation methods iteratively select or eliminate a set of entities using the metric of the classification / regression model. In built-in methods, feature selection is an integral part of the classification / regression model.
2. It depends on the characteristics of the dataset and the time allotted. In fact, depending on the volume of the dataset, a small data strategy is applied for a dataset of less than 8 GB, fully in memory. Otherwise, a big data strategy is applied.
3. In a small data situation, a first filtering approach is carried out consisting in filtering the variables of zero variance, the duplicated variables, the intercorrelated variables beyond 99% and the variables correlated to the target variable beyond 99% . Depending on the time remaining available, a second so-called encapsulation method is carried out using a LASSO-type regularization on the entire dataset by cross validation with the aim of optimizing the metric selected when the use case is launched.
4. In a big data situation, as time permits, several row and column samplings are carried out and the stages of filtering, encapsulation method and integrated methods completed by a reinforcement learning strategy are successively launched. . The variables are then ranked in order of priority according to the different approaches tested and the top variables, at the threshold defined by the user, are sent to the various algorithms of Prevision.io.

Tabular text similarity introduction

Even if considered as a training type for tabular data type, text similarity usecases are particular and need specific training options.

First, instead of a holdout dataset, a queries dataset (optional) can be selected.

Tabular text similarity dataset configuration

The second step of a text similarity training will allow you to configure the fields for the training dataset and the queries dataset (if added on the previous step).

- Description dataset configuration :
 - description column : textual description column of the dataset
 - ID column : item ID
- Queries dataset configuration :

3. NAME & DATASET SELECTION

Next step

Usecase name

set usecase name

Usecase description (optional)

set usecase description

Dataset 

select a Dataset

Queries (optional)

select a Dataset

Fields configuration

DESCRIPTION DATASET CONFIGURATION

Description column

ID column

QUERIES DATASET CONFIGURATION

Query column

Matching ID column in the description dataset

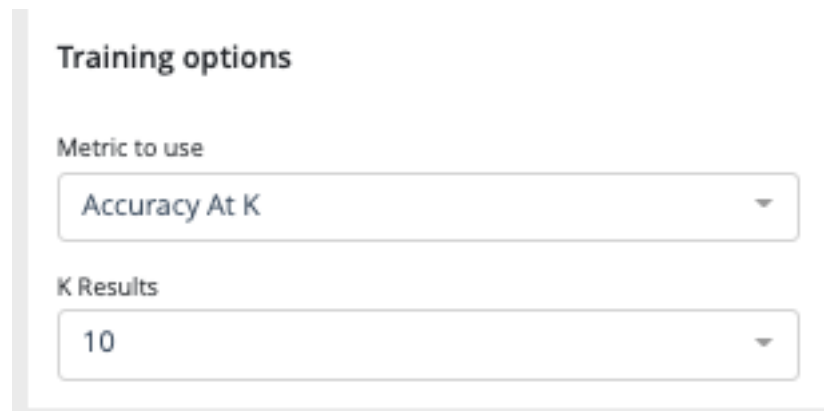
ID column (optional)

- Query column : column containing the queries
- Matching ID column : column allowing the platform to match a query from the queries dataset to a description from the description dataset
- ID column (optional) : ID of the queries

Please note that only these 3 columns from your queries dataset will be considered during the training.

Tabular text similarity advanced option

text similarity training options



The image shows a 'Training options' panel with two dropdown menus. The first dropdown, labeled 'Metric to use', has 'Accuracy At K' selected. The second dropdown, labeled 'K Results', has '10' selected.

Two training options dropdown have to be configured before training :

- the metric to use in order to compute the models scores
 - **Accuracy at k** : Is the real item corresponding to a query present in the search result, among the k items returned? The value is a percentage calculated on a set of queries.
 - **Mean Reciprocal Rank (MRR) at k** : Similar to accuracy at k. However the score for each query is divided by the rank of appearance of the corresponding item. Example : If for a query the corresponding item appears in third position in the returned list, then the score will be $\frac{1}{3}$. If it appears in second position the score will be $\frac{1}{2}$, in first position the score will be 1, etc. https://en.wikipedia.org/wiki/Mean_reciprocal_rank
- K results : the number of query like items that the tool must return during a search. Value between 1 and 100.

Text similarity Preprocessing

Several preprocessing options are available :

- Language : you can force the training dataset language to english or french or, let the platform determines by itself between these two languages
- TF-IDF preprocessing :
 - Stop words treatment : you can choose if the platform has to ignore or consider the stopwords during the training. As for the language, you can also let the system makes it own decision by selecting “automatic”
 - Word stemming : **stemming** is the process of reducing inflected (or sometimes derived) words to their **word stem**, base or **root** form—generally a written word form.
 - Ignore punctuation : by activating this option, the punctuation will not be considered during the training

Text similarity model selection

Text similarity module is composed of 2 kinds of models :

- embedding model to make a vector representation of your data
- search models to find proximity between your queries and product database

Embedding model / word vectorization : (https://fr.wikipedia.org/wiki/Word_embedding)

Term Frequency - Inverse Document Frequency (TF-IDF) : Model representing a text only according to the occurrence of words. Words not very present in the corpus of texts will have a greater impact. <https://fr.wikipedia.org/wiki/TF-IDF>

Transform : Model representing a text according to the meaning of words. In particular, the same word will have a different representation according to the other words surrounding it.

Preprocessing

Language

Automatic

☒ TF-IDF Preprocessing

Stop words treatment

☐ Ignore

☐ Consider

☒ Automatic

☒ Word stemming

☐ Ignore punctuation

Model selection ?

| | TF-IDF | TRANSFORMERS | TRANSFORMERS FINE-TUNED |
|-----------------|-------------------------------------|-------------------------------------|-------------------------------------|
| Brute force | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> |
| Cluster pruning | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| IVF OPQ | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| LSH | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| HKM | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |

[https://en.wikipedia.org/wiki/Transformer_\(machine_learning_model\)](https://en.wikipedia.org/wiki/Transformer_(machine_learning_model))

Transformer feature-based : Transformer that has been trained upstream on a large volume of data, but has not been re-trained on the corpus in question.

Fine-tuned transform : A transform that has been trained on a large volume of data and then re-trained on the text corpus in question.

Search models :

Brute Force : Exhaustive search, i.e. each query is compared to the set of item descriptions.

Locality sensitive hashing (LSH) : exhaustive search. Vectors are compressed to speed up distance calculations.

https://fr.wikipedia.org/wiki/Locality_sensitive_hashing

Cluster Pruning : non-exhaustive research. Item descriptions are grouped by cluster according to their similarity. Each query is compared only to the queries of the closest group.

<https://nlp.stanford.edu/IR-book/html/htmledition/cluster-pruning-1.html>

Hierarchical k-means (HKM) : non-exhaustive research. The idea is the same as for the previous model, but the method used to group the items is different.

Inverted File and Optimized Product Quantization (IVFOPQ) : non-exhaustive search. The idea is the same as for the two previous models, but the method used to group the items is different. Vectors are also compressed to speed up distance calculations.

As you can see in the DAG, in order to train a text similarity model, text embedders have to be done but some of them are not compatible with models. Here is all the combination text embedder/model you will be able to perform in the platform :

Please note that in order to guarantee the performance of IVF-OPQ models, a minimum of 1000 unique IDs in the train dataset is required.

Datatype time series

Introduction

In the prevision.io platform you have the possibility to train time series usecase in order to do forecasting predictions. By selecting in the new usecase screen the timeseries data type you will access the timeseries usecase configuration.

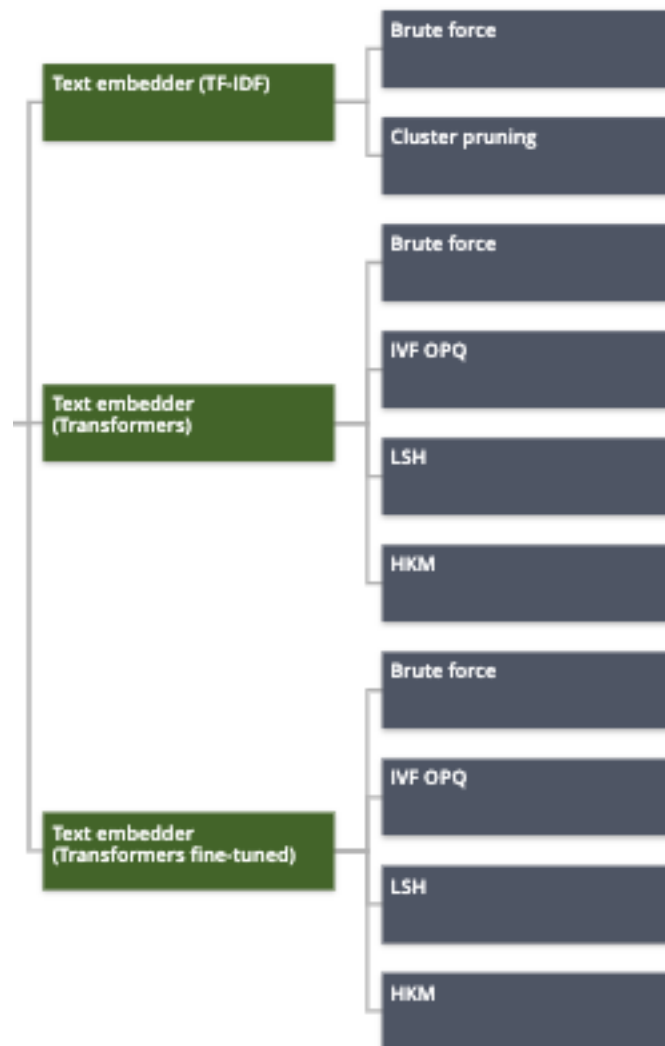
Timeserie usecase configuration

Time series is very similar to tabular usecase except :


- There is no hold out
- There is no weight
- There is no fold (in this case, Prevision.io use temporal stratification)


However, you will find some new notions :


- Temporal column : the feature that contains the time reference of the time series. Since date formats can be complex, Prevision.io supports ISO 8601 (https://fr.wikipedia.org/wiki/ISO_8601) as well as standard formats (e.g. DD/MM/YYYY or DD-MM-YYYY hh:mm).
- Time step : period between 2 events (within the same group) from the temporal column (automatically detected)
- Observation window : illustrate the period in the past that you have for each prediction * Start of observation window : the maximum time step multiple in the past that you'll have data from for each prediction (inclusive, 30 by default) * End of the observation window : the last time step multiple in the past that you'll have data from for each prediction (inclusive, 0 by default that means that the immediate values before the prediction time step is known)




1. DATA TYPE

 Tabular

 Timeseries

 Images

2. TRAINING TYPE

 Regression

- Prediction window : illustrate the period in the future that you want to predict * Start of the prediction window : the first time step multiple you want to predict (inclusive, 1 by default which means we will predict starting at the next value) * End of the prediction window : the last time stamp multiple you want to predict (inclusive, 10 by default which means we will predict up to the 10th next value)
- A priori features : features whose value is known in the future (customer number, calendar, public holidays, weather...)
- Group features : features that identify a unique time serie (e.g. you want to predict your sales by store and by product. If you have 2 stores selling 3 products, there are 6 time series in your file. Selecting features « store » and « product in the group column allows Prevision.io to take into account these multiple series)

Please note that advanced options work the same way than for tabular usecases. Please read the corresponding readthedoc section in order to configure your time series usecase.

Usecases images

In the prevision.io platform, you can train several usecases type using images :

regression classification multi-classification image detection

For the first three kinds of training, the user flow is similar to the corresponding tabular usecases. You will just need in addition to add an image folder corresponding to the train dataset in the new usecase screen.

Image usecases configuration

The image usecases training workflow is similar to the tabular corresponding usecases (except for image detection). Please refer to the tabular datatype training read the doc section in order to get information about the train settings. However this similarity, some differences are notable. On the field configuration options, an image path is required. This image path is the link between the tabular train dataset and the corresponding images.

Image detection usecases *Introduction*

In the prevision.io platform, a particular kind of image usecase allows you to train models that are able to recognize and boxing on an image a particular object.

In order to train image detection usecases you will need to have an image folder and a tabular document including :

- the image path
- the object label
- the bounding box coordinates

The image detection usecase training configuration is simpler than for other training. In advance options, only training performances choice between quick, normal & advanced is available.

versioning of a usecase

In the prevision.IO platform you can create multiple versions of one usecase allowing you to search for optimal performance training and, deploy and switch any model from any version of the same usecase.

In order to do that, several possibilities :

- From the usecase list, by clicking on the “action button” of an entry and selecting “new version”
- one a usecase page, by clicking on the “action” button and selecting “new version”
- On the version menu from a usecase and selecting “new version” in the list action button

Then, you will be redirected to the “new usecase” page but with limited option. First of all, you can not change the datatype and training type between version

duplication of a usecase

In order to duplicate a usecase, there is two options :

- by using the action button right side of the usecase list

Fields configuration

Target column

qt_traffic_ref

SET YOUR TIME LIMITS

Temporal column

dt_mois

Detected timestep

1 month



More info on timeseries usecases

PAST

What data history do you have for each prediction?

timeseries.start_dw (-start_dw)

30

× 1 month

+Add an observation time

FUTURE

On which period in the future do you want to predict?

timeseries.end_fw (end_fw)

10

× 1 month


+Add an observation delay





OPTIONAL CONFIGURATION

ID column


1. DATA TYPE


 Tabular


 Timeseries


 Images

2. TRAINING TYPE

 Regression

 Classification

 Multi-Classification

 Object-Detection

3. NAME & DATASET SELECTION

Next step

Usecase name

Usecase description (optional)

Dataset ?

Image Folder

Fields configuration

Target column

Image path

Fields configuration

Image path

Object class column

Top

Bottom

Left

Right

— by using the “action button” on top right of any usecase page and select “duplicate usecase”
By doing this, the new usecase screen will appear keeping the duplicated usecase configuration.

Usecase general navigation

navigation

For each training type (except image detection usecases) you will find the same navigation allowing you to explore your usecase and models.

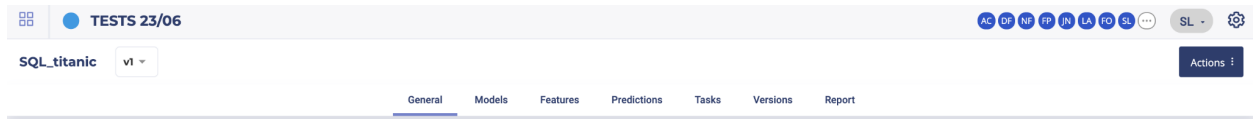
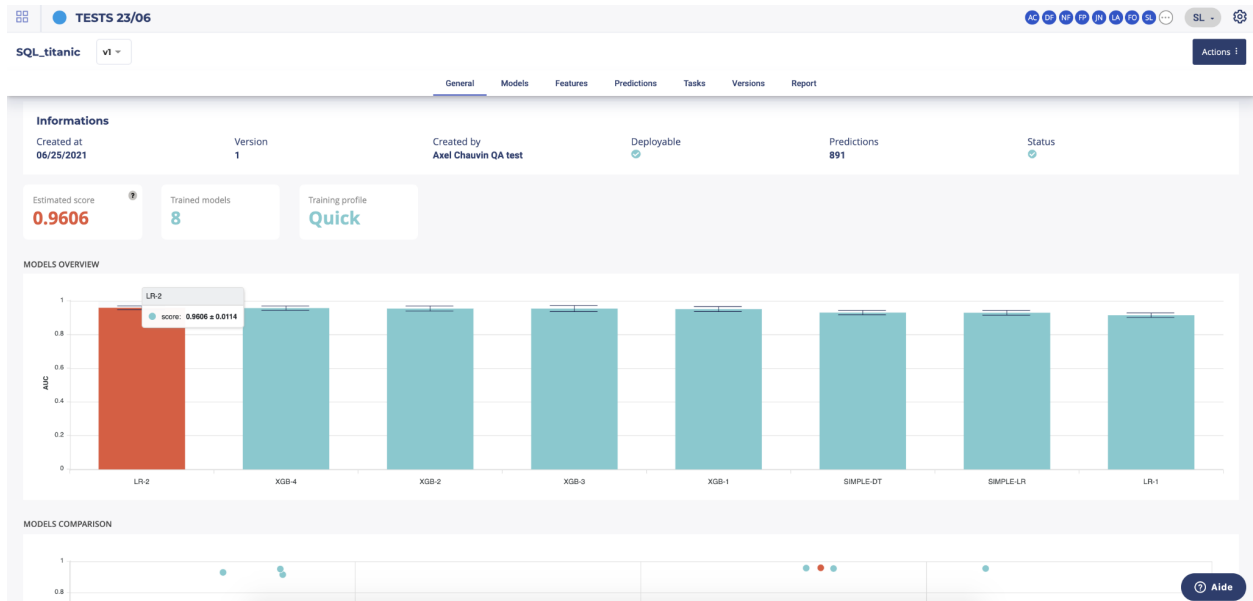


Fig. 5 – image alt text

- General : general information and comparison of your models in terms of performances
- models : list view of the created models and information about the trained models
- Features : information about the dataset used for the training and the configuration of the feature engineering
- Prediction : create bulk predict using CSV files and view all bulk prediction done for this usecase
- Task : DAG and listing of all operations done during training
- Versions : list of all version of the selected usecase
- Report : generate PDF reports explaining the models/usecases

General

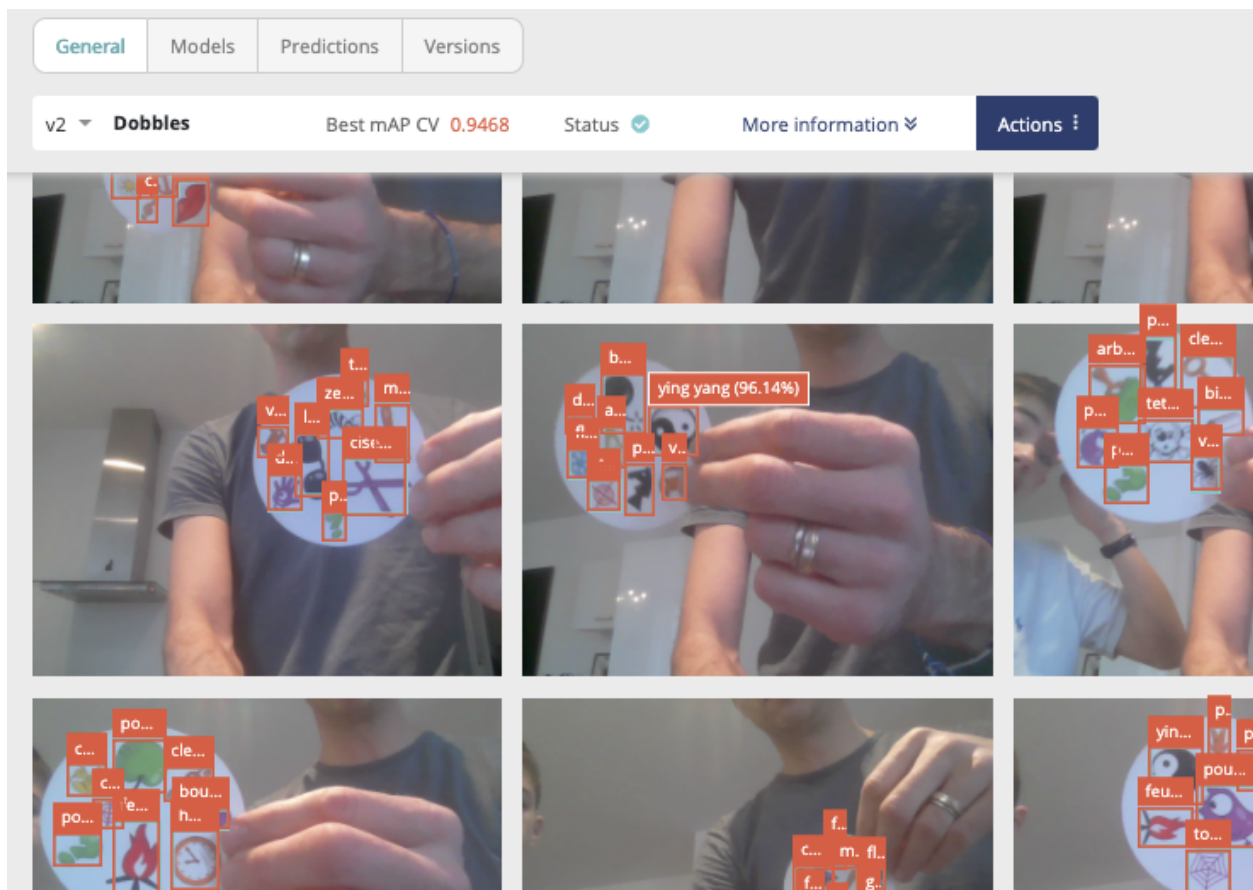
In this menu, you will find general information about your usecase such as the list of created models.



- The information header gives you the important information regarding your usecase. You can expand this panel by clicking on “more information”. Please note that by default, by entering a usecase, the information will be all about the latest version of the usecase. You can navigate through the versions using the dropdown list on the left side of this panel.

- Action button : on the right side of the information panel, you will find the actions buttons allowing you to :
 - edit the name and description of the usecase version
 - create a new version
 - duplicate the usecase
 - delete the usecase
- Under the information panel, cards displaying information regarding your usecase are displayed. Please note that the holdout score card will be displayed only if a holdout was selected during training configuration
- Two graphs are displayed on the general page of a usecase showing :
 - The models ranked by score. By clicking on a model chart bar, you can access to the selected model details
 - Models score vs. estimated prediction time

Please note that for object detection, the general screen is quite different from the other use cases types. On the image detection general menu you will find a sample of images used during the train in orange, the predicted bounding boxes using cross validation and in blue, the true bounding boxes.



models

By clicking on the models menu of top usecase navigation, you will access the model list trained for this usecase version. you will also, at the bottom of the page, find information regarding the model selected for this train.

By clicking on the model name in the list, you will be redirected to the model detail page. Please note that a toggle button is available on the right side of the list for each model. This toggle allows you to tag a model as deployable. In order to know how to deploy a model, please go to the dedicated section.

TESTS 23/06

S3_titanic v1

General Models Features Predictions Tasks Versions Report

MODELS DONE

| NAME | TECHNOLOGY | TYPE | SCORE | TRAINING DURATIONS | PREDICT DURATIONS | ARRIVAL TIME | DEPLOYABLE |
|-----------------------------------|----------------|------|-----------------|--------------------|-------------------|----------------------|--------------------------|
| LR-2 <i>Best performance</i> | Logistic model | Base | 0.9587 ± 0.0046 | 2.6s | 166ms | 06/25/2021, 14:44:05 | <input type="checkbox"/> |
| XGB-1 | XGBoost | Base | 0.9527 ± 0.0052 | 3.4s | 57ms | 06/25/2021, 14:43:32 | <input type="checkbox"/> |
| XGB-3 | XGBoost | Base | 0.952 ± 0.007 | 3.3s | 166ms | 06/25/2021, 14:44:20 | <input type="checkbox"/> |
| XGB-4 | XGBoost | Base | 0.9495 ± 0.0068 | 3.4s | 170ms | 06/25/2021, 14:45:17 | <input type="checkbox"/> |
| XGB-2 | XGBoost | Base | 0.9493 ± 0.0043 | 3.3s | 171ms | 06/25/2021, 14:44:00 | <input type="checkbox"/> |
| SIMPLE-LR | Logistic model | Base | 0.9285 ± 0.0108 | 2.6s | 34ms | 06/25/2021, 14:44:06 | <input type="checkbox"/> |
| SIMPLE-DT <i>Fastest model</i> | Decision Tree | Base | 0.927 ± 0.008 | 3.1s | 32ms | 06/25/2021, 14:43:56 | <input type="checkbox"/> |
| LR-1 | Logistic model | Base | 0.9139 ± 0.0107 | 2.9s | 54ms | 06/25/2021, 14:43:32 | <input type="checkbox"/> |

rows per page: 25

MODEL SELECTION

Simple models:

✓ Logistic model ✓ Decision Tree

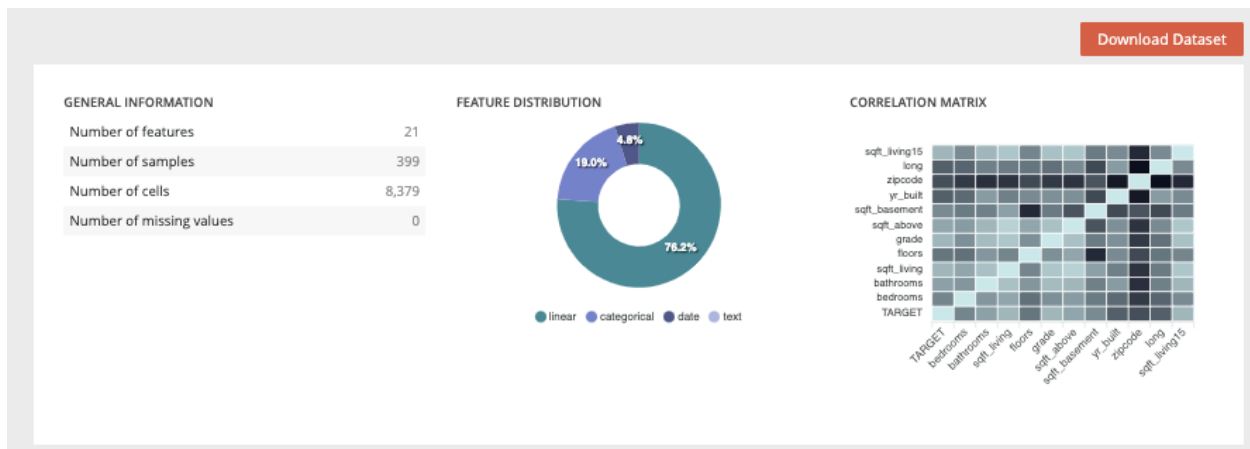
Normal models:

✓ XGBoost ✓ Logistic model ✗ Extra Trees ✗ Random Forest ✗ LightGBM ✗ Neural Network ✗ Naive Bayes ✗ CATBoost

Aide

features

In this section you will find any information relative to the dataset used during the train.



On the top panel, you will find generic information about the dataset used during the train such as the number of columns, number of samples and number of cells or the usecases using this dataset.

You can also download the dataset used for the training by clicking on the “download dataset” button on top of the page.

Two graph are also displayed showing :

- the feature distribution regarding the feature type (linear, categorical, date or text). This distribution is automatically calculated when uploading a dataset into the platform
- correlation matrix showing the correlation coefficients between variables. Each cell in the table shows the correlation between two variables

Under this top panel, three tabs are available :

- Features analysis : table displaying features information calculated after the upload of the dataset such as the % of missing value.

- Drop features : In this tab, you will find a list of all features you dropped for the usecase training during the usecase configuration
- Configuration : list of all feature engineering applied on the dataset during training

predictions

The predictions menu allows you to do bulk predictions using a previously loaded dataset and see holdout predictions made during training.

SELECT A MODEL
XGB-3 (158,625)

SELECT A DATASET

☐ confidence

Launch Prediction

In order to do a new prediction, you have to first select a model from the dedicated dropdown list and then a dataset uploaded on the project. Then, by clicking on the “launch prediction” button, the system will compute and generate a prediction file downloadable by clicking on the right side button on the prediction list below.

Prediction explain

You can make more in-depth analyses of the predictions made this way by generating explanations for each of the predictions of the dataset. You can access the explanation screen by clicking on the **image alt text** menu on the table entry for the prediction you wish to explain. (Depending on the complexity of the model and the size of your dataset, the screen may take a few seconds to load).

FILTER DATASET
age =
cancel filters filter

SAMPLES
Prediction: probability that the target column equals **yes**.

| ID | Prediction | target |
|----|------------|--------|
| 1 | 0.038 | no |
| 2 | 0.0375 | no |
| 3 | 0.0356 | no |
| 4 | 0.0334 | no |
| 5 | 0.2147 | no |
| 6 | 0.5375 | no |
| 7 | 0.0328 | no |
| 8 | 0.0349 | no |
| 9 | 0.0299 | no |
| 10 | 0.0531 | no |
| 11 | 0.0319 | no |
| 12 | 0.0322 | no |
| 13 | 0.0322 | no |
| 14 | 0.155 | no |

Explanation - Sample # 1
bank-marketing

PREDICTION
0.038029
yes

TARGET
no

The model has predicted 0.038029 for the sample 1.
Features **nr.employed**, **euribor3m**, **month**, **duration**, and **emp.var.rate** had a positive influence.

FEATURE INFLUENCE

VARIABLES simulate

age: 39 (min 18, max 88)
job: services
marital: single
education: high.school
default: no
housing: no

The explanation screen is composed of three parts :

- The “filter dataset”, on the left, that allows you to select a specific prediction from your dataset to be explained, as well as apply specific filters to the dataset to select predictions you are interested in. A filter is defined as :
 - a variable present in your dataset (selected from the dropdown)
 - an operator
 - a value. All rows matching the clause will be returned. You can apply up to two filters, and select whether both filters should be applied (“and”), or if a row matching any of the two filters should also be returned (“or”).

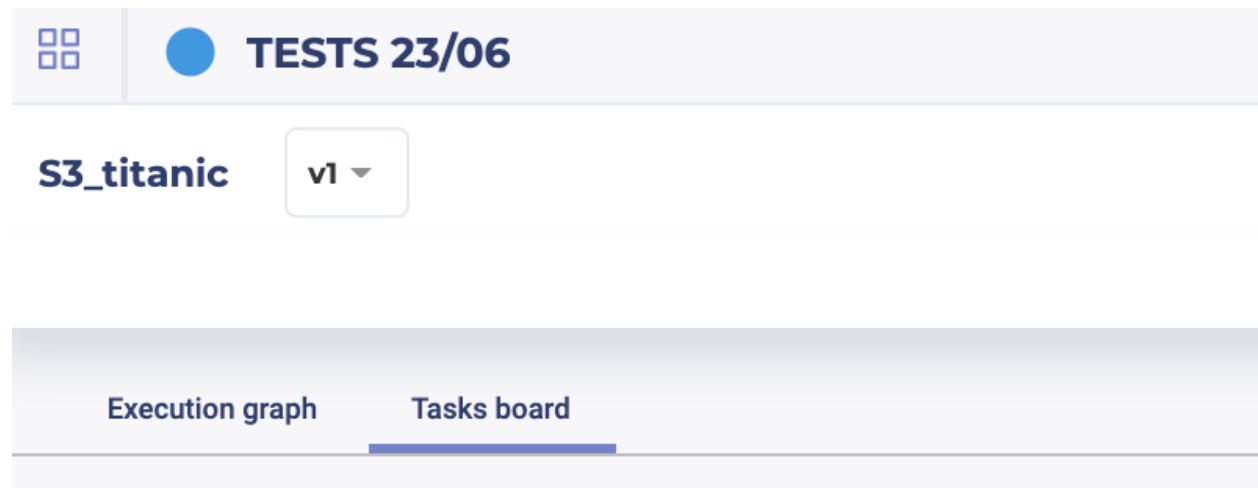
- The “explanation”, on the top right, displays the prediction for the currently selected line, the actual value of the target variable (if it was present in the dataset), as well as the explanation, shown as the relative impact (which can be positive or negative) of the different variables on the final decision.
- The “variables” part, on the bottom right, allows you to conduct “what-if” analyses and see how the prediction and explanations can evolve when the values of the variables are adjusted. When you click on the “simulate” button, the prediction & explanation above will be updated.

tasks

In this menu you will find an overview of all tasks made by the platform during the usecase training and their status. The aim of this screen is to help you to better understand the operations made during the training and, if errors occurred, at which level it happened.

For that, two views are available :

- Liste view : list all single operations done
- DAG view : graphical view of single operations and their relation



You can switch between these views by clicking on the execution graph / tasks board tabs.

versions

In the Prevision.IO platform you can iterate versions of your usecases. To do that, three possibilities :

- On the usecase list of a project, by clicking on the right side action button of the usecase you want to iterate and select “new version”
- On any pages of a usecase by clicking on the top right “actions” button and select new version
- On the “Version” menu of a usecase, by clicking on the action button right side of a version listed and select “new version”

Then, on the version menu of a usecase, you will find the list of all trained versions for this usecase. By clicking on the version number, left side of this list, you will access the selected usecase version page. You can also navigate through versions by using the dropdown list top left of the information banner on any page of a usecase.

After clicking on a new version button, you will be redirected to the usecase configuration menu. The usecase version configuration you selected for your iteration will be automatically loaded in order for you to know what configuration was done and select the changement you want to apply.

TIPS : when creating a new usecase or a new version, add a description to your usecase at the first screen of new usecase configuration. It will help you finding the version you want to work with later.

| VERSION | DESCRIPTION | CREATED AT | CREATED BY | SCORE | MODELS | PREDICTIONS | STATUS |
|---------|-------------|-------------------|----------------------|-----------------|--------|-------------|--------|
| V2 | | 06/28/2021, 16:06 | Simon Levacher | - | 0 | 0 | |
| V1 | | 06/25/2021, 14:42 | Axel Chauvin QA test | 0.9587 (auc) ★★ | 8 | 0 | |

report

In this menu, you can generate PDF reports regarding models from the usecase. To do that, once on the dedicated model menu, you will have to choose from the drop down the models you want to appear in the generated report and the feature importance count. You also can select explanations by check/uncheck the show explanation checkbox. Then, by clicking on the generate button, you will get an overview of the report. By clicking on the “print” button on the top of the overview, you will download the PDF report.

models pages

Each model page is specific to the datatype/training type you choose for the usecase training. Screens and functionality for each training type will be explained in the following sections. You can access a model page by two ways :

- by clicking on a graph entry from the general usecase page
- by clicking on a list entry from the models top navigation bar entry

Then you will land on the selected model page splitted in different parts regarding the training type.

tabular usecases - general information

For each kind of tabular training type, the model general information will be displayed on the top of the screen. Three sections will be available.

| MODEL INFORMATIONS | | HYPERPARAMETERS | | SELECTED FEATURE ENGINEERINGS |
|---------------------------|----------------------|------------------|--------------------|--|
| Holdout | 153,732 | colsample_bytree | 0.8112003946594193 | ✓ One hot encoding of categorical features ⓘ |
| model type | XGB | eta | 0.05 | ✓ Linear feature scaling ⓘ |
| score | 158,625 | eval_metric | rmse | |
| metric | rmse | max_depth | 9 | |
| metric standard deviation | 45,409 | min_child_weight | 5 | |
| train duration | 4.3s | objective | reg:linear | |
| predict response time | 252ms | reg_lambda | 17.16949112255041 | |
| deployable | Yes | silent | 1 | |
| model used for blend | No | subsample | 0.8706304835460907 | |
| arrival time | 05/18/2021, 12:48:21 | feature_selected | ["ohe", "lin"] | |
| | | seed | 481873 | |
| | | num_boost_round | 358 | |

- Model information : information about the trained model such as the selected metric and the model score
- Hyperparameters : downloadable list of hyperparameters applied on this model during the training
- Selected feature engineerings (for regression, classification & multi-classification) : features engineerings applied during the training

- Preprocessing (for text similarity usecases) : list of pre-processing applied on textual features

Please note that for following usecases types, the general information parts is different than from others :

- Image detection usecases : no feature engineering
- text similarity usecases : preprocessing are displayed instead of feature engineering

Model page - Graphical analysis

In order to better understand the selected model, several graphical analyses are displayed on a model page. Depending on the nature of the usecase, the displayed graphs change. Here an overview of displayed analysis depending on the usecase type.

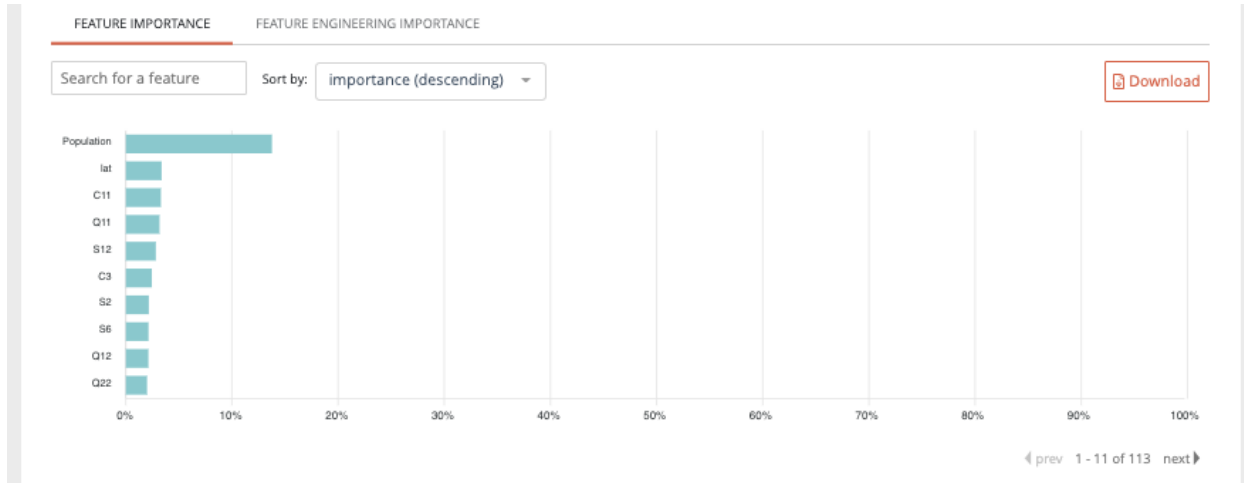
Tableau 4 – Type of Training. Tabular Data

| | Tabular regression | Tabular classification | Tabular multi-classification | Tabular text similarity | Time series regression | Image regression | Image classification | Image multi-classification | Image detection |
|------------------------------|--------------------|------------------------|------------------------------|-------------------------|------------------------|------------------|----------------------|----------------------------|-----------------|
| Scatter plot graph | Yes | No | No | No | Yes | Yes | No | No | No |
| Residual errors distribution | Yes | No | No | No | Yes | Yes | No | No | No |
| Score table (textual) | Yes | No | No | No | Yes | Yes | No | No | No |
| Residual errors distribution | No | No | No | No | No | No | No | No | No |
| Score table (overall) | No | No | Yes | No | No | No | No | Yes | No |
| Cost matrix | No | Yes | No | No | No | No | Yes | No | No |
| Density chart | No | Yes | No | No | No | No | Yes | No | No |
| Confusion matrix | No | Yes | Yes | No | No | No | Yes | Yes | No |
| Score table (by class) | No | Yes | Yes | No | No | No | Yes | Yes | No |
| Gain chart | No | Yes | No | No | No | No | Yes | No | No |
| Decision chart | No | Yes | No | No | No | No | Yes | No | No |
| lift per bin | No | Yes | No | No | No | No | Yes | No | No |
| Cumulated lift | No | Yes | No | No | No | No | Yes | No | No |
| ROC curve | No | Yes | Yes | No | No | No | Yes | Yes | No |
| Accuracy VS K results | No | No | No | Yes | No | No | No | No | No |

Model page - graphs explanation

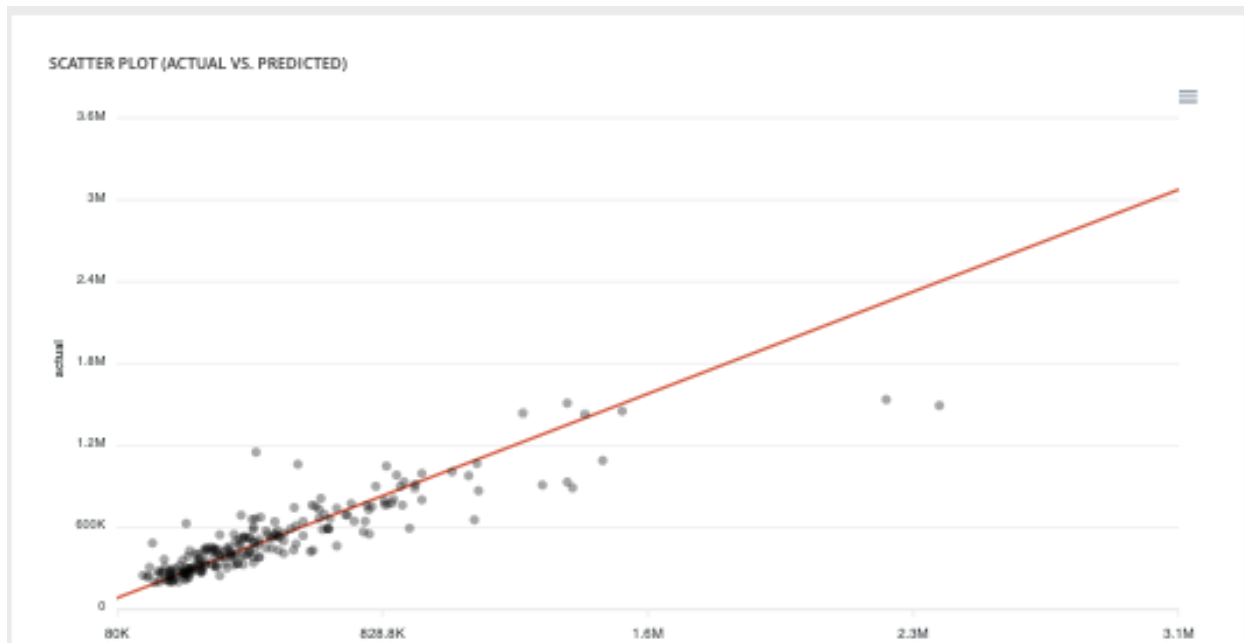
Then the feature graphs will be displayed (not for text similarity) allowing you to see the influence of features for the selected model. Two graphs are accessible through the two features tabs :

- Feature importance : graph showing you the importance of the dataset features. By clicking on the chart, you will be redirected to the dedicated feature page.
- Feature engineering importance : showing you the importance of selected feature engineering.

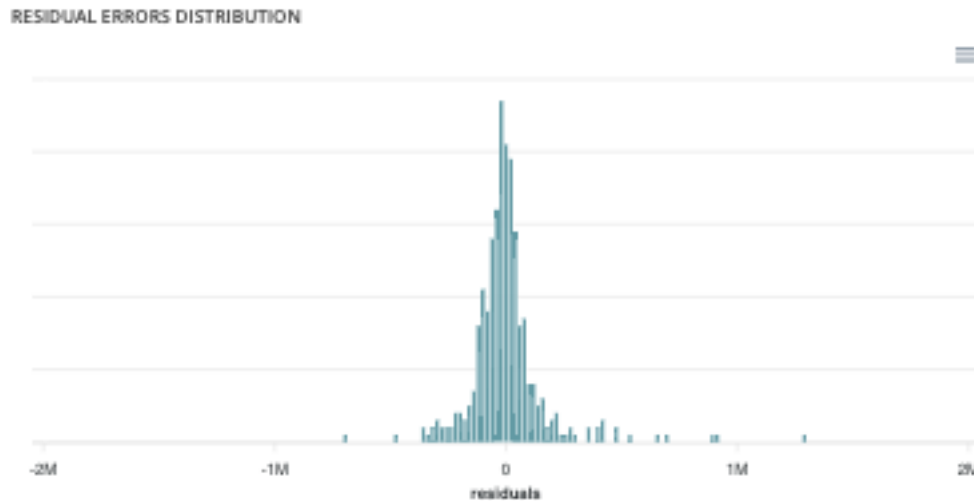


Please note that the feature importance graph also takes into account the feature engineering importance. For example, if a feature n°1 has not so much influence by itself regarding the model but, after feature engineering has a great influence, it will be represented on the feature importance graph.

- Scatter plot graph : This graph illustrates the actual values versus the values predicted by the model. A powerful model gathers the point cloud around the orange line.



- Residual errors distribution : This graph illustrates the dispersion of errors, i.e. residuals. A successful model displays centered and symmetric residues around 0.
- Score table (textual) : Among the displayed metrics, we have :



- The mean square error (MSE)
- The root of the mean square error (RMSE)
- The mean absolute error (MAE)
- The coefficient of determination (R2)
- The mean absolute percentage error (MAPE)

SCORE TABLE

| | |
|--------------------------------|----------------|
| Mean squared error | 27,223,899,929 |
| Root mean squared error | 164,997 |
| Mean absolute error | 94,770 |
| R2 | 99.95% |
| Mean absolute percentage error | 18.55% |

Please note that you can download every graph displayed in the interface by clicking on the top right button of each graph and selecting the format you want.

- Slider : For a binary classification, some graphs and scores may vary according to a probability threshold in relation to which the upper values are considered positive and the lower values negative. This is the case for :
 - The scores
 - The confusion matrix
 - The cost matrix

Thus, you can define the optimal threshold according to your preferences. By default, the threshold corresponds to the one that minimizes the F1-Score. Should you change the position of the threshold, you can click on the « back to optimal » link to position the cursor back to the probability that maximizes the F1-Score.



- Cost matrix : Provided that you can quantify the gains or losses associated with true positives, false positives, false negatives, and true negatives, the cost matrix works as an estimator of the average gain for a prediction made by your classifier. In the case explained below, each prediction yields an average of €2.83.

The matrix is initiated with default values that can be freely modified.

- Density chart : The density graph allows you to understand the density of positives and negatives among the

COST MATRIX

| | | | |
|-----------------|-------------------|-----------|-------------------|
| value = true/1 | predict = true/1 | gain = 10 | expected = 0.248 |
| | predict = false/0 | gain = -5 | expected = -0.804 |
| value = false/1 | predict = true/1 | gain = -5 | expected = -0.088 |
| | predict = false/0 | gain = 5 | expected = 3.984 |

predictions. The more efficient your classifier is, the more the 2 density curves are disjointed and centered around 0 and 1.

- Confusion matrix : The confusion matrix helps to understand the distribution of true positives, false positives, true negatives and false negatives according to the probability threshold. The boxes in the matrix are darker for large quantities and lighter for small quantities.

Ideally, most classified individuals should be located on the diagonal of your matrix.

- Score table (graphical) : Among the displayed metrics, we have :
 - Accuracy : The sum of true positives and true negatives divided by the number of individuals
 - F1-Score : Harmonic mean of the precision and the recall
 - Precision : True positives divided by the sum of positives
 - Recall : True positives divided by the sum of true positives and false negatives
- Gain chart : The gain graph allows you to quickly visualize the optimal threshold to select in order to maximise the gain as defined in the cost matrix.
- Decision chart : The decision graph allows you to quickly visualize all the proposed metrics, regardless of the probability threshold. Thus, one can visualize at what point the maximum of each metric is reached, making it possible for one to choose its selection threshold.

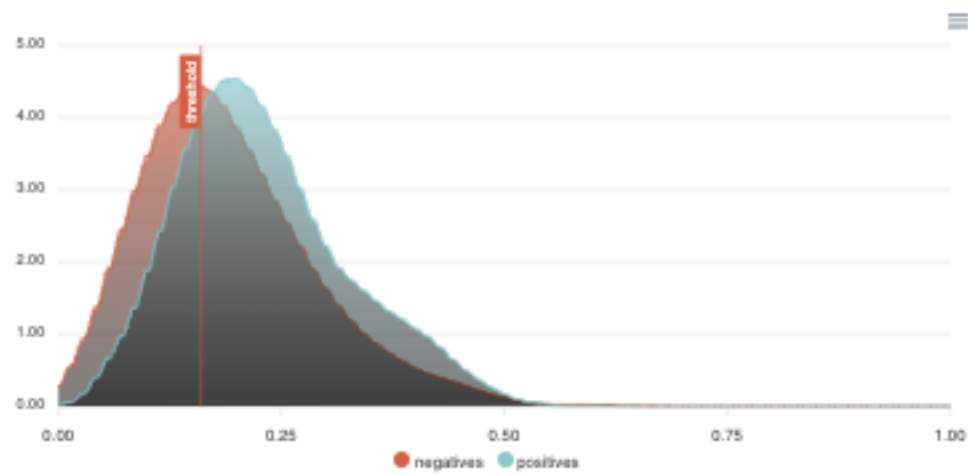
It should be noted that the discontinuous line curve illustrates the expected gain by prediction. It is therefore totally linked to the cost matrix and will be updated if you change the gain of one of the 4 possible cases in the matrix.

- lift per bin : The predictions are sorted in descending order and the lift of each decile (bin) is indicated in the graph. Example : A lift of 4 means that there are 4 times more positives in the considered decile than on average in the population.

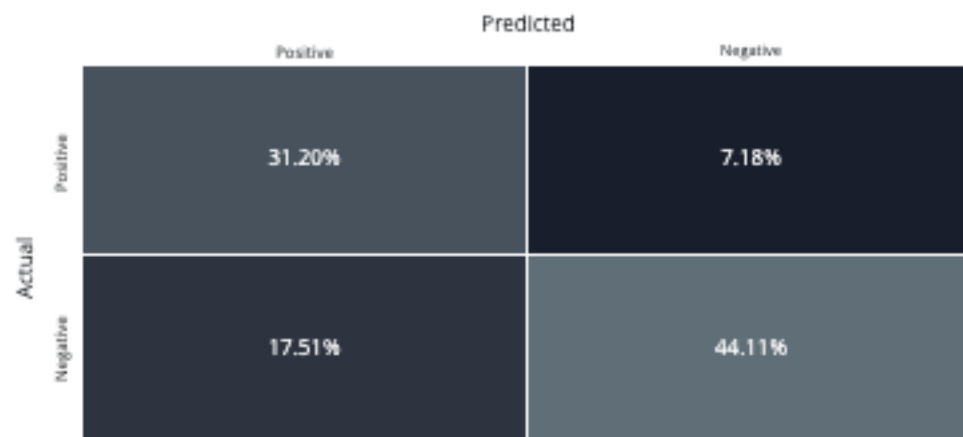
The orange horizontal line shows a lift at 1.

- Cumulated lift : The objective of this curve is to measure what proportion of the positives can be achieved by targeting only a subsample of the population. It therefore illustrates the proportion of positives according to the proportion of the selected sub-population.

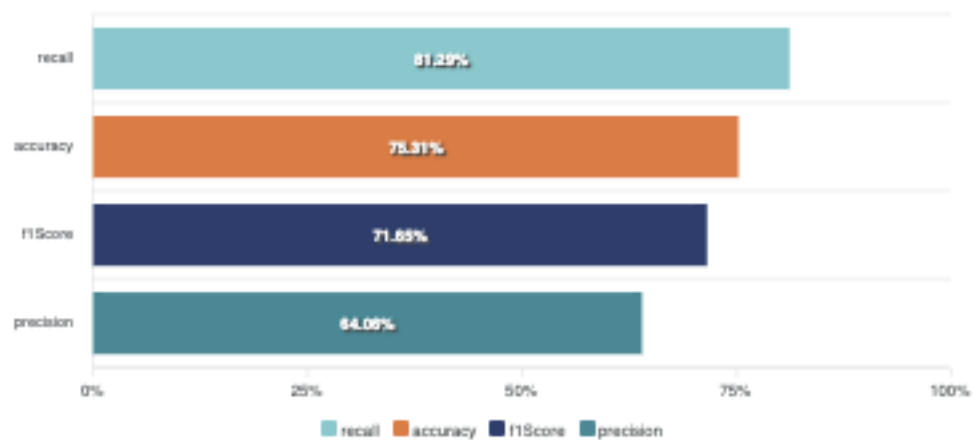
DENSITY CHART

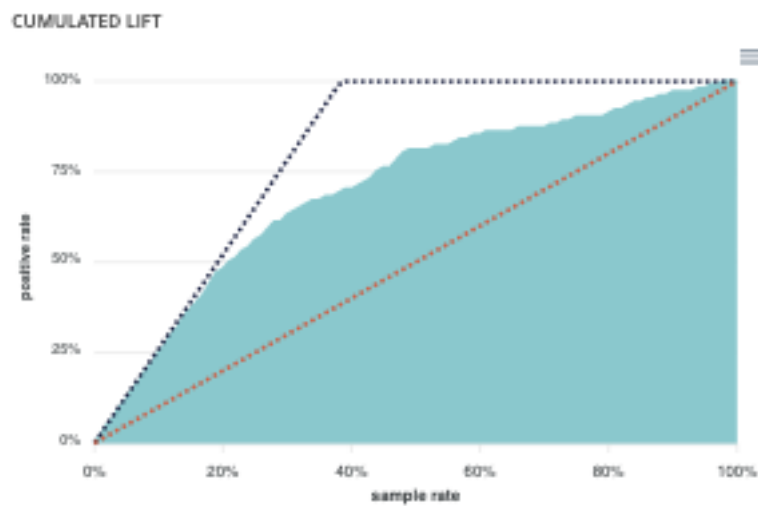
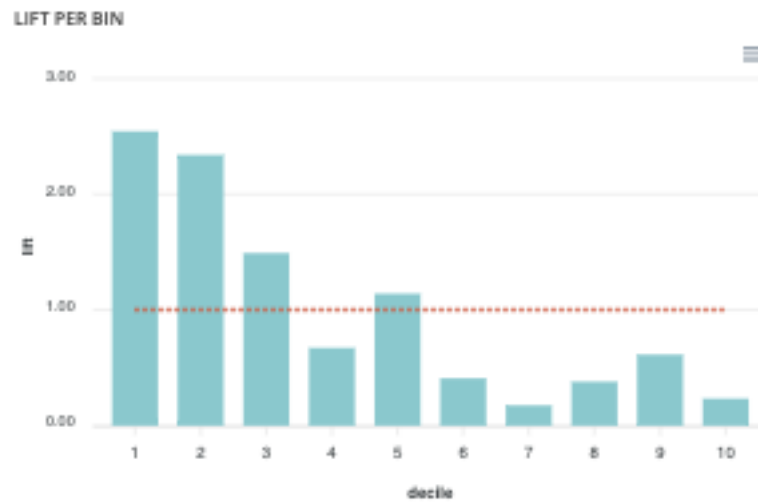
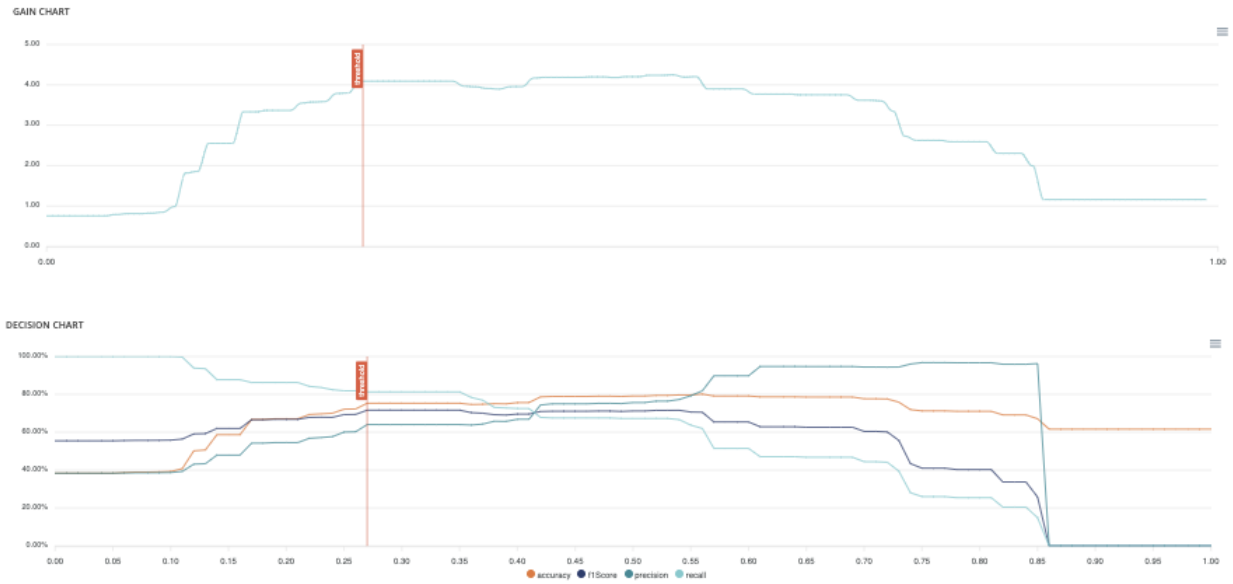


CONFUSION MATRIX



SCORE TABLE

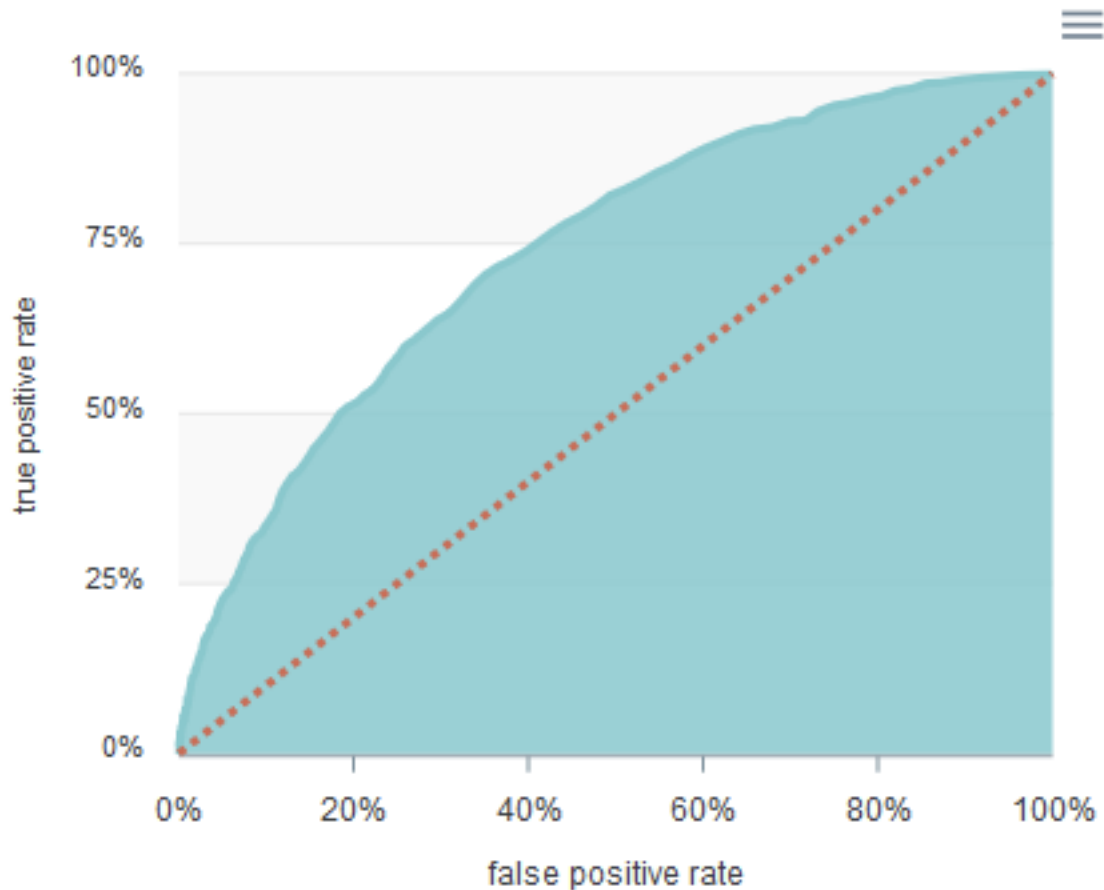




A diagonal line (orange) illustrates a random pattern (= x % of the positives are obtained by randomly drawing x % of the population). A segmented line (blue) illustrates a perfect model (= 100% of positives are obtained by targeting only the population's positive rate).

- ROC curve : The ROC curve illustrates the overall performance of the classifier (more info : https://en.wikipedia.org/wiki/Receiver_operating_characteristic). The more the curve appears linear, the closer the quality of the classifier is to a random process. The more the curve tends towards the upper left side, the closer the quality of your classifier is to perfection.

ROC CURVE (AUC = 0.7358)



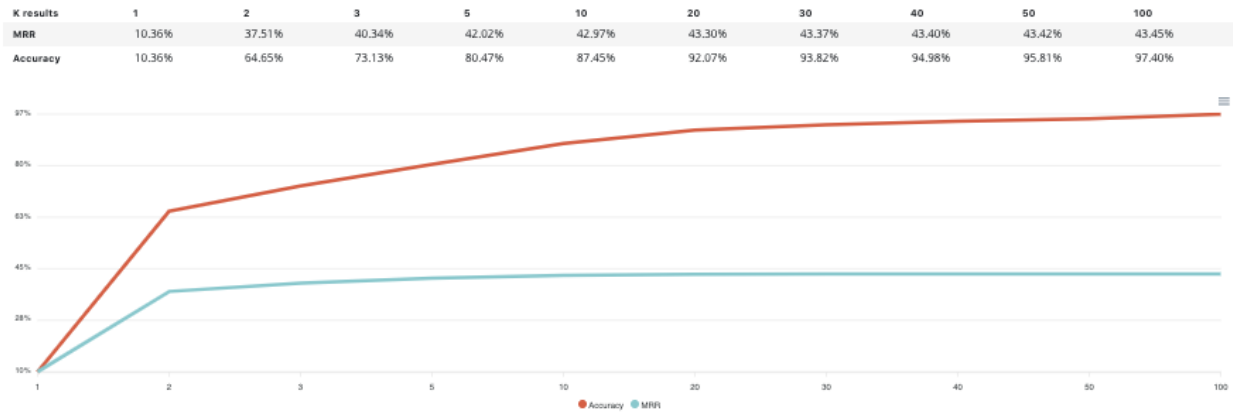
- Accuracy VS K results : this graph shows the evolution of accuracy and MRR for several value of K results

2.2.4 Pipelines

liste

pipeline components

pipeline template



pipeline runs

2.2.5 Contributors

By clicking on contributors on the main menu of a project, you will access the list of all users of the project. If you have enough rights on this project, you will be able to add & delete users from the project and modify the rights of the users.

Rôles & rules

- Viewer : you can access to all pages (except project settings) with no possibility of creation or edition
- Contributor : viewer rights + you can edit and create resources inside the project
- Admin : contributor rights + you can manage users and modify project properties

Add & delete

If you are admin in a project, you can manage users into your project.

In order to add a user, you have to enter into the top left field the collaborator email, set his right using the dropdown menu and click on the “invite this collaborator”. Please note that you can only invite collaborators that have already a prevision.io account.

In order to change the rights of a user into the project, into the list, you just have to select the new role using the dropdown. In order to be sure that the project and users properties can be manage at least one collaborator have to be admin of the project.

In order to remove a collaborator from the project, use the trash button on the left side of the list.

Project's collaborators
Share this project with collaborators

| NAME | EMAIL | ROLE | |
|----------------------|-----------------------------|-------|--|
| Axel Chauvin QA test | axel.chauvin@prevision.io | admin | |
| Simon Levacher | simon.levacher@prevision.io | admin | |

Project settings

If you are admin on a project, the project settings button is enabled and, by clicking on it, you will access the project setting page.

The screenshot shows the 'Project's settings' page. On the left, under 'Project's settings', there is a form to 'Edit the project's settings'. It includes a 'Project's name' field with the value 'QA - Multiclassification' (24 / 40 characters), a 'Project's description (optional)' field (0 / 210 characters), and a 'Set a color to this project' section with a purple color picker and a 'Change color' button. At the bottom are three buttons: 'Cancel', 'Save new project's settings', and 'Delete the project'. On the right, the 'PREVIEW' section shows the project details: 'QA - Multiclassification', user 'axel.chauvin@prevision.io', and timestamp '05/18/2021, 11:06'. Below this, it shows '2' documents and '4' datasets, and a list of collaborators: 'AC', 'SL', and 'UU'.

You can on this page :

- update name, description and color of your project
- delete the project. Please note that if you delete a project, all ressources linked to the project will be deleted (usecases, datasets, deployed models...)

2.2.6 Notebooks

Introduction

Prevision.io offers various tools to enable data science use cases to be carried out. Among these tools, there are notebooks and production tools. Notebooks are not scoped into projects. You can access notebooks by clicking on the notebook button on the left main menu. Then you will be redirected to the following page.

Jupyter (python)

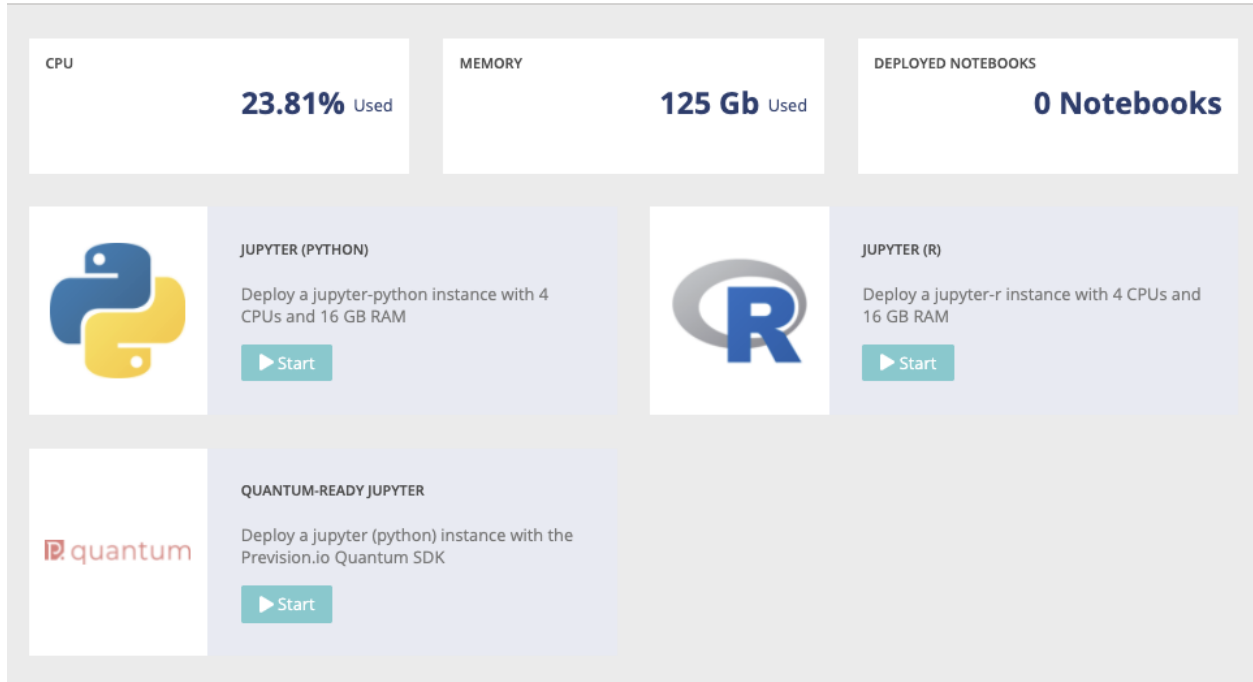
For Python users, a JUPYTERLAB environment (<https://github.com/jupyterlab/jupyterlab>) is available in Prevision.io

Note that a set of packages is preinstalled on your instance (list : https://previsionio.readthedocs.io/fr/latest/_static/ressources/packages_python.txt), particularly the previsionio package that encapsulates the functions using the tool's native APIs. Package documentation link : <https://prevision-python.readthedocs.io/en/latest/>

Jupyter (R studio)

For R users, a R STUDIO environment (<https://www.rstudio.com>) is available in Prevision.io

Note that a set of packages is preinstalled on your instance (list : https://previsionio.readthedocs.io/fr/latest/_static/ressources/packages_R.txt), particularly the previsionio package that encapsulates the functions that use the tool's native APIs. Package documentation link : https://previsionio.readthedocs.io/fr/latest/_static/ressources/previsionio.pdf



2.2.7 Help

By clicking on help on the main menu, you will be redirected to the prevision's resources helping you through the application and usecases. Four sections are available :

- type of problem : helping you to define what kind of usecase type is the most suitable for your issue
- videos : centralisation of tutorials & data-science resources
- medium post : our datascience dedicated posts published on medium
- documentation : link to the application documentation such as the readthedoc or the SDK documentation

2.2.8 User

- Language : switch the language between french or english
- Profile : navigate to your profile information such as email and password
- administration & API key : available only for admins
- documentation : ReadTheDoc redirection
- Terms and conditions
- log out

