Technical Specification Double-side Page

1. **TECHNICAL SCOPE:** Summarise the mock-up devised during the EXPLORE phase: how have you addressed the challenge/Theme Challenges and tackled with its requirements and data. Include a diagram.

As B2Metric we have an Automated Machine Learning (Auto-ML) platform solution that provides low code modelling solutions for supervised, unsupervised learning and deep learning models. B2ML Studio provides an insurance specific product called Hunter, which runs Auto-ML for insurance customer risk scoring, churn prediction and fraud detection problem solutions. B2ML Studio Hunter enables users to automatically discover, visualise and narrate important findings (such as correlations, exceptions, clusters, drivers and predictions) in datasets, without requiring people to build data visualisations, create models or write algorithms. Augmented capabilities are differentiating features in platforms across D&A. They're also a key factor accelerating the convergence between analytics and data science.

B2ML Studio brings end-to-end solutions and meets these main data science situations: data preparation, data wrangling, feature engineering, selection of algorithms, training and parameter tuning, then understandable insights with reporting at clean BI dashboards.

At first, we analysed the data and did project planning for the MVP version of B2Metric Hunter Almersy's Fraud Detection platform. It's a SAAS online platforma that can be instantly implemented as an on-premises solution to any Healthcare Insurance companies and Almerys's software systems via B2ML API proxies.

As B2Metric we have 2 platform product that we run Almersy's dataset for the REACH-2021-READYMADE-ALMERYS_1 use case of DATA SCIENCE/ DATA MANIPULATION IN ORDER TO GAIN INSIGHTS FROM THE MARKET with the data of '*Historical data of reimbursement requests from opticians to insurance companies*'.

For this project we have developed and trained 3 clustering, 2 dimension reduction (factor analysis, discriminant analysis), 1 outlier detection (Local Outlier Factor) and 1 classification model and in total 6 models with our DSML team. Moreover, we run the same data in the B2ML Studio Auto-ML pipeline solution to compare it with our ML Engineering team results. We give Explainable AI module results in the previous pages. However, results of these models' results and interpretations are indicated in the ALGORITHMS, TOOLS AND CONCLUSIONS section of this document.

1-) B2Metric Auto-ML Studio & Explainable AI https://app.b2metric.com/

2) B2M IQ Analytics & Real Time BI Dashboard Reporting: https://analytics.b2metric.com/

B2ML Studio Url:	https://app.b2metric.com/
Username:	almerys@b2metric.com
Password:	Almerys1234!!

B2Metric has been recognized by world's leading management consultancy company Gartner. You can read Gartner Peer insights in B2Metric's review here:
https://www.gartner.com/reviews/market/data-and-analytics-others/vendor/b2metric

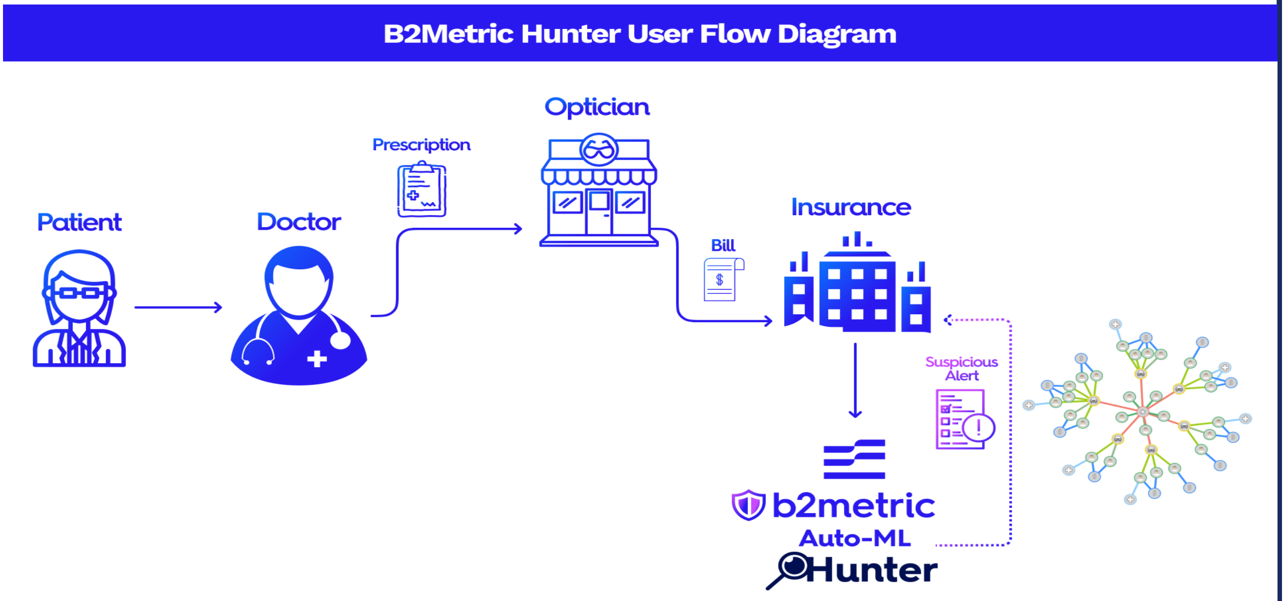
The results screen and explanations have been detailed in the sections below.

First of all, we tried to understand the variables representing various behaviours and properties of optics and their relationships with each other. It also has some logic errors like negative customer ages or negative brute prices, so they all have to be cleaned up before exploratory data analysis and ML modelling.

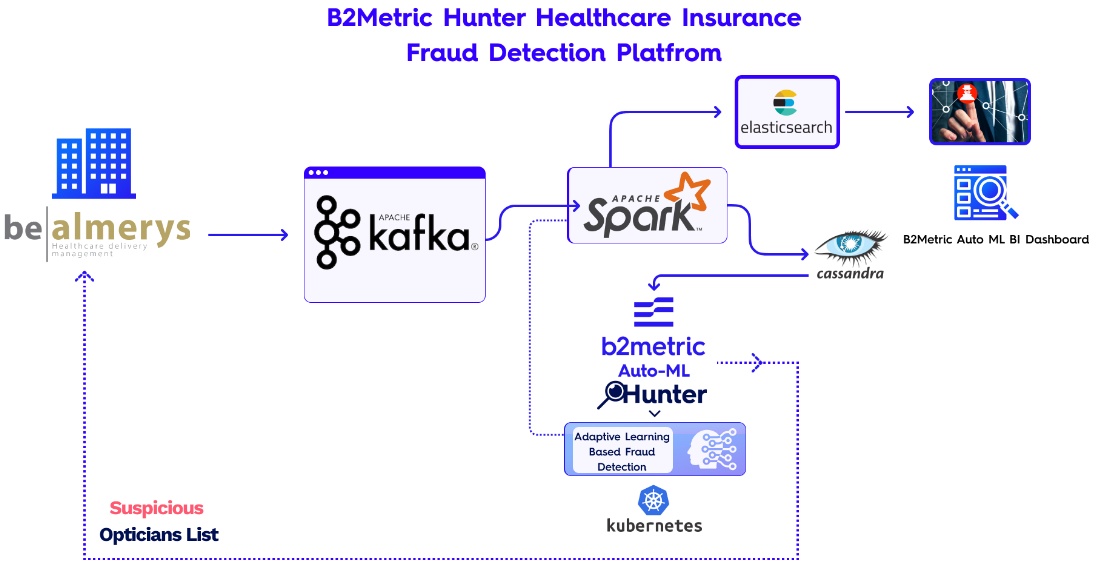
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

We have ingested Almerys datasets into B2Metric ML Studio Data Ingestor then run Fraud models and the details of models are explained in detail in the sections below.

B2Metric Auto-ML Hunter fraud use case user flow diagram indicates briefly the user personas and flows diagram below.



We have finalised successfully the MVP version of the B2Metric Hunter AutoML based on Healthcare Insurance Fraud Detection project with opticians market insights and fraud datasets of Almerys. Systems' high level system architecture design has been drawn below.



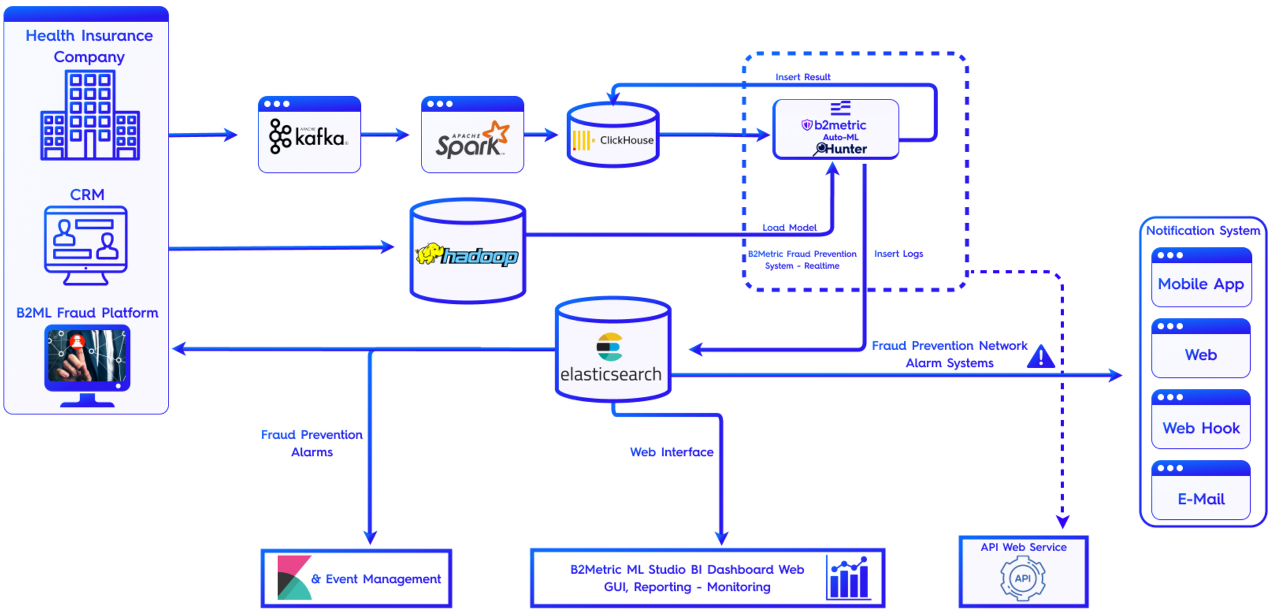
The data in .csv format is ingested in B2ML Studio's by using Kafka instances, then pre-processing step executes in Apache Spark Instance to store valid and ready to fraud prediction step in ClickHouse DB. After storing the data in clickhouse, our B2Metric Auto-ML Hunter decides which fraudulent action is in those real time streaming data by using a fraud model which was stored in Hadoop. Our Application sends all action logs and also results in ElasticSearch Clusters then triggers Notification System and BI Dashboard. This fraud alarm notification is also sent to Almersys' CRM/SAP System if necessary. If any reason Almersys' team can deeply search all logs in this pipeline by using BI Dashboard and Kibana interface, System Log Management

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

Instance.

B2Metric Auto-ML Fraud models training, retraining and prediction mechanism works the system below. This training module runs these following steps:

B2Metric Auto ML Hunter Almersy Fraud Alert Platform Design



1. The First Training Module may trigger two reasons. One of them is, it is triggered by a time scheduler if data density is sufficient to refresh the model. The other of them is, it is triggered by any unusual data which an online model has never seen before, gathered in a real time stream. Our experience shows that the first step is mostly used besides that, the second step may be used to send an alarm in the notification system that Almerys's team monitors in B2Metric BI Dashboard.
2. In this step, the Auto Feature Extraction module is run in Auto Feature Engineer Module. Three main flows are run in this step, they are Numerical, Categorical and Texture Feature Extraction.
    a. Numerical Feature Extraction:

       Skewness, Kurtosis, Shapiro-Wilk tests are calculated to decide which feature is normally distributed. If not, Standard & MinMax scalers are used separately to normalise fields.

    b. Categorical Feature Extraction:

       Categorical Features are separated by using cardinality. If there is low cardinality, Onehot or ordinal encoder is used; otherwise, embedding, hash or binning encoder is used. Domain knowledge is the most important thing because it helps us to decide which field is ordinal or nominal field.

    c. Texture Feature Extraction:

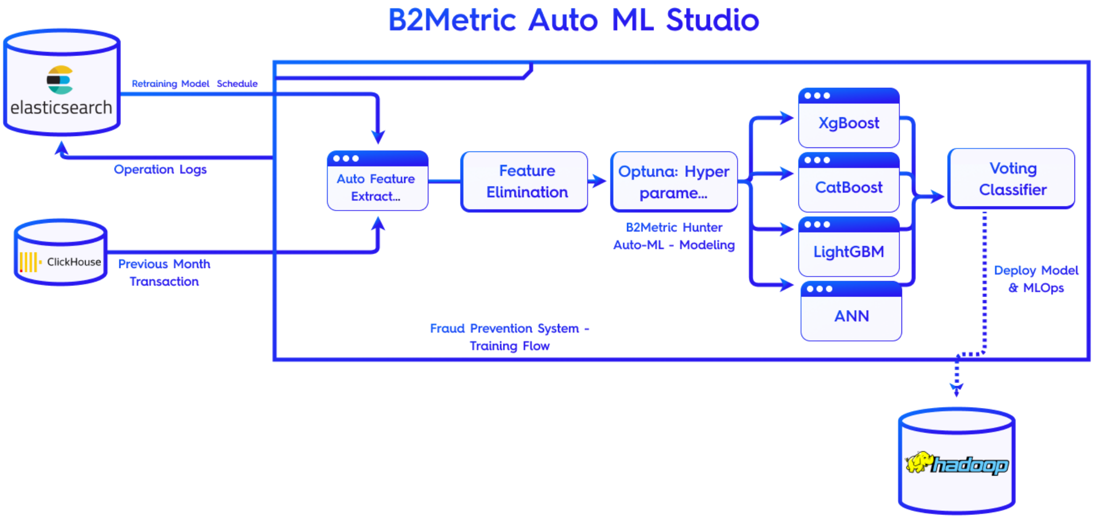
       TF-IDF, Bag of Words, Word2vec and Glove methods are used for feature extraction. Our application already supports English language but French will be included.

3. In Feature Selection, LOFO (Leave One Feature Out) or RFE (Recursive Feature Elimination) is used to eliminate feature selection. Default elimination method is LOFO.
4. Optuna is an open source hyperparameter optimization framework to automate hyperparameter search. It efficiently searches large spaces and prunes unpromising trials for faster results and also

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

Parallelize hyperparameter searches over multiple threads or processes without modifying code.

5. In the Modelling section, our application runs all selected algorithms distributedly and optimises with their parameters. After training multiple classifiers, our application uses a Voting Classifier because the voting classifier aggregates the predicted class or predicted probability on the basis of hard voting or soft voting. So if we feed a variety of base models to the voting classifier it makes sure to resolve the error by any model.
6. After modelling the pipeline, the result model is saved with its configurations and reports in the hadoop environment.



.

B2METRIC AUTO-ML HUNTER ALMERSY OPTICIANS FRAUD MODELS & EXPLAINABLE AI IMPLEMENTATION SCREENS

We ran Supervised based multiple regression, unsupervised based clustering and anomaly detection models on the B2Metric ML Studio platform with Almerys data.

For multiple regression, we ran algorithms such as Linear Discriminant Analysis, SVM, CatBoost, XGBoost, LightGBM, B2M Ensemble model. The explainable AI model results of these algorithms are reported below.

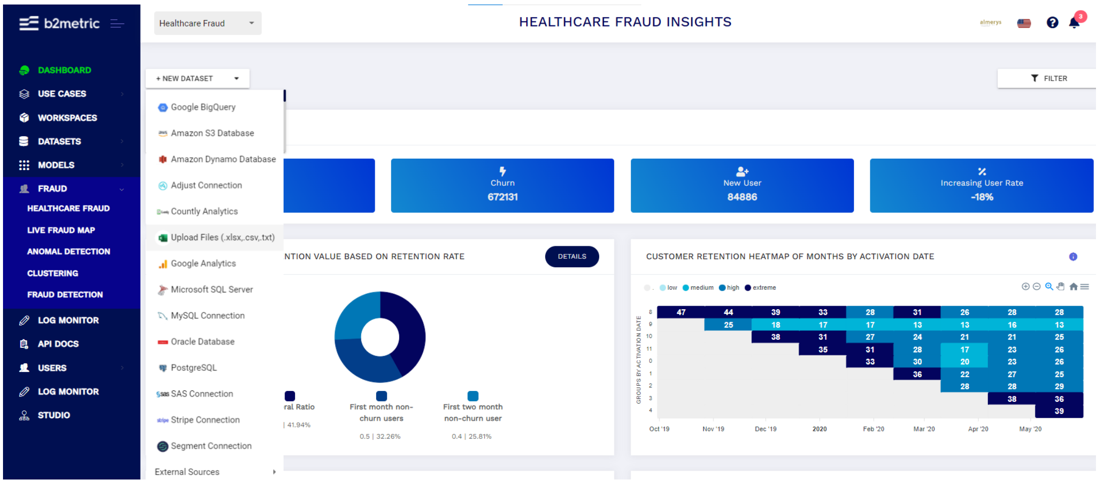
For Clustering, we ran KMeans, Agglomerative and DBSCAN algorithms on B2Metric ML Studio platform.

Isolation Forest, One Class SVM and Local Outlier Factor (LOF) algorithms also run for Anomaly Detection model.

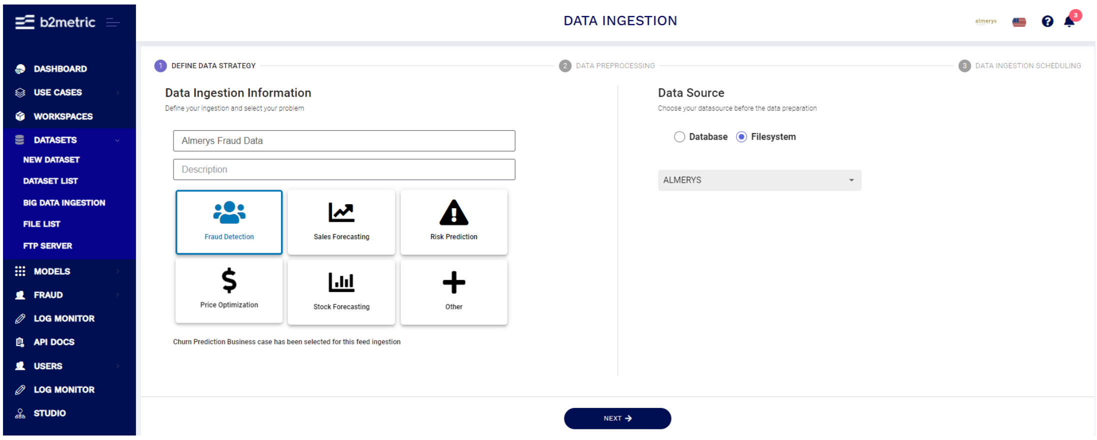
You can access B2Metric ML Studio Fraud Model Explainable AI reports and other reports in platform via these informations below:

B2ML Studio Url:	https://app.b2metric.com/
Username:	almerys@b2metric.com
Password:	Almerys1234!!
How B2Metric AutoML Solution Works?	https://www.youtube.com/watch?v=wPxOxb7-xtg
B2Metric ML Studio Documentation	https://drive.google.com/drive/u/0/folders/1xC_WTWHC64L8UPJbSz8hgVYhWLcYdhAy

DATA INGESTION



- The platform allows you to upload data to be modelled from your database like Amazon DynamoDB, Oracle, PostgreSQL, MsSQL, MySQL, Stripe, Google Analytics, Google Big Query, SAS or from your computer as a CSV, XLS(X), TXT, PARQUET file.



Data Ingestion Information
Define your ingestion and select your problem

Almerys Fraud Data

Description

Churn Prediction Business case has been selected for this feed ingestion

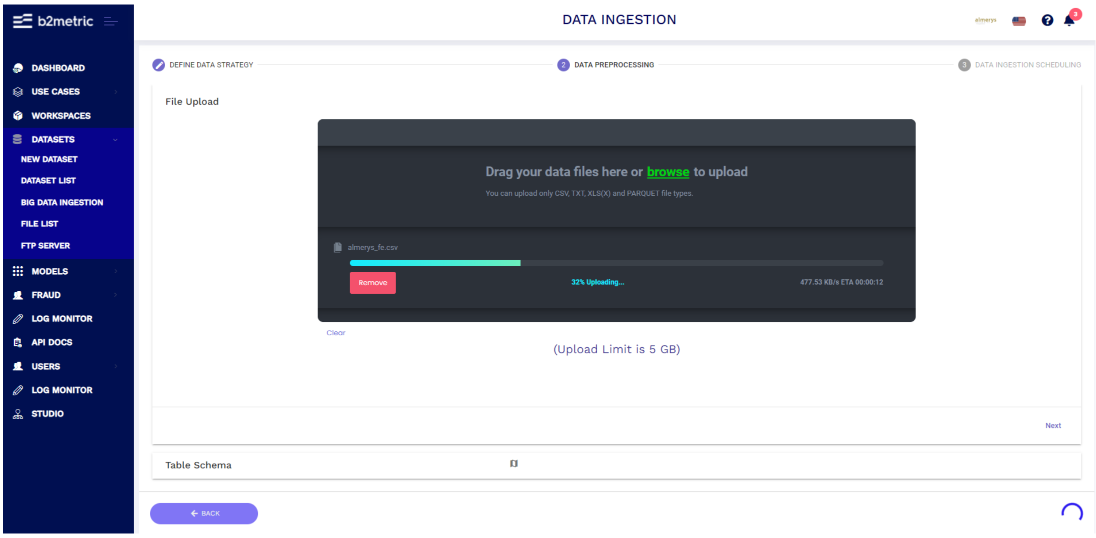
Data Source
Choose your datasource before the data preparation

○ Database ● Filesystem

ALMERYS

- The first step for data ingestion, named the data and select data sources like file upload from computer or file upload from DB connection.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981



DATA INGESTION

File Upload

Drag your data files here or browse to upload

You can upload only CSV, TXT, XLS(X) and PARQUET file types.

almerys_fe.csv

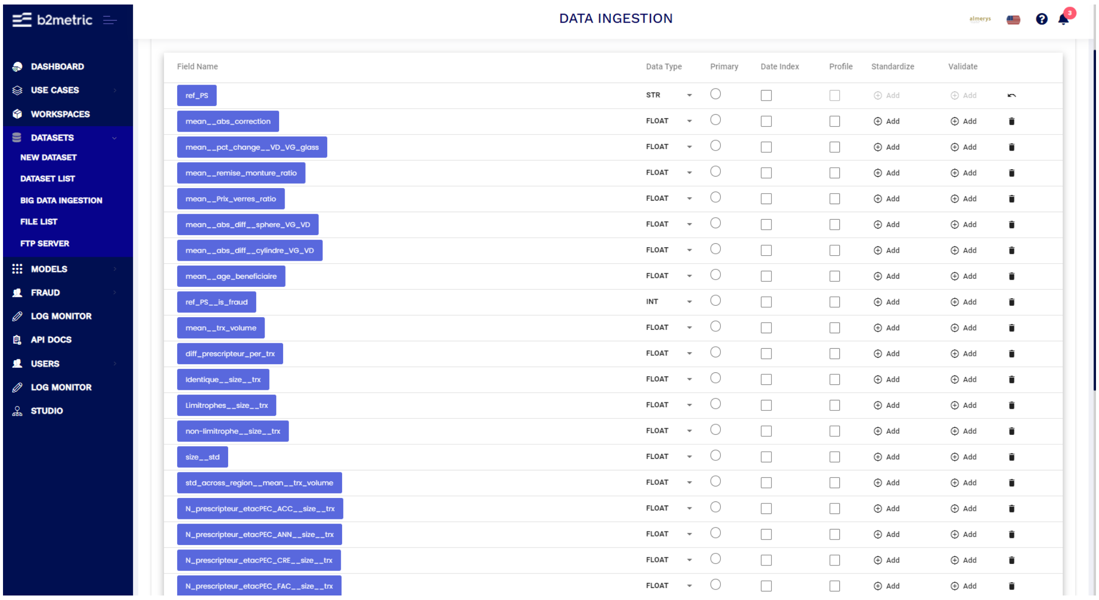
32% Uploading...

477.53 KB/s ETA 00:00:12

(Upload Limit is 5 GB)

Table Schema

- Then select and upload Almerys data file from the computer and wait until the file upload is done.

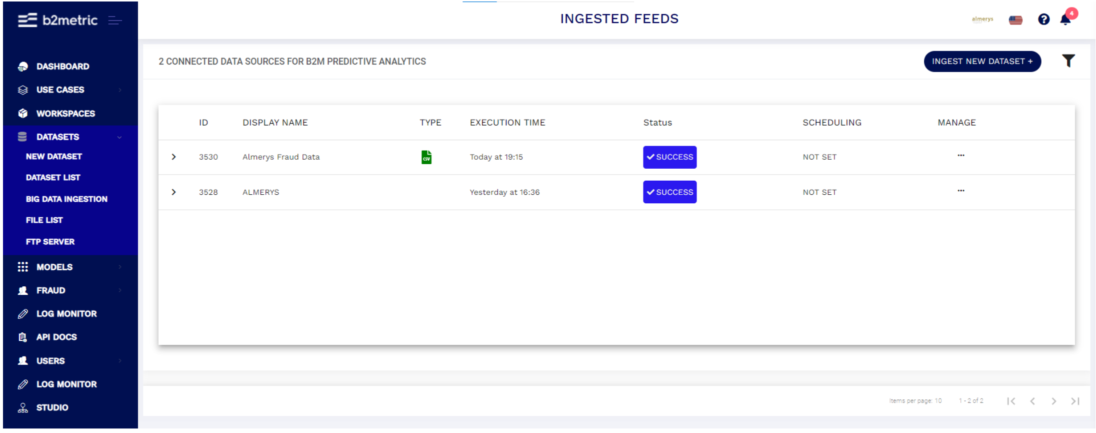


DATA INGESTION

Field Name	Data Type	Primary	Date Index	Profile	Standardize	Validate	
ref_PS	STR	○	□	□	⊕ Add	⊕ Add	↶
mean__abs_correction	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
mean__pct_change__VD_VG_glass	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
mean__remise_monture_ratio	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
mean__Prix_verres_ratio	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
mean__abs_diff__sphere_VG_VD	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
mean__abs_diff__cylindre_VG_VD	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
mean__age_beneficiaire	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
ref_PS__is_fraud	INT	○	□	□	⊕ Add	⊕ Add	🗑
mean__trx_volume	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
diff_prescripteur_per_trx	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
Identique__size__trx	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
Limitrophes__size__trx	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
non-limitrophe__size__trx	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
size__std	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
std_across_region__mean__trx_volume	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
N_prescripteur_etacPEC_ACC__size__trx	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
N_prescripteur_etacPEC_ANN__size__trx	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
N_prescripteur_etacPEC_CRE__size__trx	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑
N_prescripteur_etacPEC_FAC__size__trx	FLOAT	○	□	□	⊕ Add	⊕ Add	🗑

- In B2Metric AutoML Platform Data Ingestor Module, you can do some basic data preprocessing steps like deleting some columns, validating and standardising or labelling a column as primary and date index.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981



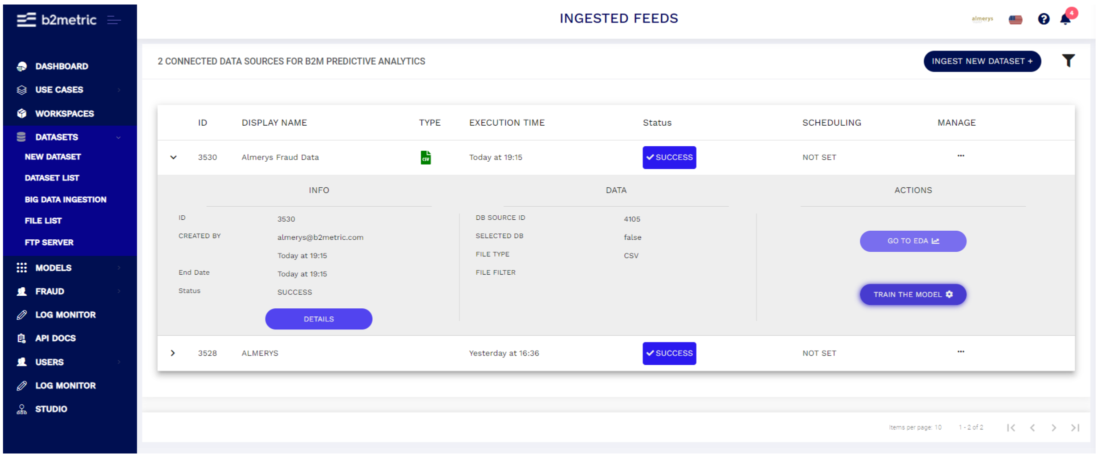
INGESTED FEEDS

2 CONNECTED DATA SOURCES FOR B2M PREDICTIVE ANALYTICS

	ID	DISPLAY NAME	TYPE	EXECUTION TIME	Status	SCHEDULING	MANAGE
›	3530	Almerys Fraud Data	csv	Today at 19:15	✓ SUCCESS	NOT SET	···
›	3528	ALMERYS		Yesterday at 16:36	✓ SUCCESS	NOT SET	···

- Uploaded Almerys data are listed in Ingested Feeds.

MODELLING



INGESTED FEEDS

2 CONNECTED DATA SOURCES FOR B2M PREDICTIVE ANALYTICS

	ID	DISPLAY NAME	TYPE	EXECUTION TIME	Status	SCHEDULING	MANAGE
⌄	3530	Almerys Fraud Data	csv	Today at 19:15	✓ SUCCESS	NOT SET	···
›	3528	ALMERYS		Yesterday at 16:36	✓ SUCCESS	NOT SET	···

INFO

ID	3530
CREATED BY	almerys@b2metric.com
	Today at 19:15
End Date	Today at 19:15
Status	SUCCESS

DATA

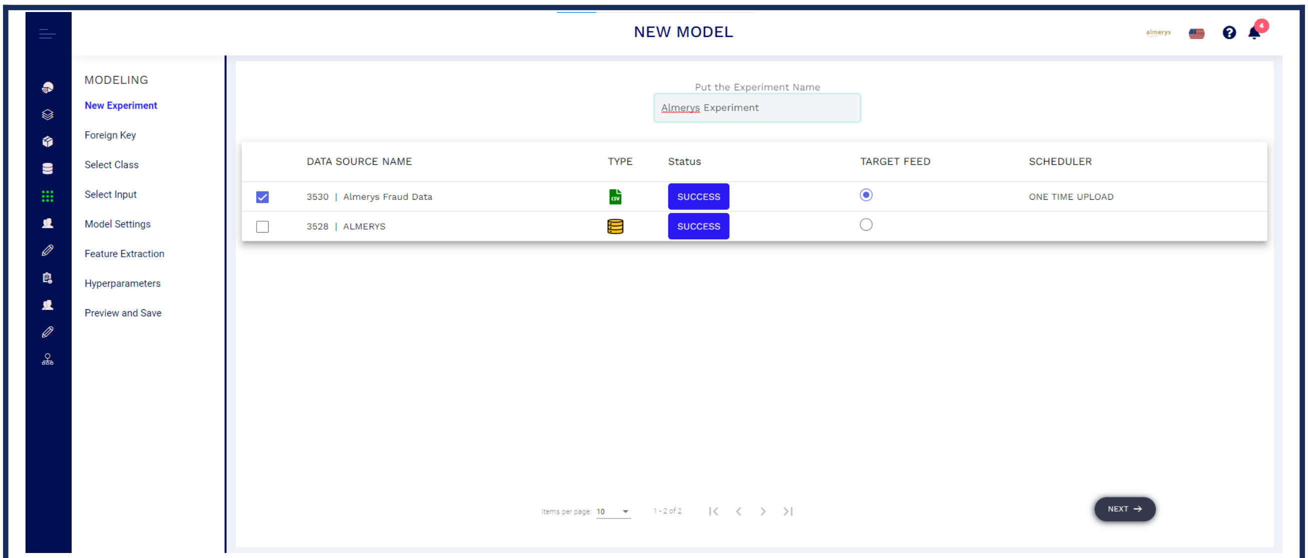
DB SOURCE ID	4105
SELECTED DB	false
FILE TYPE	CSV
FILE FILTER	

ACTIONS

- You can run the supervised based classification or regression models, unsupervised based clustering models and anomaly detection models on the platform.

  From here, the screenshots given from the platform belong to the creation stages of the fraud prediction model, which is modelled as binary classification (supervised).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

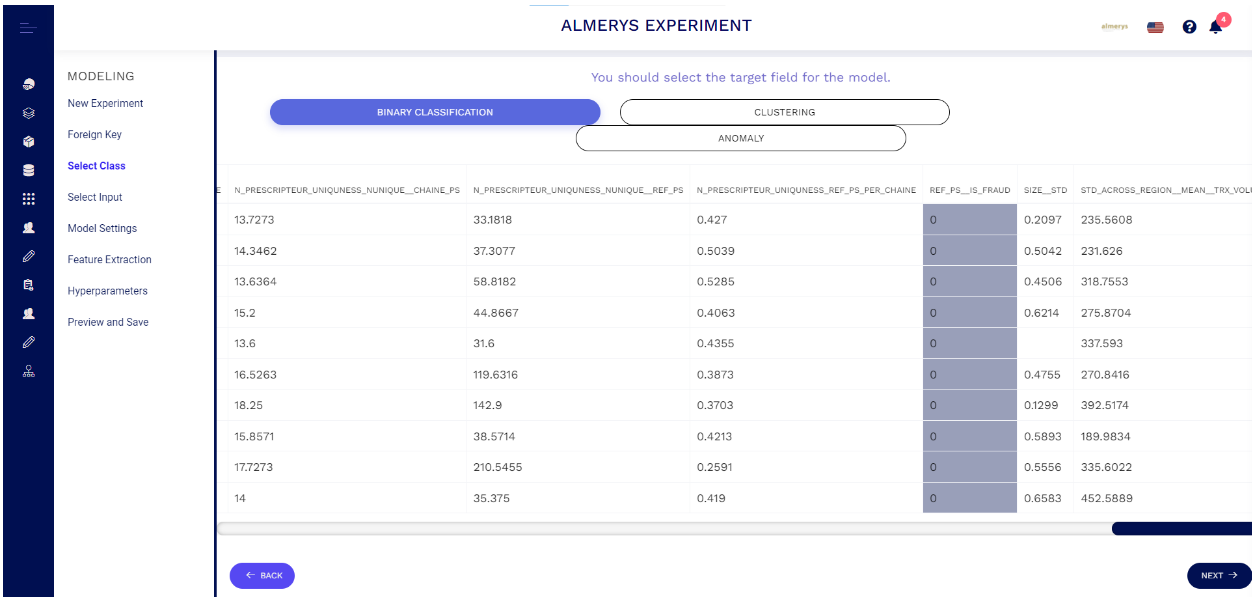


Put the Experiment Name

Almerys Experiment

	DATA SOURCE NAME	TYPE	Status	TARGET FEED	SCHEDULER
☑	3530 | Almerys Fraud Data	csv	SUCCESS	●	ONE TIME UPLOAD
□	3528 | ALMERYS		SUCCESS	○	

- The modelling phase begins by naming the classification model to be modelled on Almerys data.



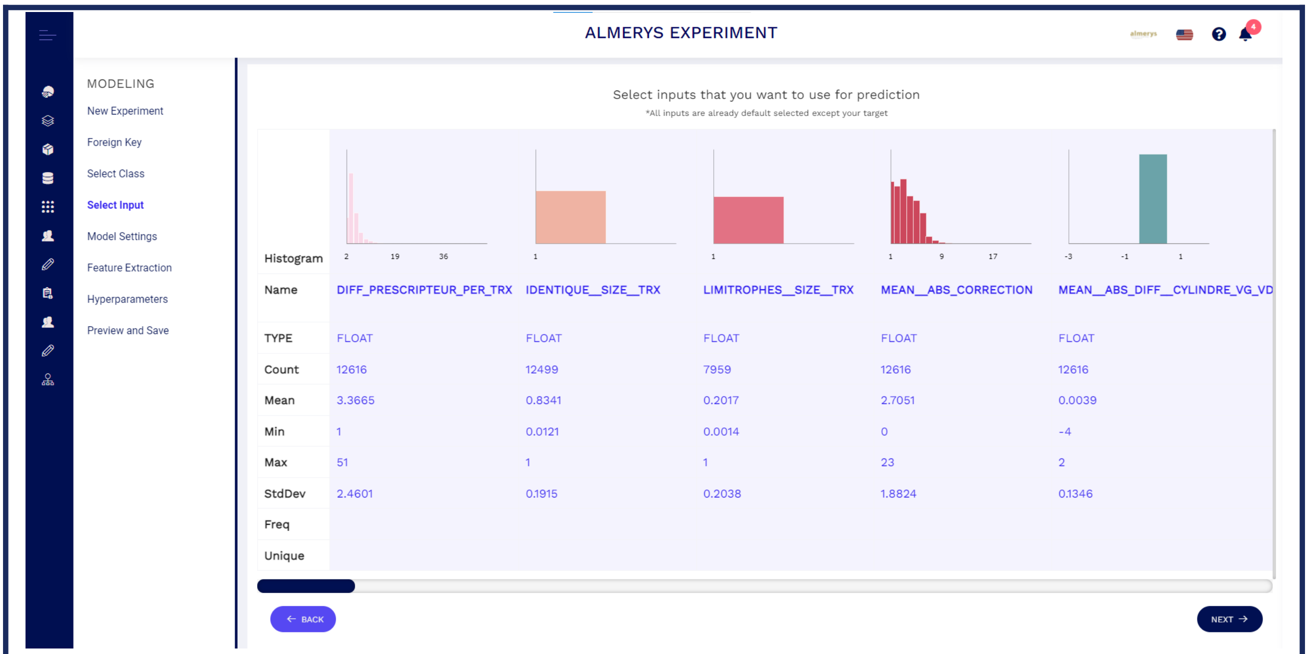
You should select the target field for the model.

N_PRESCRIPTEUR_UNIQUNESS_NUNIQUE__CHAINE_PS	N_PRESCRIPTEUR_UNIQUNESS_NUNIQUE__REF_PS	N_PRESCRIPTEUR_UNIQUNESS_REF_PS_PER_CHAINE	REF_PS__IS_FRAUD	SIZE__STD	STD_ACROSS_REGION__MEAN__TRX_VOLI
13.7273	33.1818	0.427	0	0.2097	235.5608
14.3462	37.3077	0.5039	0	0.5042	231.626
13.6364	58.8182	0.5285	0	0.4506	318.7553
15.2	44.8667	0.4063	0	0.6214	275.8704
13.6	31.6	0.4355	0		337.593
16.5263	119.6316	0.3873	0	0.4755	270.8416
18.25	142.9	0.3703	0	0.1299	392.5174
15.8571	38.5714	0.4213	0	0.5893	189.9834
17.7273	210.5455	0.2591	0	0.5556	335.6022
14	35.375	0.419	0	0.6583	452.5889

- In the next step, target selection is made with the model, which is desired to be estimated(REF_PS_IS_FRAUD)

.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

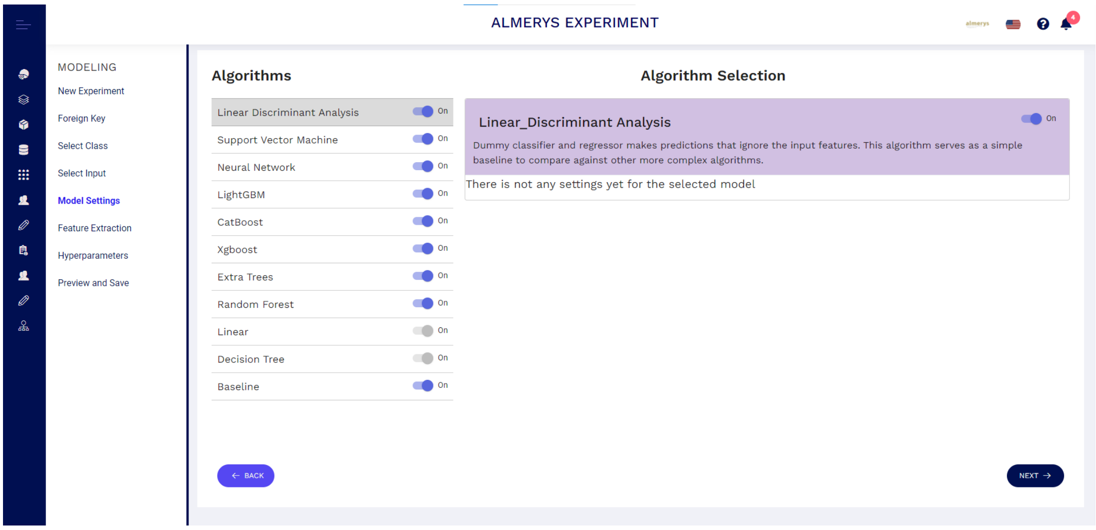


Select inputs that you want to use for prediction

*All inputs are already default selected except your target

Name	DIFF_PRESCRIPTEUR_PER_TRX	IDENTIQUE__SIZE__TRX	LIMITROPHES__SIZE__TRX	MEAN__ABS_CORRECTION	MEAN__ABS_DIFF__CYLINDRE_VG_VD
TYPE	FLOAT	FLOAT	FLOAT	FLOAT	FLOAT
Count	12616	12499	7959	12616	12616
Mean	3.3665	0.8341	0.2017	2.7051	0.0039
Min	1	0.0121	0.0014	0	-4
Max	51	1	1	23	2
StdDev	2.4601	0.1915	0.2038	1.8824	0.1346
Freq					
Unique					

- The variables to be entered into the model as input from the Almerys data are selected at this stage. All inputs except the target are already default selected. At this stage, you can exclude the unwanted variables from the model by deselecting them.



Algorithms

Linear Discriminant Analysis	● On
Support Vector Machine	● On
Neural Network	● On
LightGBM	● On
CatBoost	● On
Xgboost	● On
Extra Trees	● On
Random Forest	● On
Linear	○ On
Decision Tree	○ On
Baseline	● On

Algorithm Selection

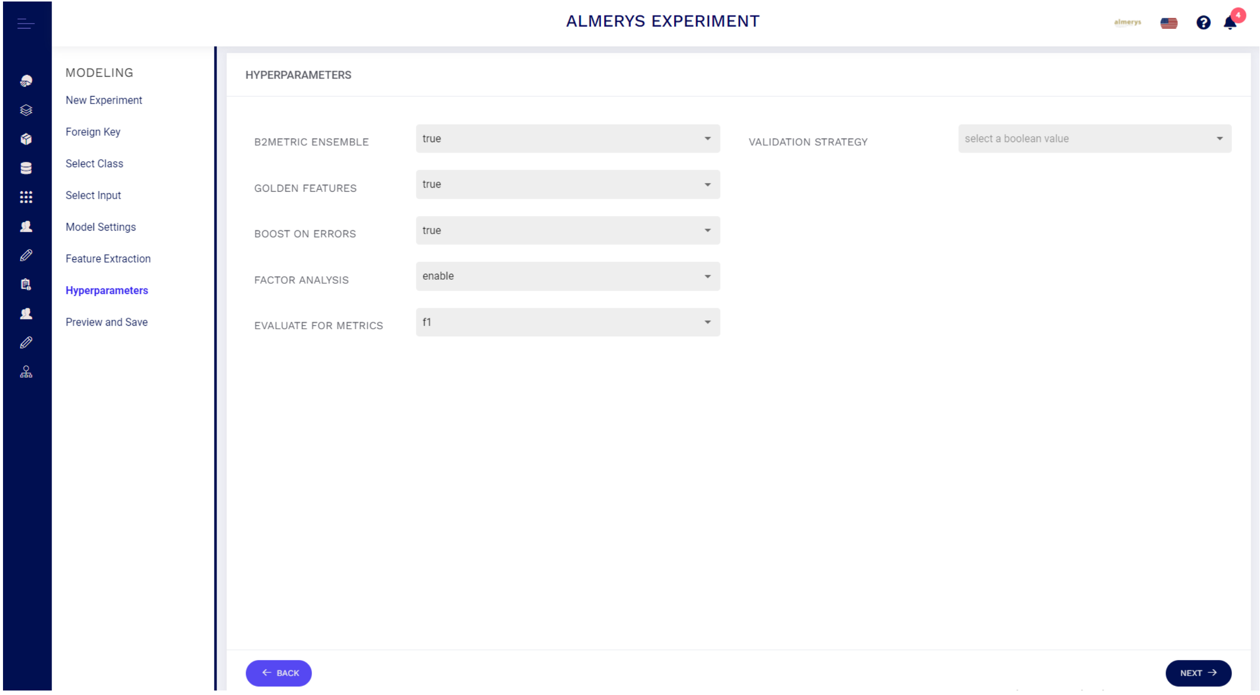
Linear_Discriminant Analysis ● On

Dummy classifier and regressor makes predictions that ignore the input features. This algorithm serves as a simple baseline to compare against other more complex algorithms.

There is not any settings yet for the selected model

- The algorithms to be run for the Almerys model are selected at this stage.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981



HYPERPARAMETERS

B2METRIC ENSEMBLE: true

GOLDEN FEATURES: true

BOOST ON ERRORS: true

FACTOR ANALYSIS: enable

EVALUATE FOR METRICS: f1

VALIDATION STRATEGY: select a boolean value

- You can do some parameter optimization processes in this step

The B2M Ensemble model chooses the best result from the run algorithms and tries to improve the model result with various parameter optimizations. For this reason, the B2M Ensemble model do the best classification among the data features.

Factor analysis can reveal which qualities, characteristics, or priorities are most essential to a specific group of clients (group)

With the golden features, the AutoML infrastructure tries to create new variables that will improve the model results by crossing the variables in the background. You can think of feature creation as a manual, golden feature created with the automatic ml method.

Shapley analysis is an analysis method we use. You can enable or disable these features optionally. You can see them later on features.

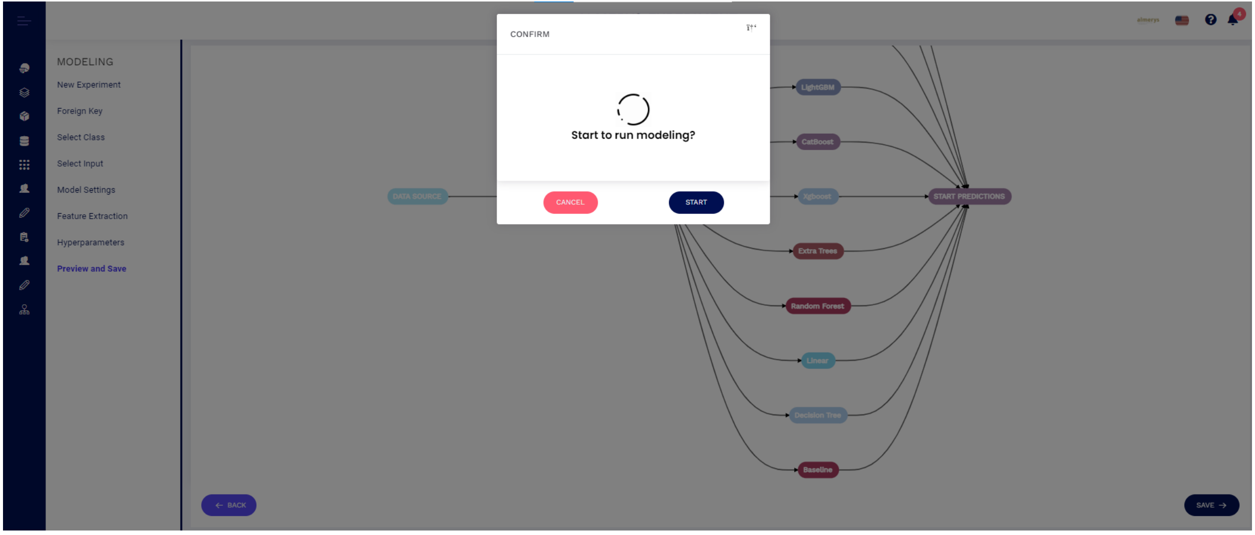
You can create your model based on this metric by selecting the model metric you want to optimise. Metrics could be accuracy, f1, log loss for classification models or RMSA, r2(r-squared), explained variance score for regression models

In the validation strategy; there are 2 options kfold and split. In the split option, your data split into 2 sets: a test and a train set. You can choose a split ratio for these test and train sets.

In the k-fold option, the data is divided into the number of folds you choose, and each fold is trained and tested separately, then the average is taken for each fold.

If no selection is made for the validation strategy, by default, 25% of the data is used for testing and 75% for training.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

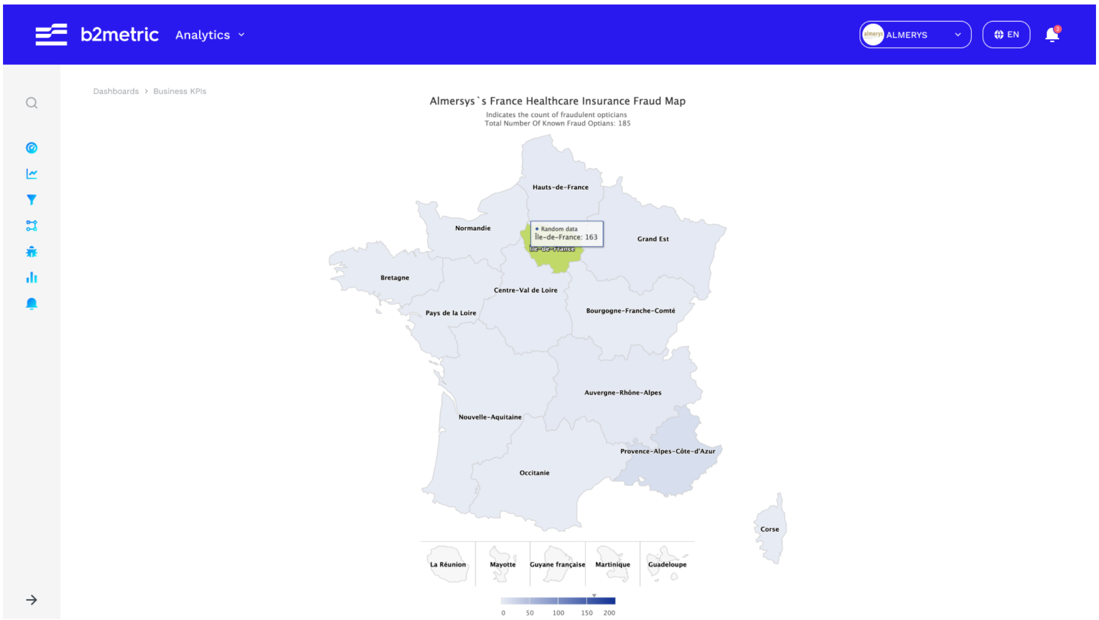


CONFIRM

Start to run modeling?

- At the last stage, you can start to run the Fraud Prediction model from the labelled Almerys data with the algorithms that you choose by Start to Run Modelling.

1- Platform starts with the current fraud frequency map. For instance, this map shows that the actual opticians fraud count is 163 in Ile-de-France.



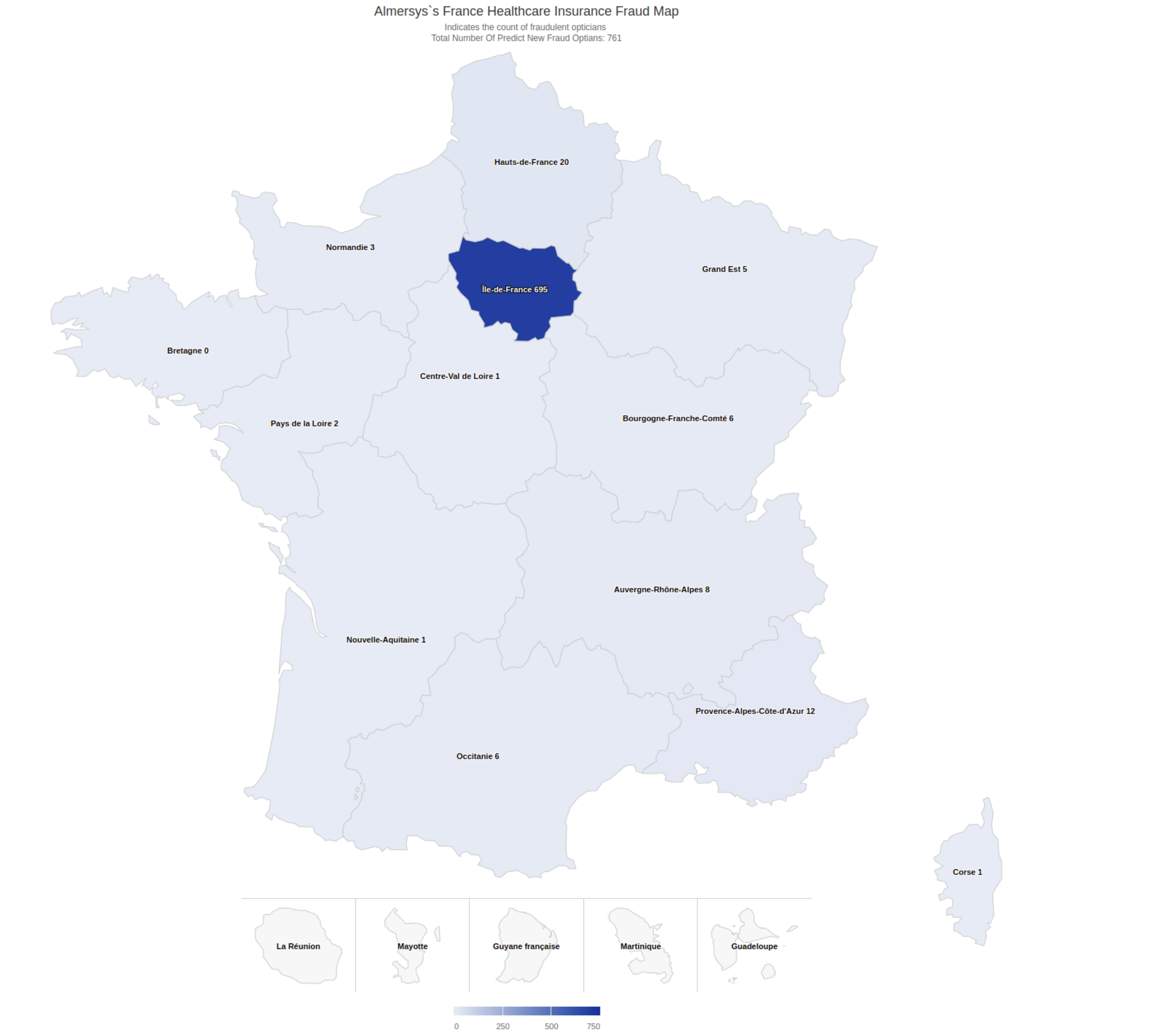
Almersys`s France Healthcare Insurance Fraud Map

Indicates the count of fraudulent opticians
Total Number Of Known Fraud Optians: 185

Random data
Île-de-France: 163

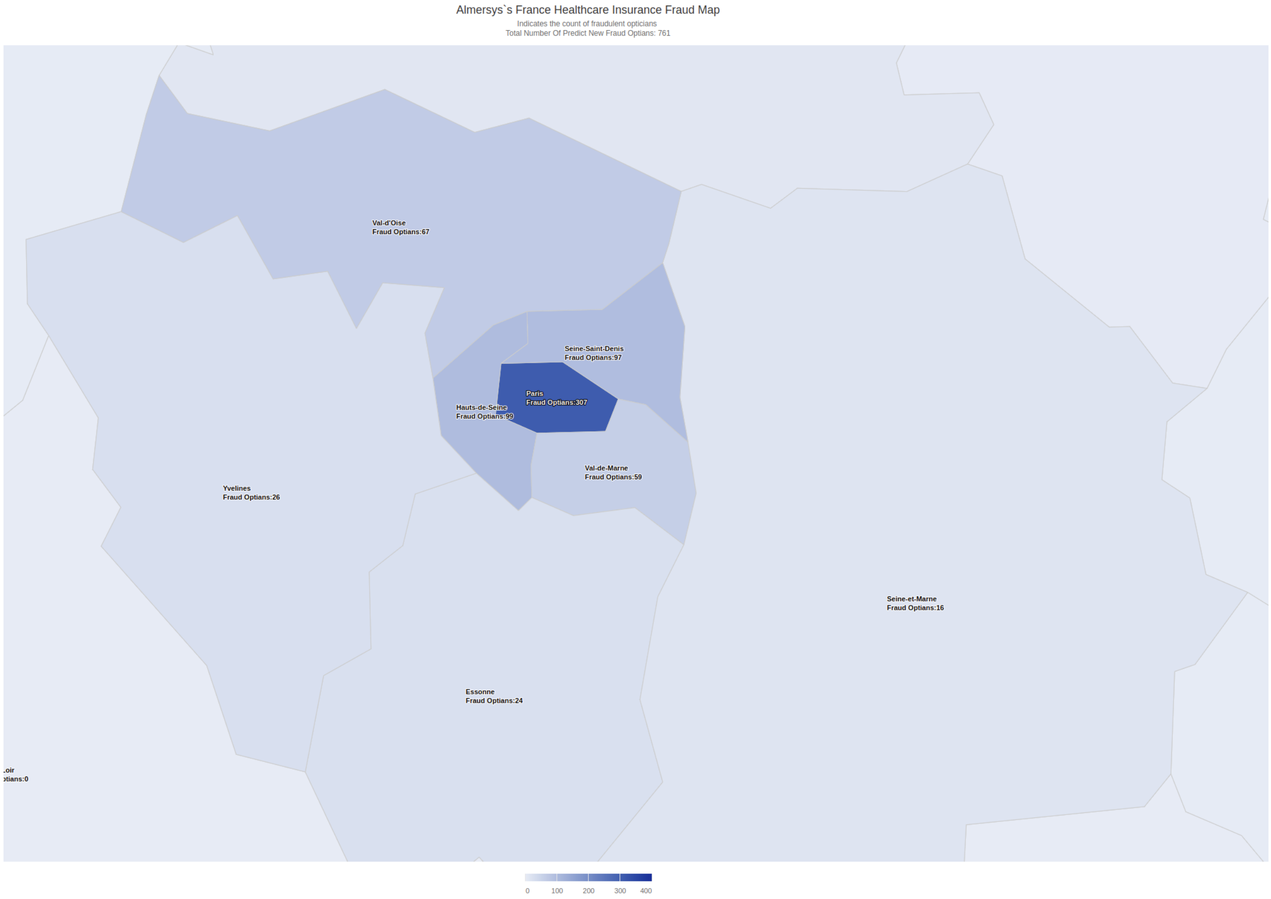
2- Then B2ML Hunter's Fraud model prediction indicates below. B2Metric's fraud model finds 557 new suspicious opticians in France.

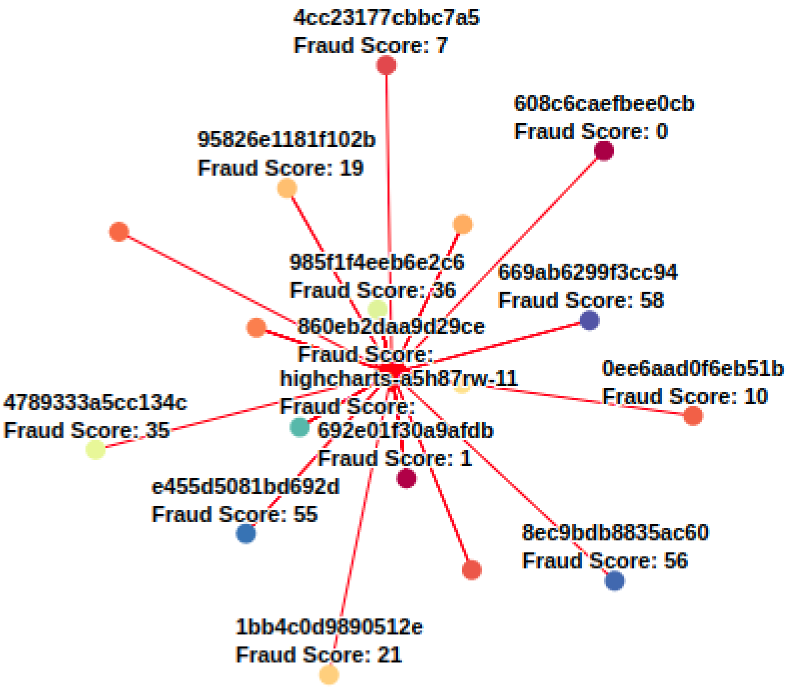
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981



3- Also the map below indicates the whole cities of France. Then B2ML Hunter’s Fraud model prediction indicates below. B2Metric’s fraud model finds 761 new suspicious opticians in France.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 951981





This graph network indicates one optician that B2Metric system detect it as a fraud. It shows suspicious transactions within all this optician’s transactions. For instance, this graph network shows the detected fraud on the **optician id 08960ca0a741f5e an he has 385** different transactions.

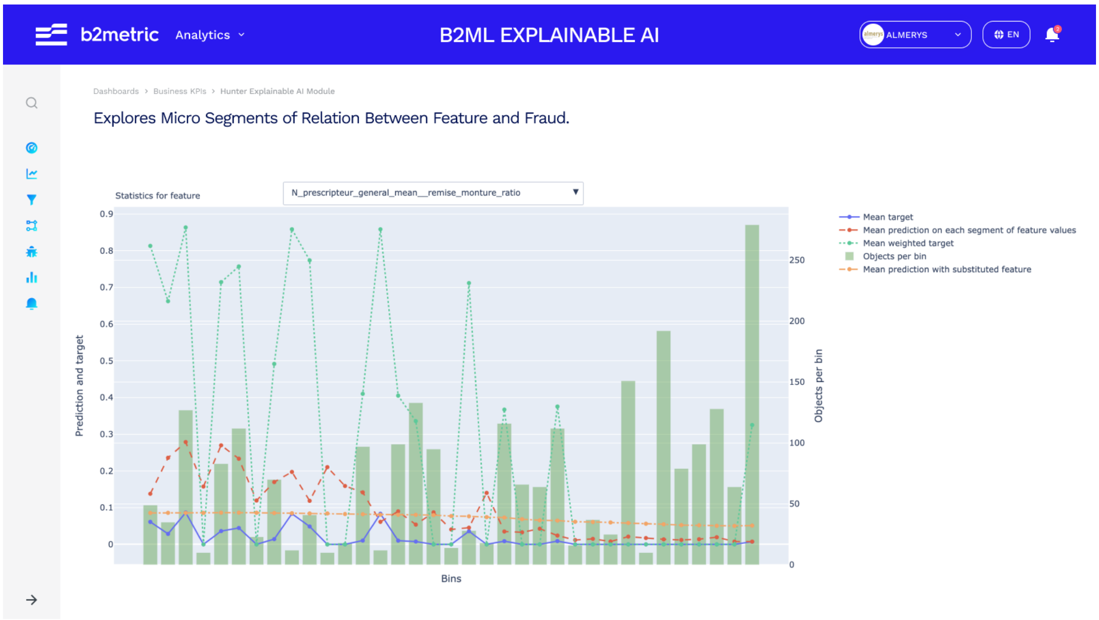
This project has received funding from the European Union’s Horizon 2020
research and innovation programme under grant agreement No 951981

Here is the list all predicted Fraud optician transaction list dataset:
https://docs.google.com/spreadsheets/d/1-JqEc_NEKHMBgqo6adskEPbali9ehkoX/edit?usp=sharing&ouid=105050475189860102624&rtpof=true&sd=true

Here is the list all predicted Fraud optician aggregated transaction dataset:
https://docs.google.com/spreadsheets/d/1MoyIf97fxhmPd6rvahealOM_8cB7AwzW/edit?usp=sharing&ouid=105050475189860102624&rtpof=true&sd=true

Here is the explanation of all features:
https://docs.google.com/spreadsheets/d/13OpKfwmj31XsLI9OYyqUqRpxQZ0-dZI9KalefgV0TLw/edit#gid=107525781

B2ML Studio Explainable AI indicates the interpretable modelling solution for Almerys's Fraud and Market Insights datasets.



B2ML EXPLAINABLE AI

Dashboards › Business KPIs › Hunter Explainable AI Module

Explores Micro Segments of Relation Between Feature and Fraud.

Statistics for feature: N_prescripteur_general_mean__remise_monture_ratio

Explainable AI: This Dash explains to find micro segments of relation between feature and fraud. User can easily navigate in between the feature list below:

mean__abs_correction, mean__pct_change__VD_VG_glass, mean__remise_monture_ratio, mean__Prix_verres_ratio, mean__abs_diff__sphere_VG_VD, mean__abs_diff__cylindre_VG_VD,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

mean__age_beneficiaire, mean__trx_volume, diff_prescripteur_per_trx, Identique__size__trx, Limitrophes__size__trx, non-limitrophe__size__trx, size__std, std_across_region__mean__trx_volume, N_prescripteur_etacPEC_ACC__size__trx, N_prescripteur_etacPEC_ANN__size__trx, N_prescripteur_etacPEC_CRE__size__trx, N_prescripteur_etacPEC_FAC__size__trx, N_prescripteur_etacPEC_REF__size__trx, N_prescripteur_etacPEC_size__std, N_prescripteur_etacPEC_std_across_region__mean__trx_volume, N_prescripteur_lien_Identique__size__trx, N_prescripteur_lien_Limitrophes__size__trx, N_prescripteur_lien_non-limitrophe__size__trx, N_prescripteur_lien_size__std, N_prescripteur_lien_std_across_region__mean__trx_volume, N_prescripteur_uniquness_nunique__ref_PS, N_prescripteur_uniquness_nunique__chaine_PS, N_prescripteur_uniquness_ref_PS_per_chaine, N_prescripteur_general_sum__Prix_total, N_prescripteur_general_size__trx, N_prescripteur_general_mean__abs_correction, N_prescripteur_general_mean__pct_change__VD_VG_glass, N_prescripteur_general_mean__remise_monture_ratio, N_prescripteur_general_mean__Prix_verres_ratio, N_prescripteur_general_mean__abs_diff__sphere_VG_VD, N_prescripteur_general_mean__abs_diff__cylindre_VG_VD, N_prescripteur_general_mean__age_beneficiaire, N_prescripteur_general_nunique__N_prescripteur

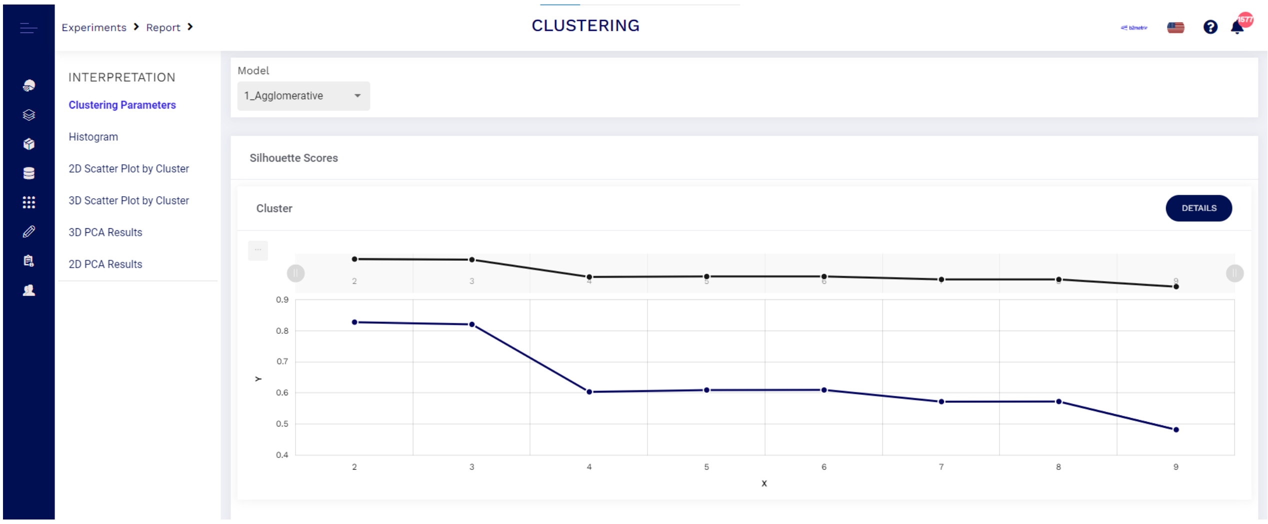
B2Metric AutoML Hunter Clustering Module

B2Metric AutoML Hunter Almerys Fraud Detection Explainable AI Clustering Module reports are indicated below. There are 3 main clustering algorithms in B2Metric AutoML Hunter Almerys Fraud Detection Clustering Module as DBSCAN, Agglomerative and KMeans.

Almerys data was separated into 2 clusters with Agglomerative algorithm and 4 clusters with KMeans algorithm.

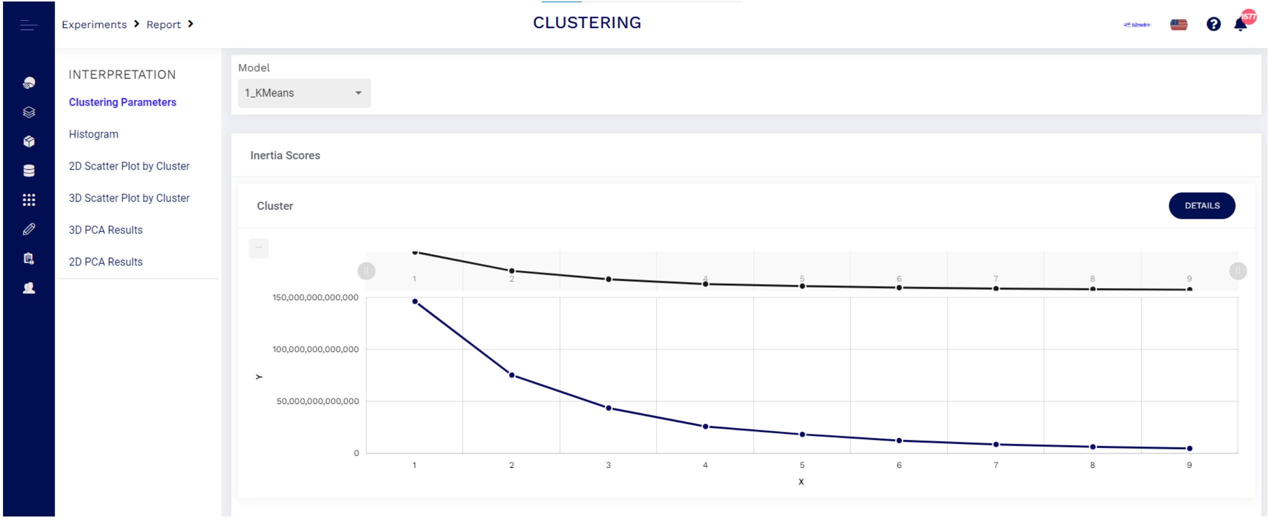
Here you can see the model parameters used for each clustering algorithm run in the B2Metric AutoML Hunter Almerys Fraud Detection - Clustering Module. The screenshots are extracted from B2Metric ML Studio Almersys instance and indicated below.

Clustering Parameters

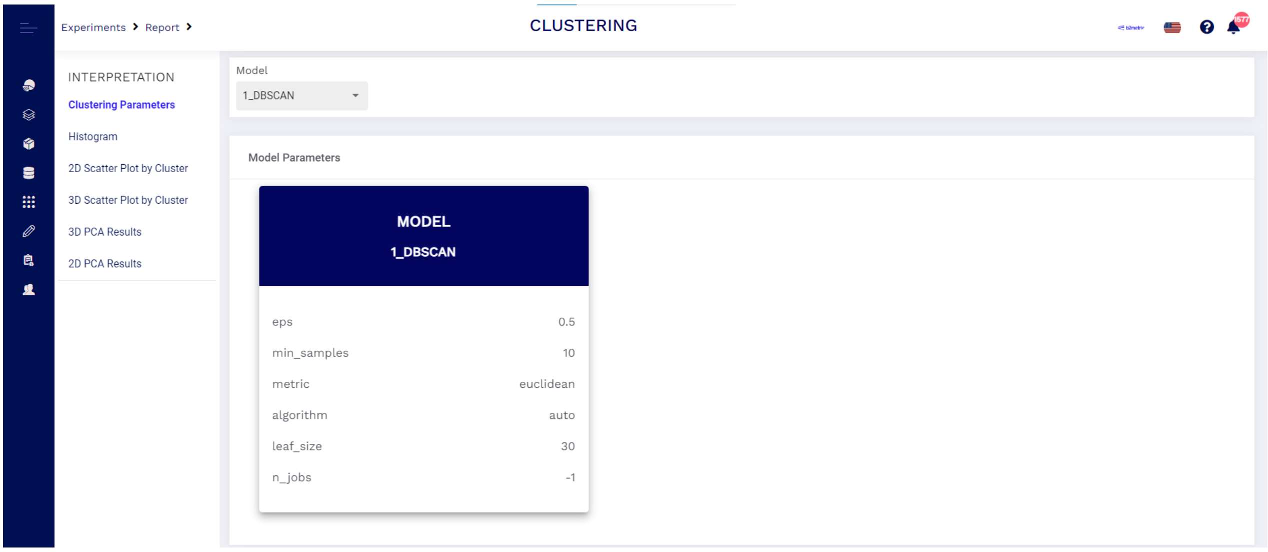


Model: 1_Agglomerative

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981



Model: 1_KMeans



Model: 1_DBSCAN

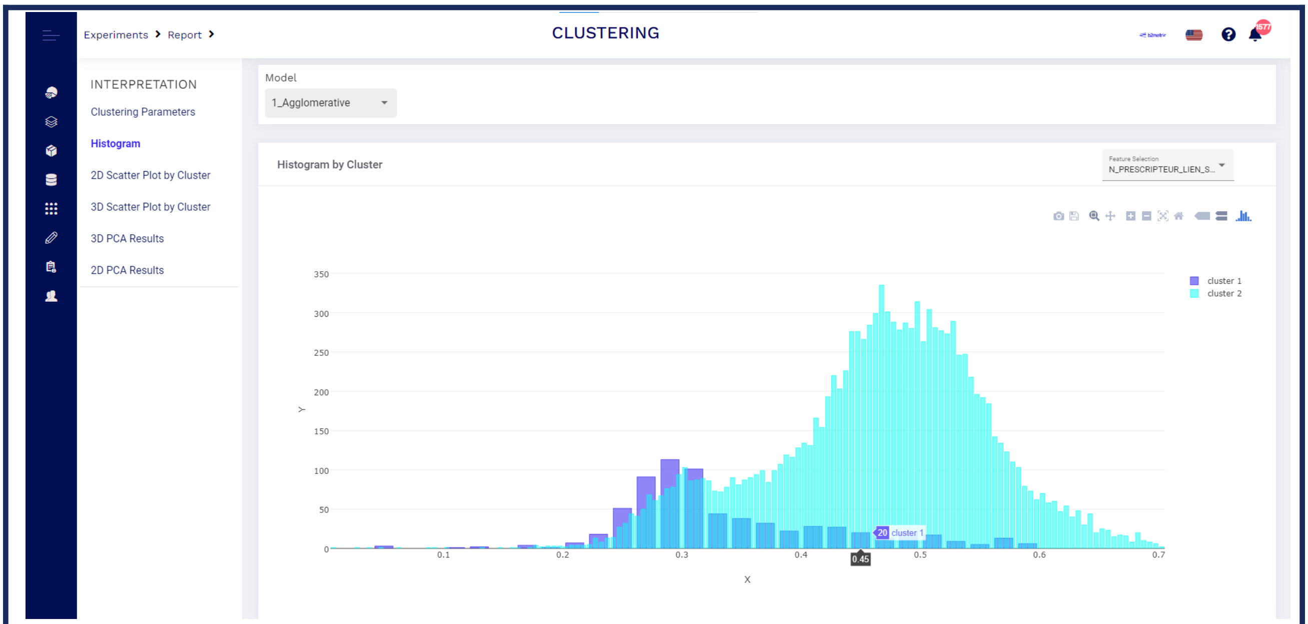
MODEL
1_DBSCAN

eps	0.5
min_samples	10
metric	euclidean
algorithm	auto
leaf_size	30
n_jobs	-1

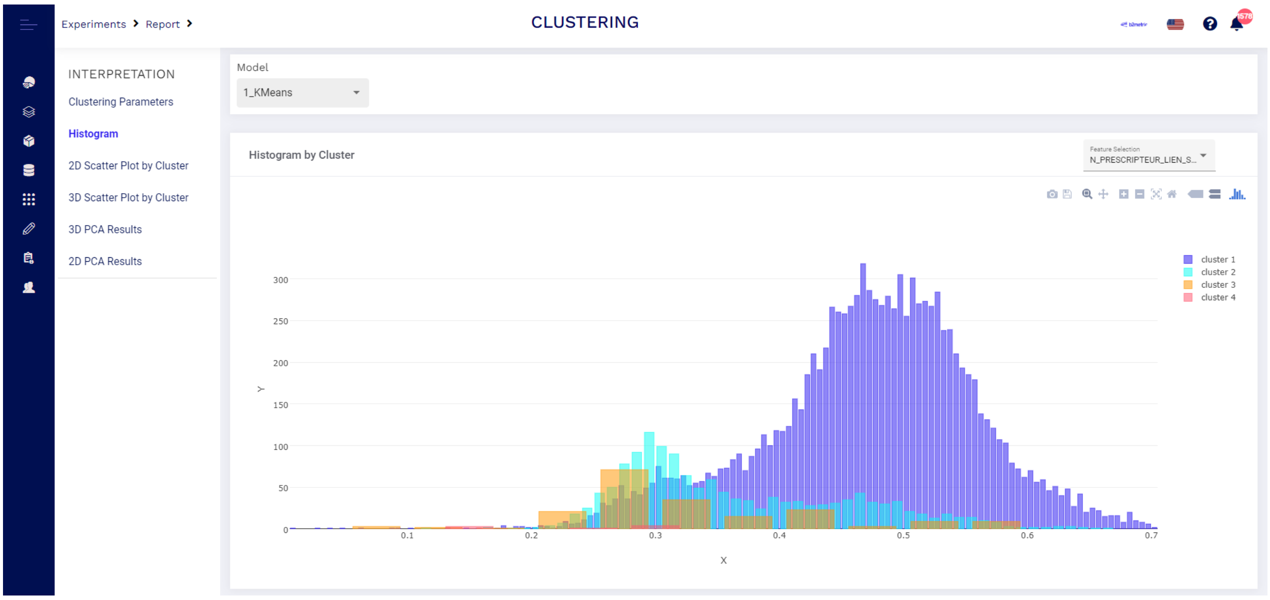
Histogram

This B2ML Studio Explainable AI Feature shows the histogram of features by cluster. You can select and change the feature that you want to see on the histogram.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 951981



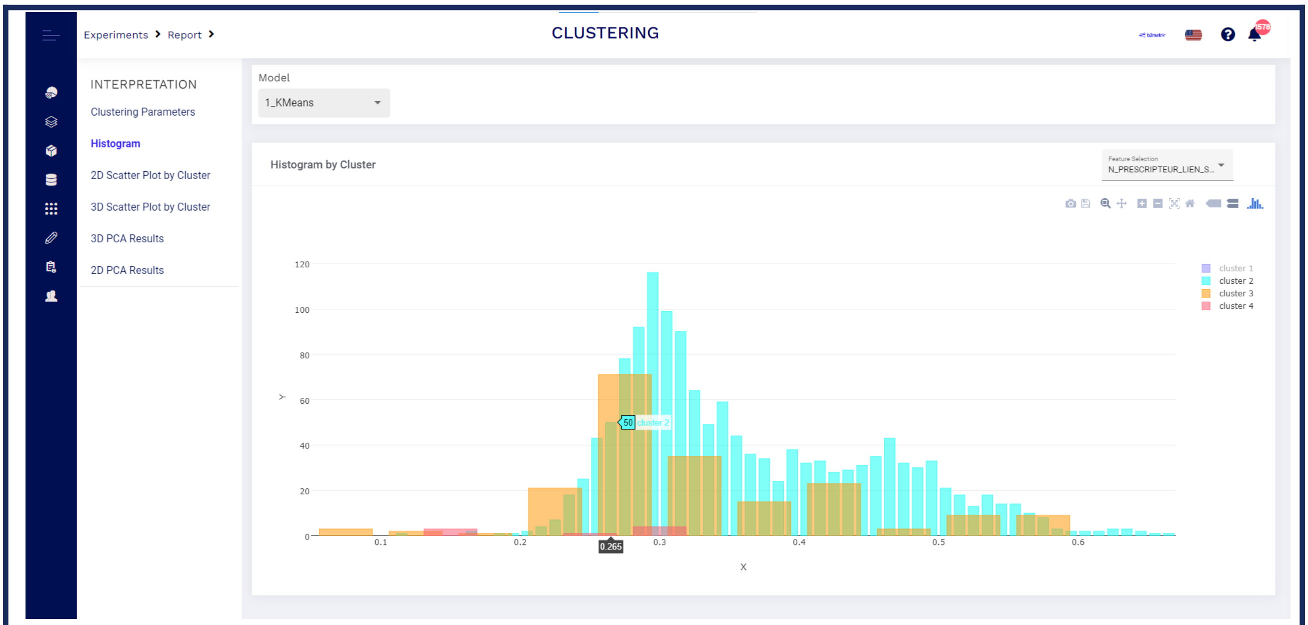
This histogram shows the Agglomerative Clustering algorithm result of Almerys data based on selected features. You can see the histogram results by cluster for each feature.



This histogram represents the distribution of selected features by K-Means clusters.

You can select & deselect a cluster on histogram.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981



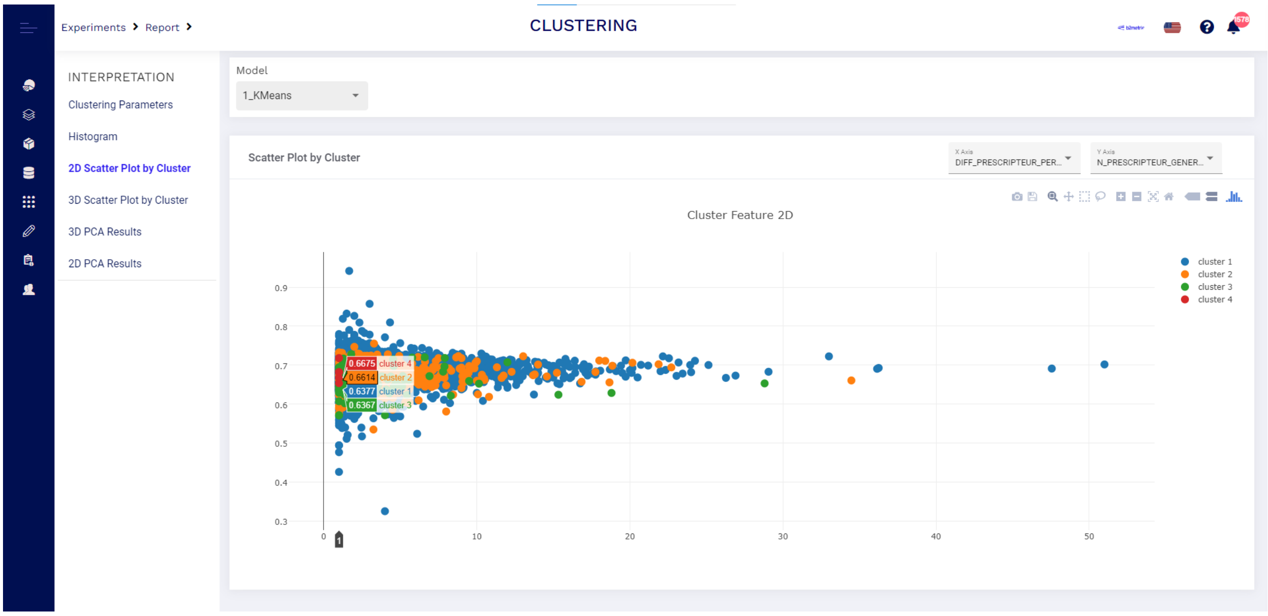
Model
1_KMeans

Feature Selection
N_PRESCRIPTEUR_LIEN_S...

Scatter Plots of Features by Cluster

2D and 3D Scatter Plots for features are reported here. In the graph; you can choose any feature for the dimension axis.

2D Scatter Plot



Model
1_KMeans

X Axis
DIFF_PRESCRIPTEUR_PER...

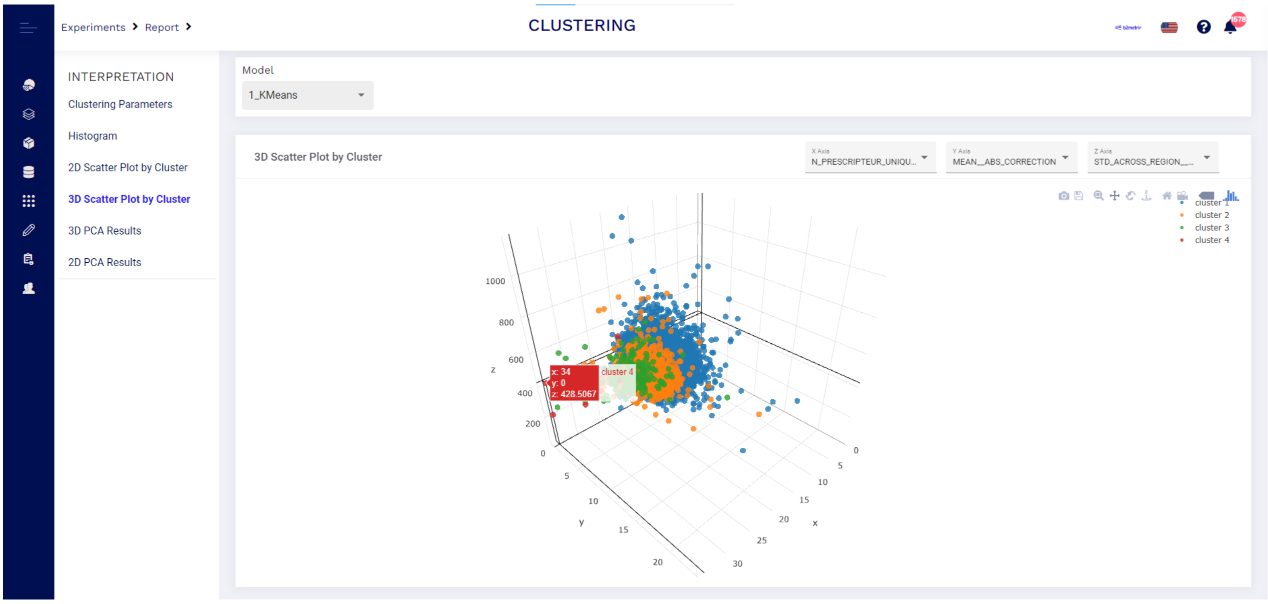
Y Axis
N_PRESCRIPTEUR_GENER...

This graph shows the 2D Scatter plot result of KMeans Clustering algorithm on Almerys data. Almerys data

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 951981

separated into 4 clusters with KMeans algorithm.

3D Scatter Plot



Model
1_KMeans

3D Scatter Plot by Cluster

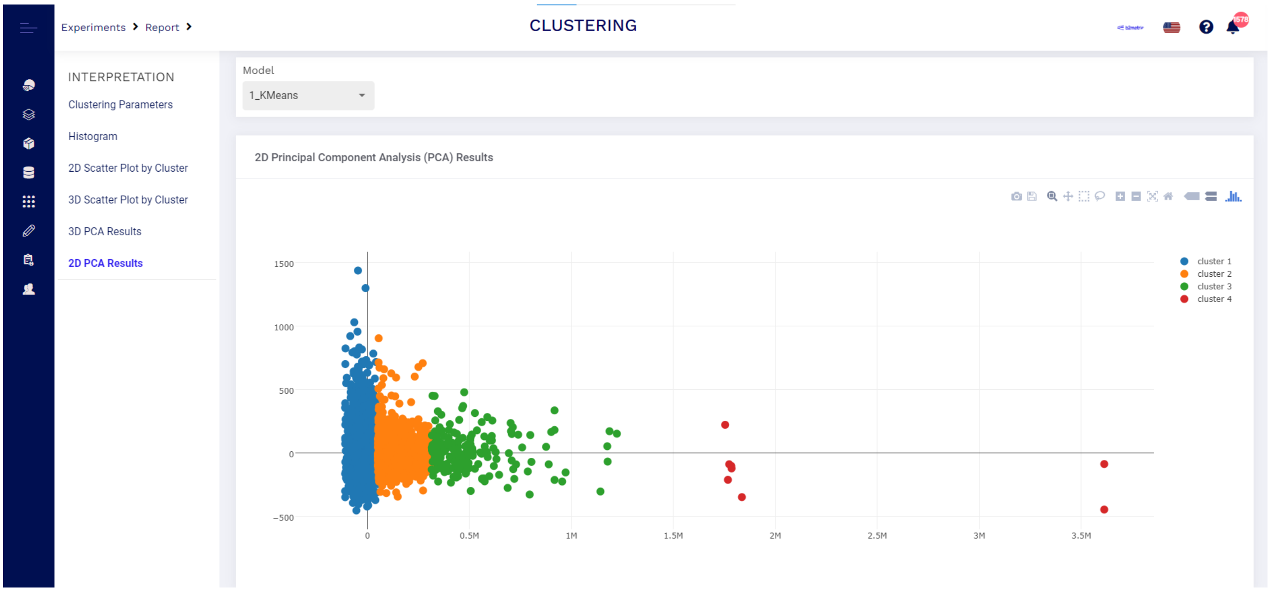
X Axis: N_PRESCRIPTEUR_UNIQU...
Y Axis: MEAN__ABS_CORRECTION
Z Axis: STD_ACROSS_REGION__...

This graph shows the 3D Scatter plot result of KMeans algorithm on Almerys data.

Principal Component Analysis (PCA) Results

You can see the distribution of 2D and 3D reduced versions of Almerys data by Principal Component Analysis by clusters here.

2D PCA Result



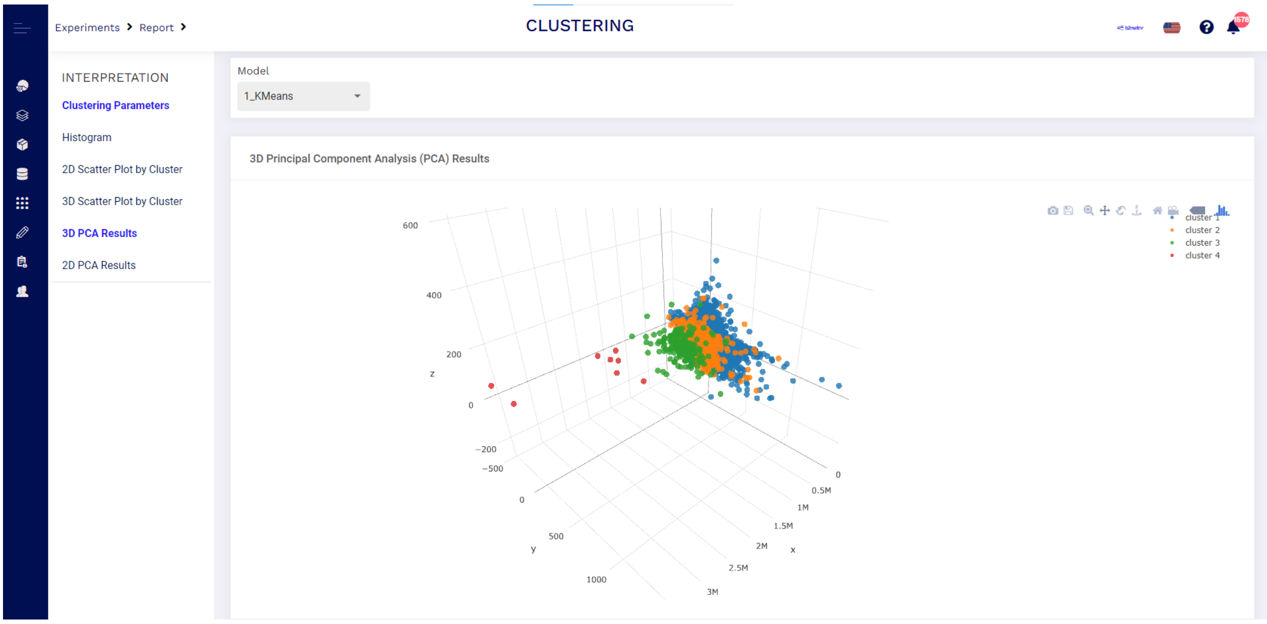
Model
1_KMeans

2D Principal Component Analysis (PCA) Results

This graph shows the 2-Dimensional Principal Component Analysis results of Almerys data with KMeans algorithm.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

3D PCA Result



This graph shows the 3-Dimensional Principal Component Analysis results of Almerys data with KMeans algorithm.

Anomaly Detection B2ML Hunter Module

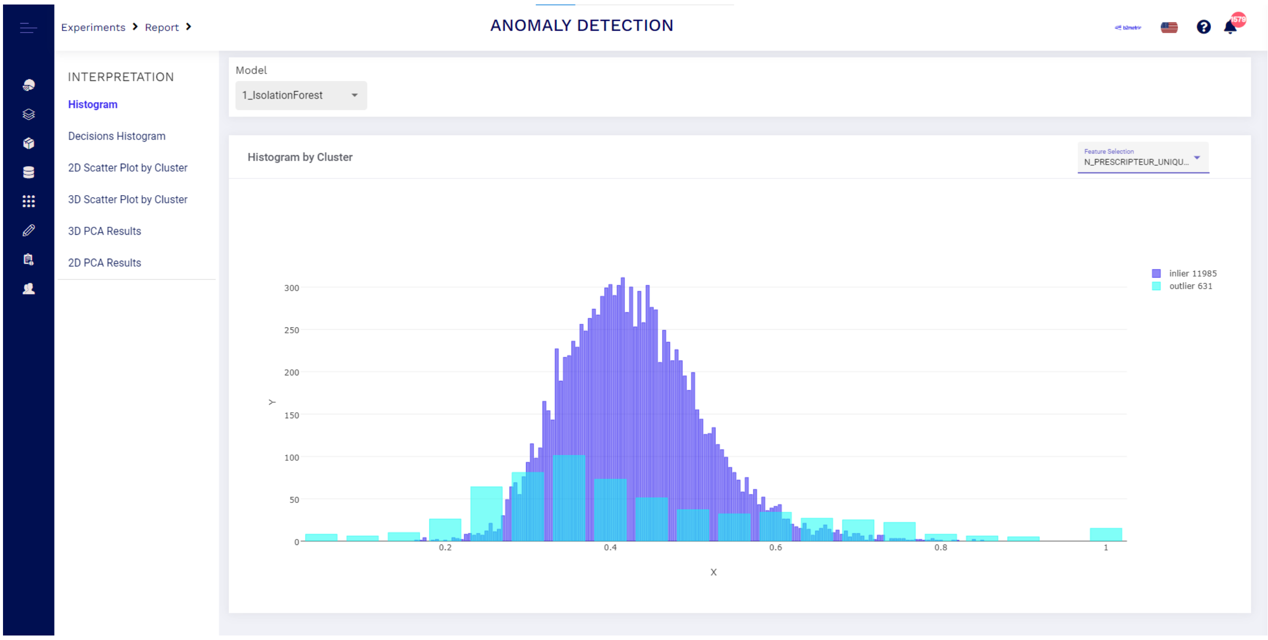
There are 3 main anomaly detection algorithms in B2Metric AutoML Hunter Almerys Anomaly Detection Module as Isolation Forest, One Class SVM and Local Outlier Factor. B2Metric AutoML Hunter Almerys Fraud Detection Explainable AI Anomaly Detection Module reports are indicated below.

B2Metric AutoML Hunter Almerys Anomaly Detection module Isolation Forest algorithm found 11985 inlier and 631 outlier observations on Almerys data.

B2Metric AutoML Hunter Almerys Anomaly Detection module Local Outlier Factor algorithm found 12321 inlier and 295 outlier observations on Almerys data.

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 951981

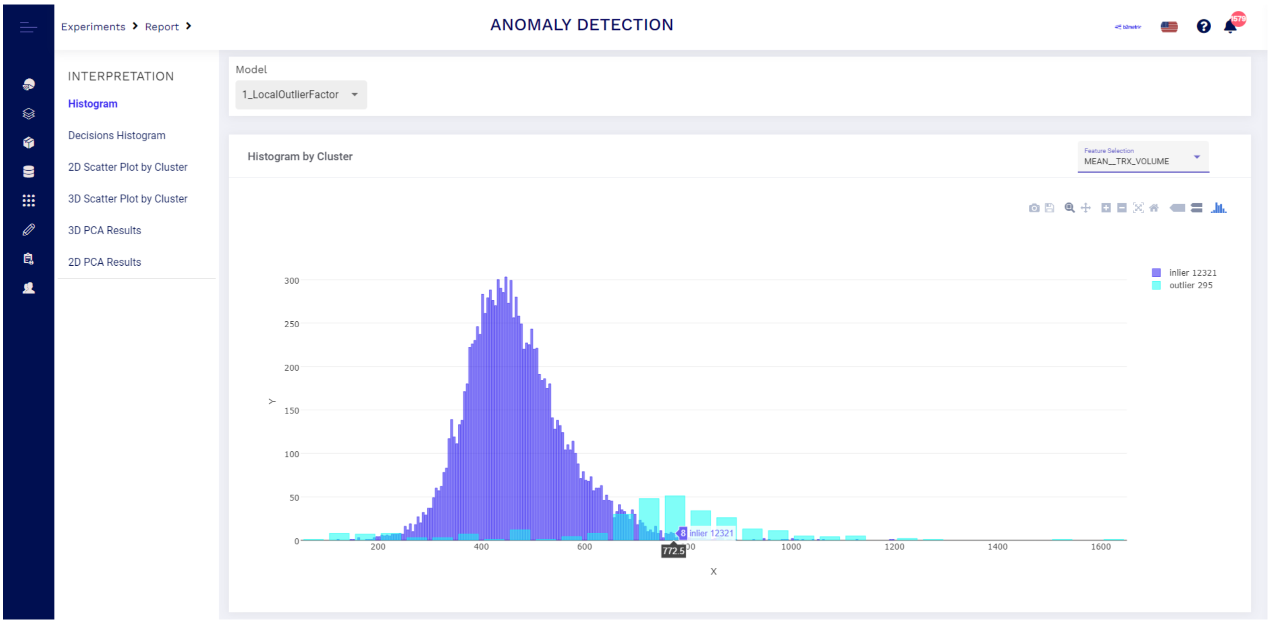
Histogram



Model: 1_IsolationForest

Feature Selection: N_PRESCRIPTEUR_UNIQU...

This histogram shows the Anomaly Detection - Isolation Forest algorithm results on Almerys data for chosen features.



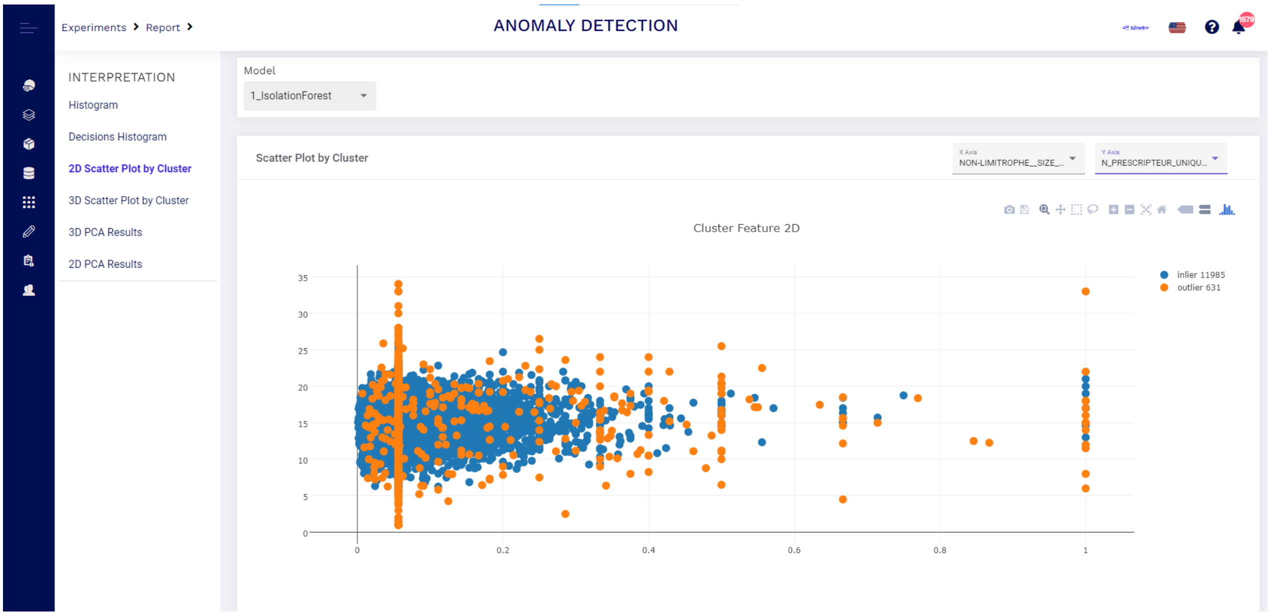
Model: 1_LocalOutlierFactor

Feature Selection: MEAN__TRX_VOLUME

This histogram shows the Anomaly Detection -Local Outlier Factor algorithm results on Almerys data for chosen features.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

2D Scatter Plot Results

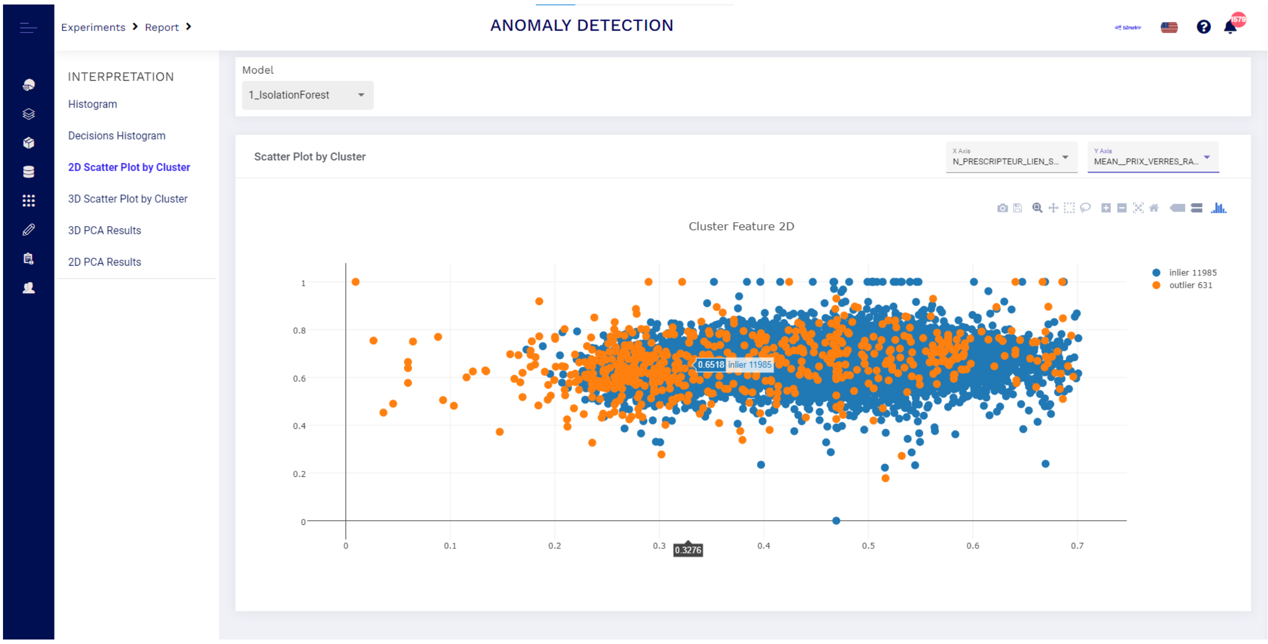


Model: 1_IsolationForest

Scatter Plot by Cluster

X Axis: NON-LIMITROPHE__SIZE_...

Y Axis: N_PRESCRIPTEUR_UNIQU...



Model: 1_IsolationForest

Scatter Plot by Cluster

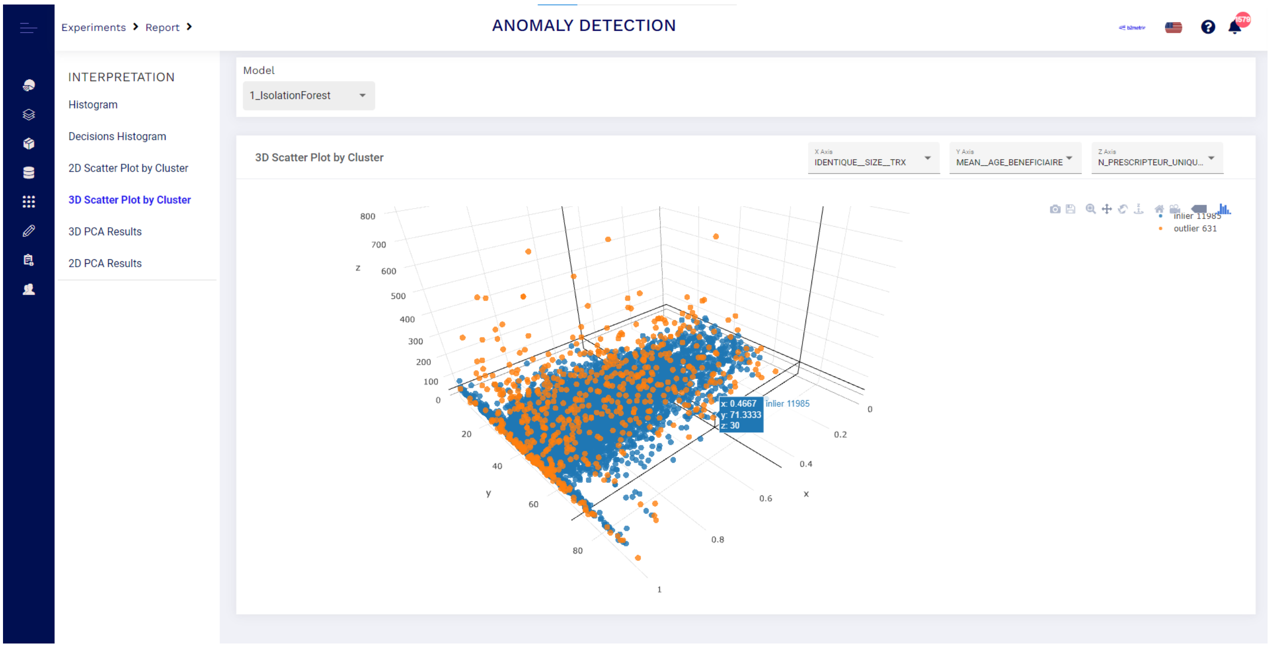
X Axis: N_PRESCRIPTEUR_LIEN_S...

Y Axis: MEAN__PRIX_VERRES_RA...

These scatter plots show the Isolation Forest algorithm results on Almerys Data for features in 2-Dimension. The Isolation Forest algorithm found 631 outliers on Almerys data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

3D Scatter Plot Results



Experiments > Report >

ANOMALY DETECTION

Model: 1_IsolationForest

3D Scatter Plot by Cluster

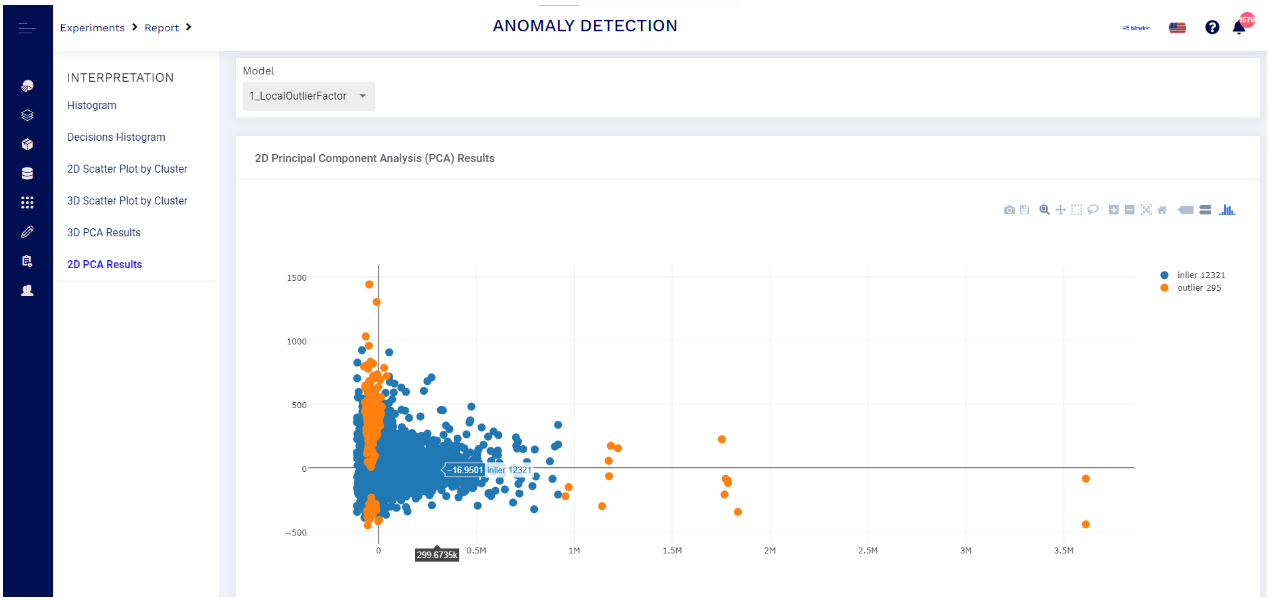
X Axis: IDENTIQUE__SIZE__TRX

Y Axis: MEAN__AGE_BENEFICIAIRE

Z Axis: N_PRESCRIPTEUR_UNIQU...

These scatter plots show the Isolation Forest algorithm results on Almerys Data for features in 3-Dimension. The Isolation Forest algorithm found 631 outliers on Almerys data.

2D PCA Results



Experiments > Report >

ANOMALY DETECTION

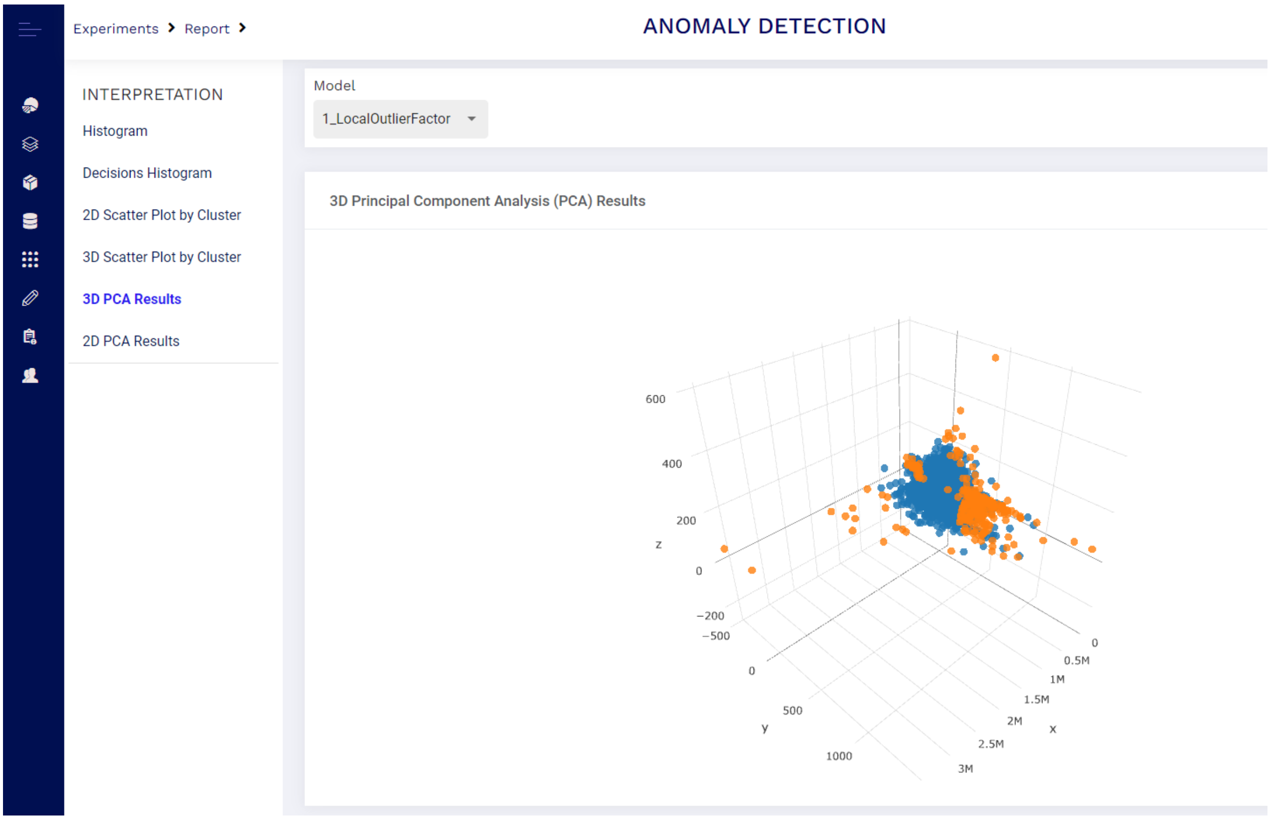
Model: 1_LocalOutlierFactor

2D Principal Component Analysis (PCA) Results

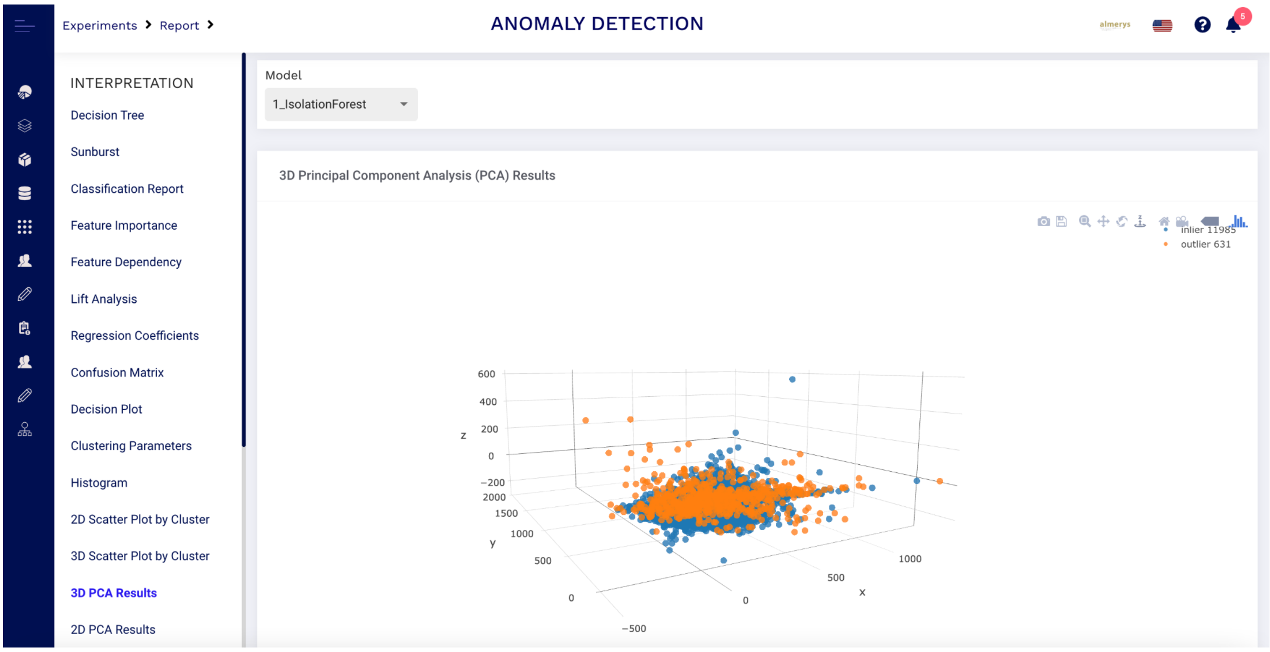
This graph shows the 2D PCA result of Almerys data based on outliers and inliers.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

3D PCA Reporting and Results



Model: 1_LocalOutlierFactor



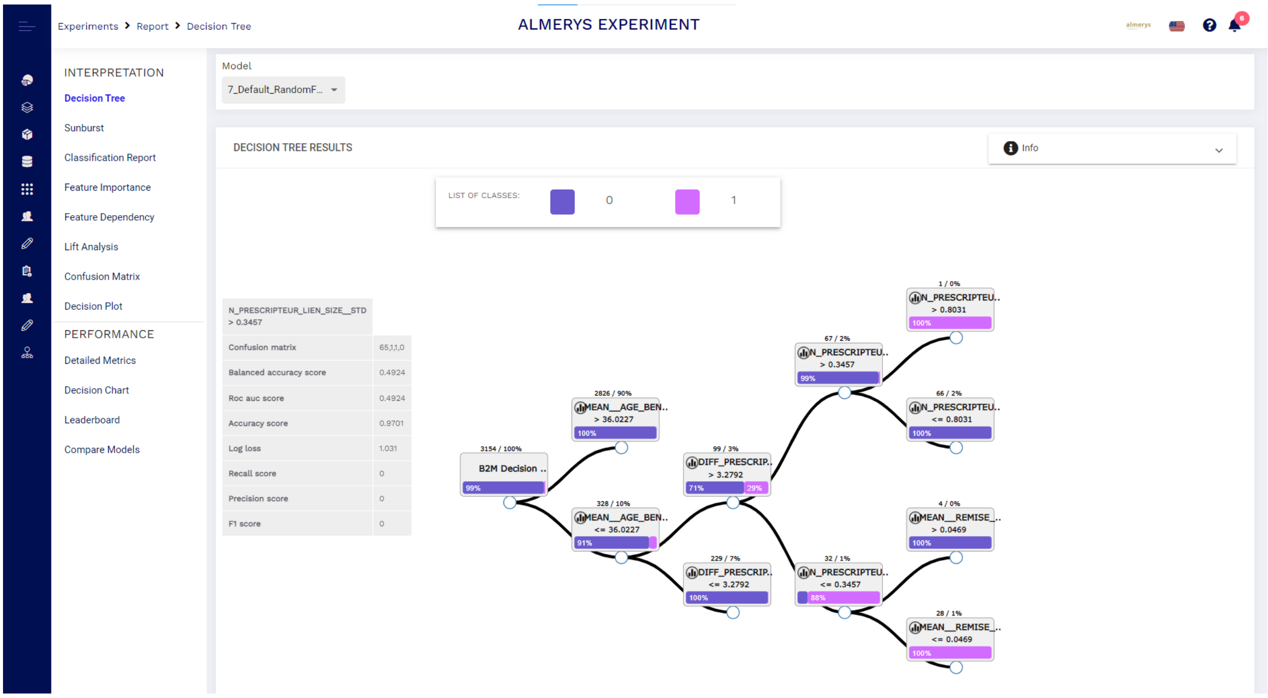
Model: 1_IsolationForest

These graphs show the 3D PCA result of Almerys data based on outliers and inliers.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 951981

B2Metric AutoML Hunter Almerys Fraud Detection Supervised Module Explainable AI Classification Model Results

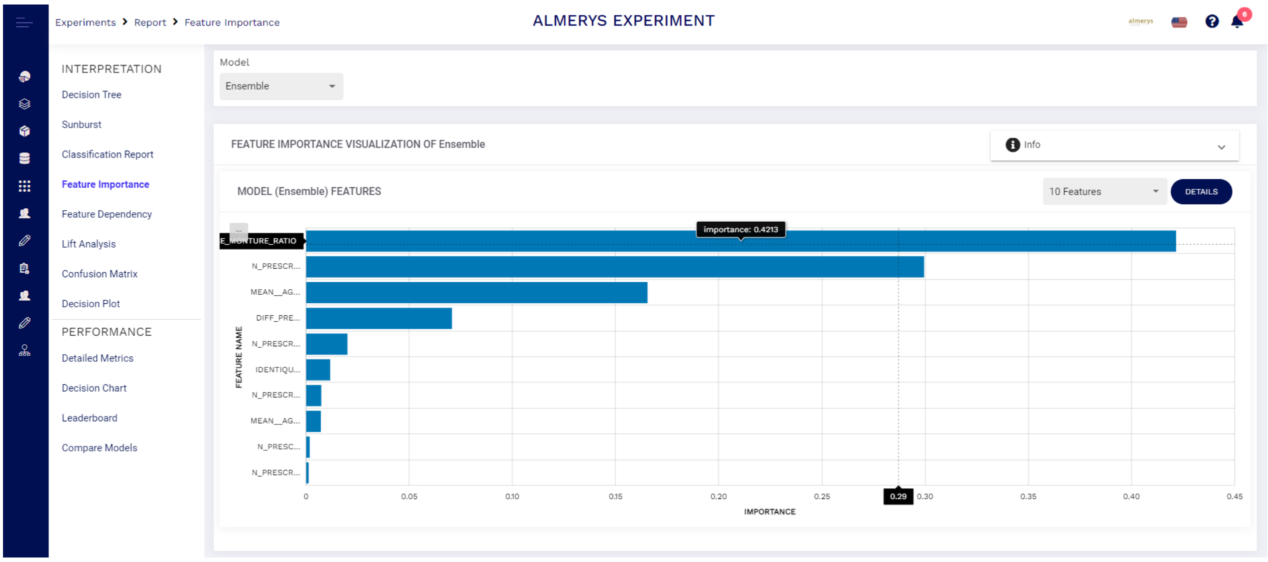
Decision Tree



N_PRESCRIPTEUR_LIEN_SIZE__STD > 0.3457	
Confusion matrix	65,1,1,0
Balanced accuracy score	0.4924
Roc auc score	0.4924
Accuracy score	0.9701
Log loss	1.031
Recall score	0
Precision score	0
F1 score	0

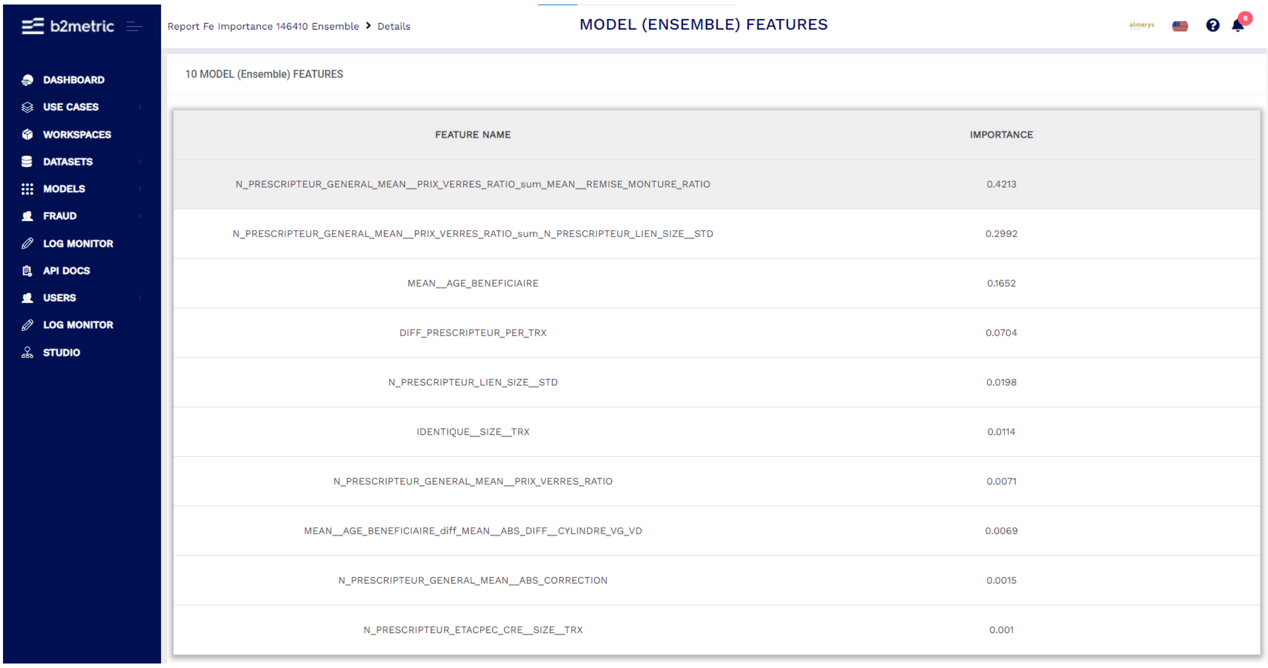
Decision tree explains how the model makes decisions by transforming the results of the data distribution learned through decision tree modelling into simple rule sets.

Feature Importance Reports



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

Feature importance, also known as permutation importance, refers to techniques that calculate all input features as a score for a particular model. These scores represent the "importance" of each feature. A higher score means that a particular feature will have a greater impact on the model outcome used to predict a particular variable.

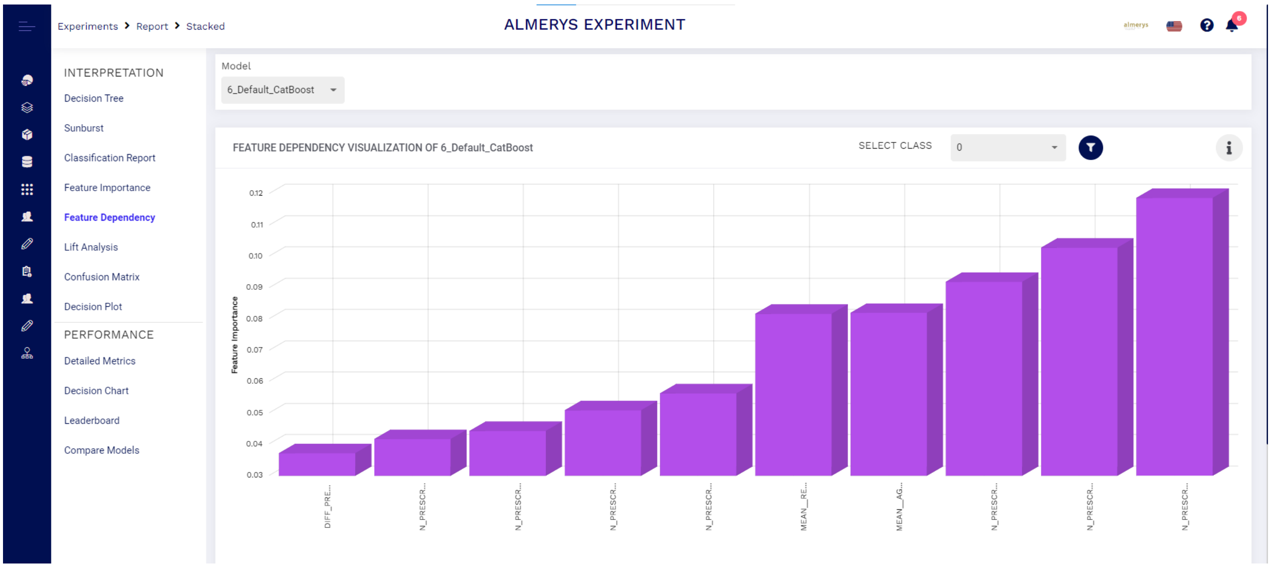


MODEL (ENSEMBLE) FEATURES

10 MODEL (Ensemble) FEATURES

FEATURE NAME	IMPORTANCE
N_PRESCRIPTEUR_GENERAL_MEAN__PRIX_VERRES_RATIO_sum_MEAN__REMISE_MONTURE_RATIO	0.4213
N_PRESCRIPTEUR_GENERAL_MEAN__PRIX_VERRES_RATIO_sum_N_PRESCRIPTEUR_LIEN_SIZE__STD	0.2992
MEAN__AGE_BENEFICIAIRE	0.1652
DIFF_PRESCRIPTEUR_PER_TRX	0.0704
N_PRESCRIPTEUR_LIEN_SIZE__STD	0.0198
IDENTIQUE__SIZE__TRX	0.0114
N_PRESCRIPTEUR_GENERAL_MEAN__PRIX_VERRES_RATIO	0.0071
MEAN__AGE_BENEFICIAIRE_diff_MEAN__ABS_DIFF__CYLINDRE_VG_VD	0.0069
N_PRESCRIPTEUR_GENERAL_MEAN__ABS_CORRECTION	0.0015
N_PRESCRIPTEUR_ETACPEC_CRE__SIZE__TRX	0.001

Feature Dependency Reports



ALMERYS EXPERIMENT

Model: 6_Default_CatBoost

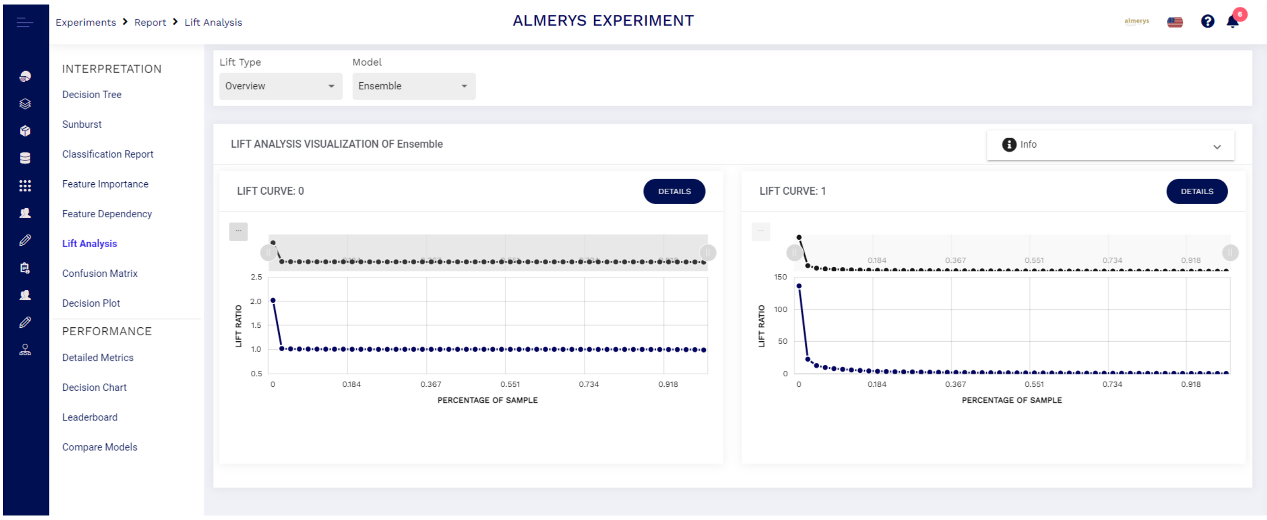
FEATURE DEPENDENCY VISUALIZATION OF 6_Default_CatBoost

SELECT CLASS: 0

Feature Dependency, also known as Shapley Importance, determines what R-Squared ratio can be attributed to each independent variable from a linear regression model. In other words, the Shapley Importance describes how much influence a particular variable has in estimating the dependent variable.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

Lift Analysis



Experiments > Report > Lift Analysis

ALMERYS EXPERIMENT

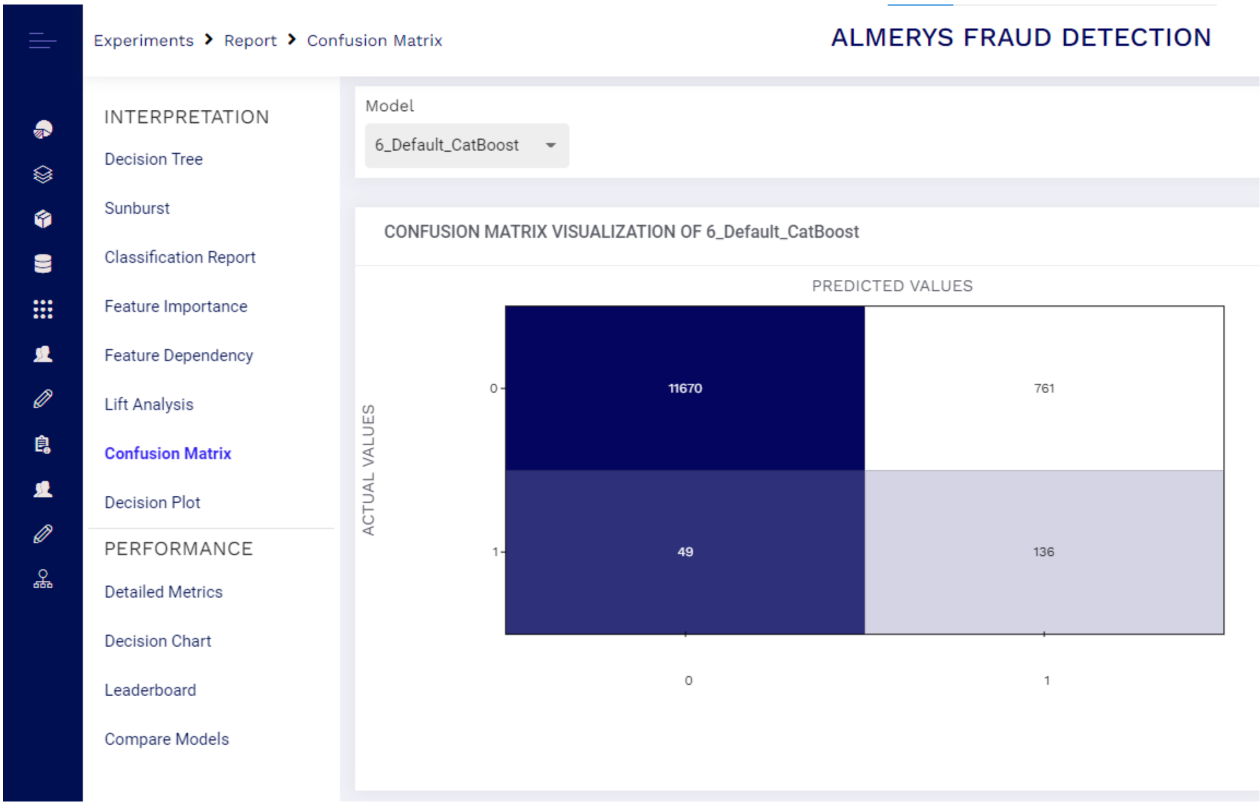
Lift Type: Overview

Model: Ensemble

LIFT ANALYSIS VISUALIZATION OF Ensemble

It represents the increase in the predictive power, that is, the success rate, of a machine learning model compared to the baseline model (dummy model).

Confusion Matrix



Experiments > Report > Confusion Matrix

ALMERYS FRAUD DETECTION

Model: 6_Default_CatBoost

CONFUSION MATRIX VISUALIZATION OF 6_Default_CatBoost

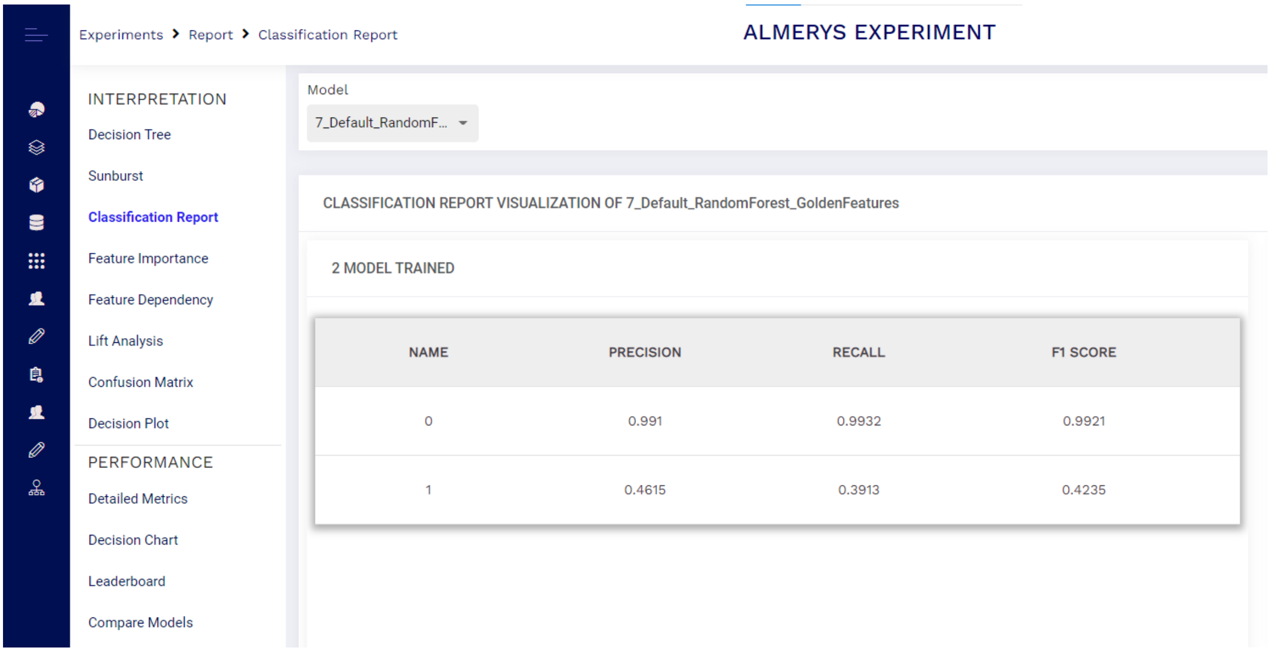
		PREDICTED VALUES	
		0	1
ACTUAL VALUES	0	11670	761
	1	49	136

The confusion matrix is a table summarising how successful the classification model is in estimating examples of related classes. One axis of the confusion matrix represents the values predicted by the model, and the other axis represents the actual values. CatBoost model predicted 136 of 185 fraud opticians as

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

fraud and 11670 of 12431 not frauds as not fraud.

Classification Report



Experiments > Report > Classification Report

ALMERYS EXPERIMENT

INTERPRETATION
- Decision Tree
- Sunburst
- Classification Report
- Feature Importance
- Feature Dependency
- Lift Analysis
- Confusion Matrix
- Decision Plot

PERFORMANCE
- Detailed Metrics
- Decision Chart
- Leaderboard
- Compare Models

Model: 7_Default_RandomF...

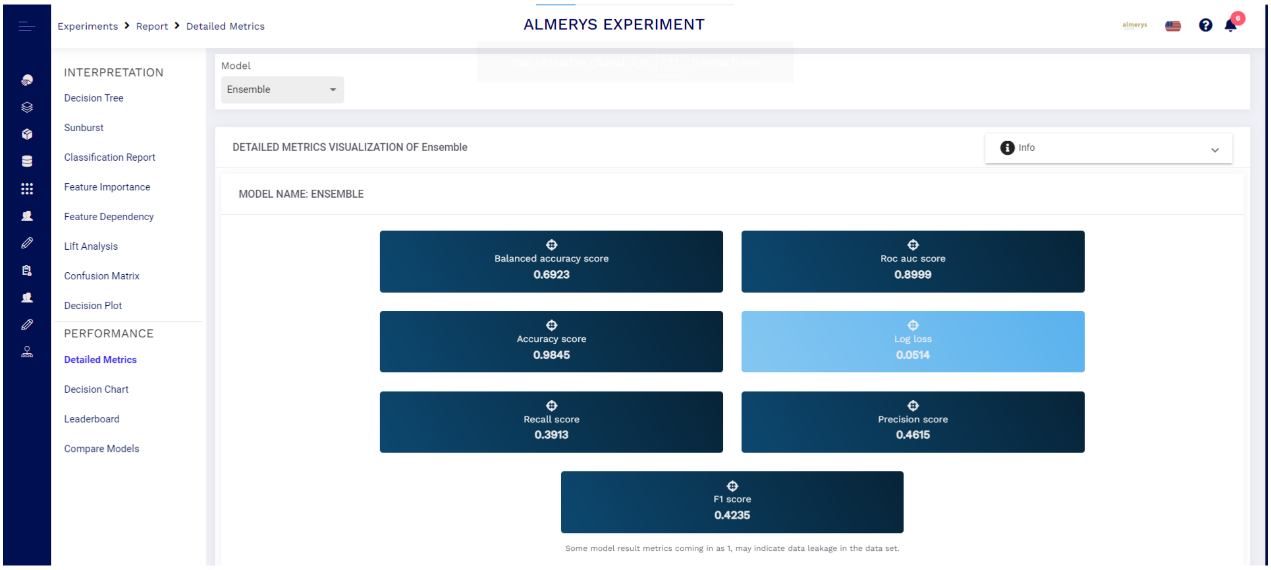
CLASSIFICATION REPORT VISUALIZATION OF 7_Default_RandomForest_GoldenFeatures

2 MODEL TRAINED

NAME	PRECISION	RECALL	F1 SCORE
0	0.991	0.9932	0.9921
1	0.4615	0.3913	0.4235

On this page, classification metrics of target variables are reported separately for each algorithm.

Detailed Metrics



Experiments > Report > Detailed Metrics

ALMERYS EXPERIMENT

almerys

INTERPRETATION
- Decision Tree
- Sunburst
- Classification Report
- Feature Importance
- Feature Dependency
- Lift Analysis
- Confusion Matrix
- Decision Plot

PERFORMANCE
- Detailed Metrics
- Decision Chart
- Leaderboard
- Compare Models

Model: Ensemble

DETAILED METRICS VISUALIZATION OF Ensemble

Info

MODEL NAME: ENSEMBLE

Balanced accuracy score 0.6923

Roc auc score 0.8999

Accuracy score 0.9845

Log loss 0.0514

Recall score 0.3913

Precision score 0.4615

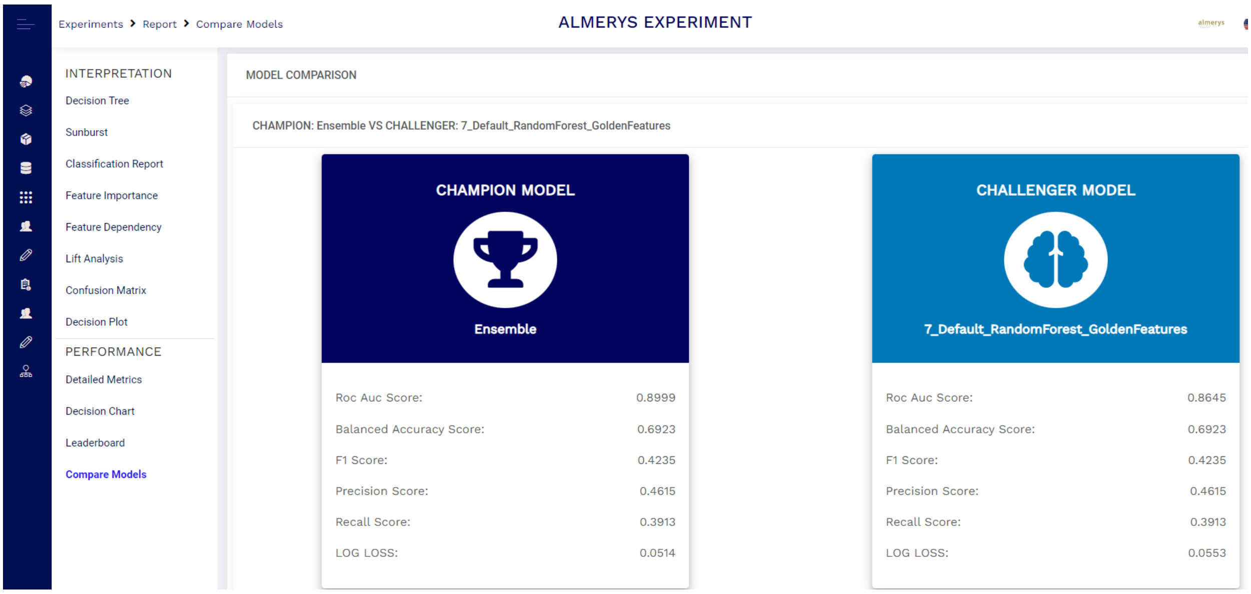
F1 score 0.4235

Some model result metrics coming in as 1, may indicate data leakage in the data set.

On the Detailed metrics page, model result metrics are reported separately for all algorithms that are run. With the model selection filter in the upper left, you can select the model you want to report the results

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

Compare Models and Model Selection Results Report



Experiments > Report > Compare Models

ALMERYS EXPERIMENT

INTERPRETATION
- Decision Tree
- Sunburst
- Classification Report
- Feature Importance
- Feature Dependency
- Lift Analysis
- Confusion Matrix
- Decision Plot

PERFORMANCE
- Detailed Metrics
- Decision Chart
- Leaderboard
- Compare Models

MODEL COMPARISON

CHAMPION: Ensemble VS CHALLENGER: 7_Default_RandomForest_GoldenFeatures

CHAMPION MODEL

Ensemble

Roc Auc Score:	0.8999
Balanced Accuracy Score:	0.6923
F1 Score:	0.4235
Precision Score:	0.4615
Recall Score:	0.3913
LOG LOSS:	0.0514

CHALLENGER MODEL

7_Default_RandomForest_GoldenFeatures

Roc Auc Score:	0.8645
Balanced Accuracy Score:	0.6923
F1 Score:	0.4235
Precision Score:	0.4615
Recall Score:	0.3913
LOG LOSS:	0.0553

On the Compare Models page provided by the B2Metric AutoML platform, the model result metrics of all algorithms run while training the Almerys Fraud Detection model are reported comparatively.

2. **ALGORITHMS, TOOLS AND CONCLUSIONS:** Detail the algorithms and tools identified to accomplish the challenge/Theme Challenges. Show clear understanding of the used REACH dataset/s and addressed challenge/Theme Challenges.

During the ML modelling and statistical insights generation for the use case of DATA SCIENCE/ DATA MANIPULATION IN ORDER TO GAIN INSIGHTS FROM THE MARKET with the data of '*Historical data of reimbursement requests from opticians to insurance companies*'. Algorithms and conclusions of technical details are listed below.

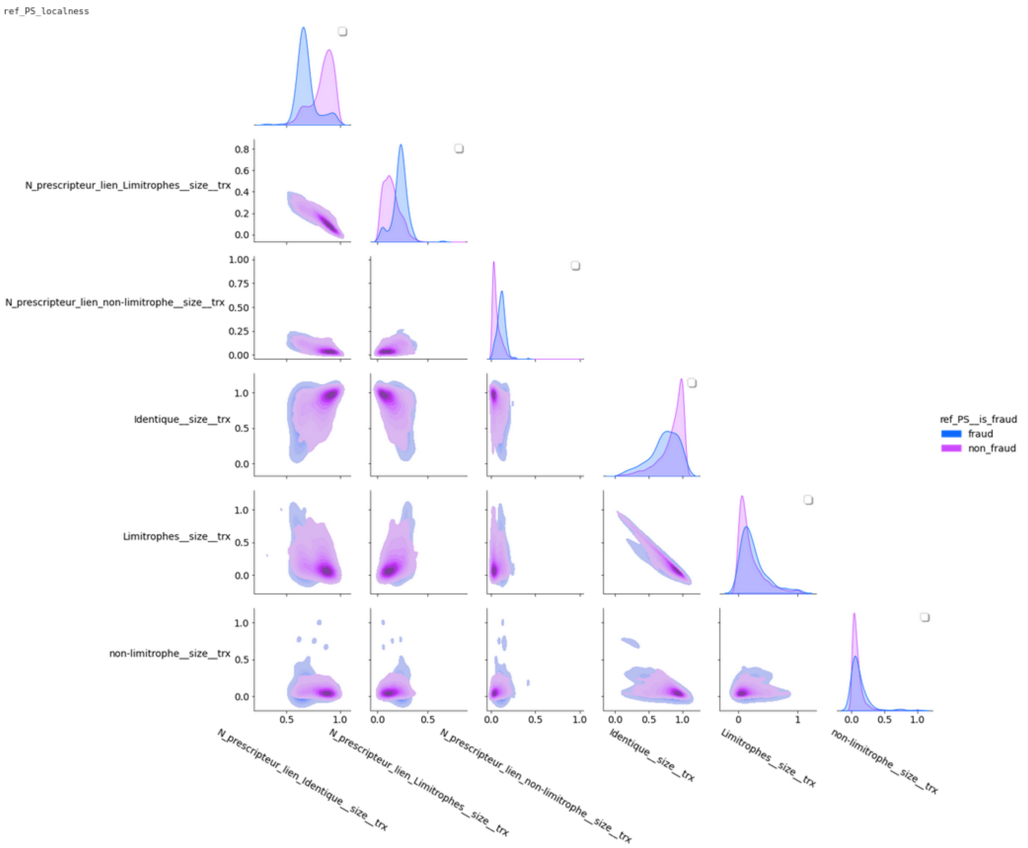
1. **Exploratory Data Analysis:** Referring to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypotheses and to check assumptions with the help of summary statistics and graphical representations.
2. **Cluster Analysis:** Identify its various customer segments, and then conduct cluster analysis to see if any such segments share similar characteristics (e.g. objectives, pain points, perceptions, demographics, preferences, etc.) that are distinctly different from other segments.
3. **Factor Analysis:** Shed light on what combination of aspects, characteristics or priorities are most important to a certain type of customers (group).
4. **Discriminant Analysis:** Predicting membership in a group (or population or cluster) based on measured characteristics of other variables. Subsequently, it is requested to be based on the previous results (market analysis, etc.) to: Extract outliers from the dataset (according to several axes of analysis). Carry out an in-depth analysis of these data to identify fraud movements
5. **Fraud Detection using Multiple Regression:** Predict the value of a variable based on changes to two or more variables. For the reason that the target feature is categorical, it will be a

classification model.

6. **Adaptive Learning & Continues Learning of Fraud Models.**

1-) Exploration Data Phase - Understanding Data

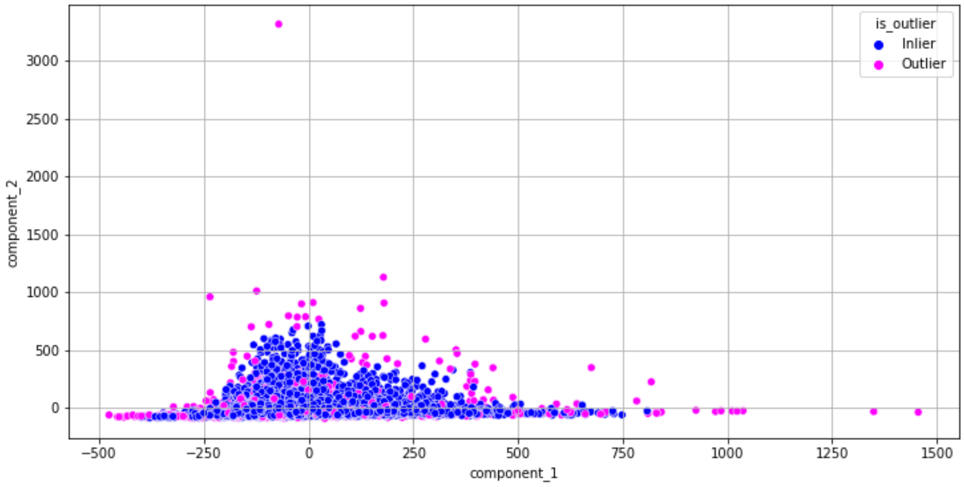
The dataset consists of behavioural descriptions of opticians, chains, and prescriptions and several qualities about customers. It also has some logic errors like negative customer ages or negative brute prices, so they all have to be cleaned up before exploratory data analysis and ML modelling. After understanding and cleaning data processes, the observations that can be outliers should be detected in order to smooth the skewness and protect the distribution of continuous features.



Above visualisation method, it shows us the combination relations between two of the most important features and target.

1.1) Local Outlier Factor

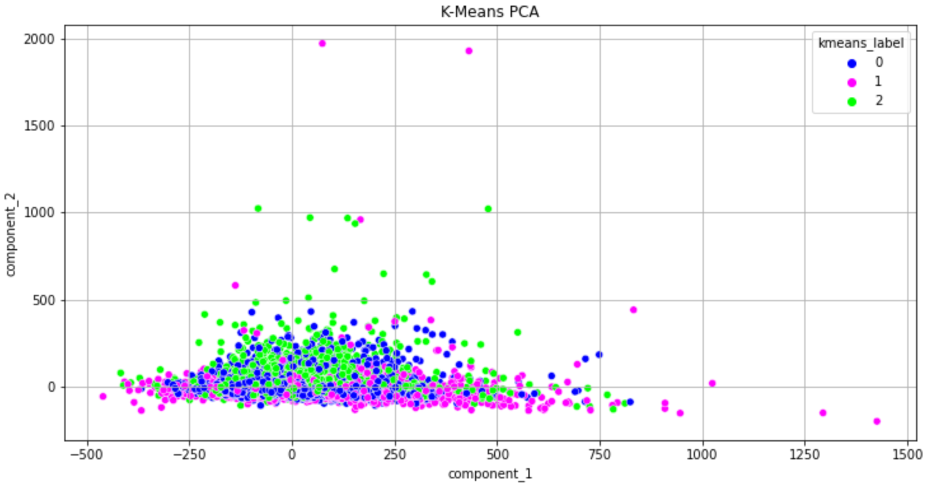
The Local Outlier Factor algorithm is an unsupervised anomaly detection method which computes the local density deviation of a given data point with respect to its neighbours. It considers as outliers the samples that have a substantially lower density than their neighbours.



1262 outliers detected from 12616 observations with Local Outlier Factor. According to the purpose of the study, different approaches may be preferred to outliers. Since fraud is an illegal behaviour, the possibility of some outliers being fraud cannot be excluded.

2.1) K-Means Clustering

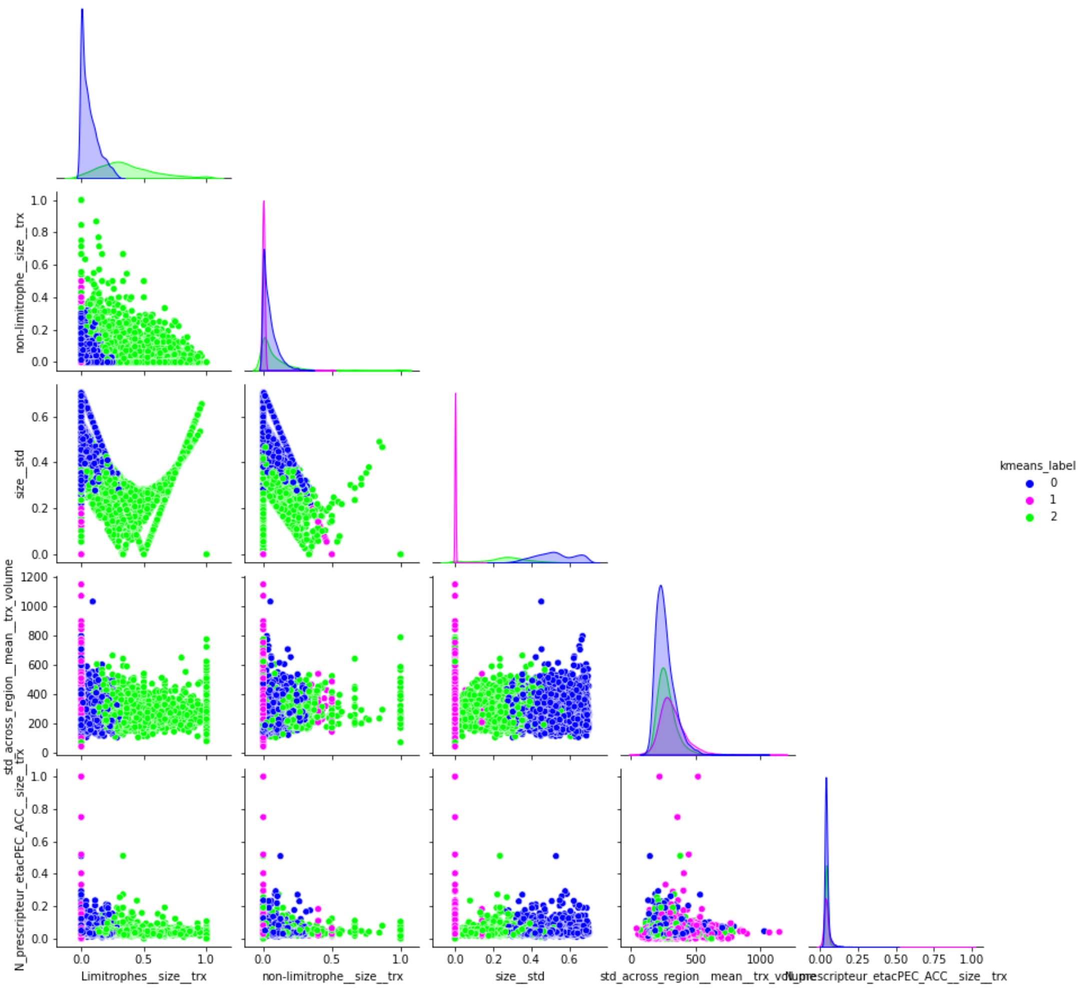
K-Means Clustering uses “centroids”, K different randomly-initiated points in the data, and assigns every data point to the nearest centroid. After every point has been assigned, the centroid is moved to the average of all of the points assigned to it. Then the process repeats: every point is assigned to its nearest centroid, centroids are moved to the average of points assigned to it. The algorithm is done when no point changes the assigned centroid. Customer segmentation is the process of dividing customers into groups based on common characteristics so companies can market to each group effectively and appropriately.



Customers registered with opticians were segmented based on similar characteristics using the K-Means clustering algorithm, and found the optimum cluster size as 3. The Silhouette Coefficient is one of the most popular model evaluation metrics for clustering algorithms. The Silhouette Coefficient for a set of samples is given as the mean of the Silhouette Coefficient for each sample. The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering. Scores around zero indicate overlapping clusters. The

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 951981

score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster. This clustering module, the Silhouette Coefficient is calculated as 0.36.

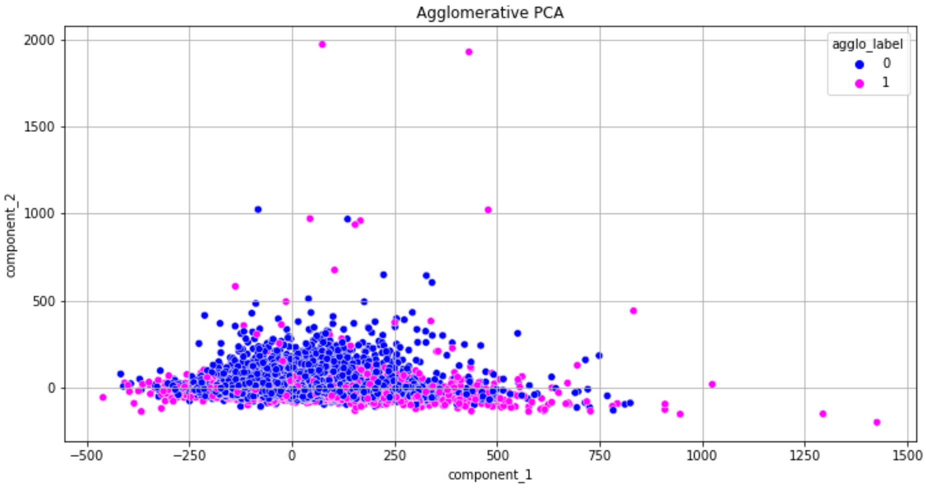


The comparison of the characteristics of customer segments, differences on the basis of some variables come to the fore compared to other variables.

2.2) Agglomerative Clustering

Agglomerative Clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It's also known as AGNES (Agglomerative Nesting). The algorithm starts by treating each object as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. In addition to K-Means Clustering, Agglomerative Clustering is an option for customer segmentation.

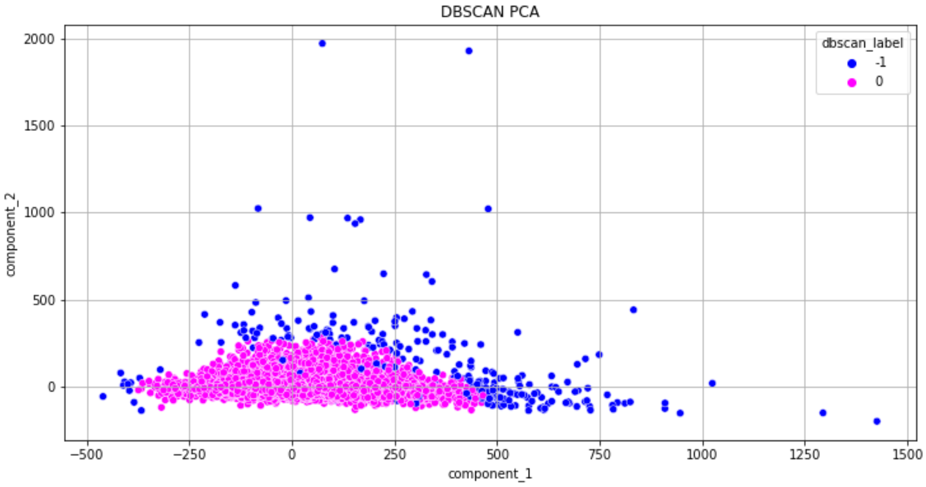
This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 951981



In K-Means Clustering, the optimum cluster size was found as 3, but in Agglomerative Clustering, it had the best results with 2 clusters and the Silhouette Coefficient calculated as 0.30.

2.3) Clustering with DBSCAN

DBSCAN is a density-based clustering algorithm that works on the assumption that clusters are dense regions in space separated by regions of lower density.For locating data points in space, DBSCAN uses Euclidean distance, although other methods can also be used (like great circle distance for geographical data). It also needs to scan through the entire dataset once, whereas in other algorithms we have to do it multiple times.



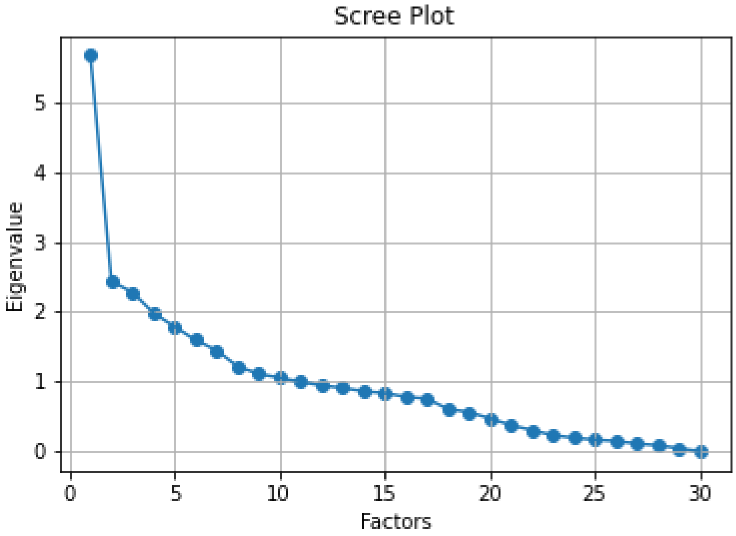
Principal Component Analysis is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. It allows observing all variables with 2 dimensions in cluster analysis of opticians.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

3) Factor Analysis

Factor analysis is a technique that is used to reduce a large number of variables into fewer numbers of factors. This technique extracts maximum common variance from all variables and puts them into a common score. As an index of all variables, we can use this score for further analysis. Factor analysis is part of the general linear model (GLM) and this method also assumes several assumptions: there is linear relationship, there is no multicollinearity, it includes relevant variables into analysis, and there is true correlation between variables and factors. Several methods are available, but principal component analysis is used most commonly.

Also, you can check 21-22-23 pages of this report B2Metric AutoML Hunter Almerys Anomaly Detection PCA results in plots.

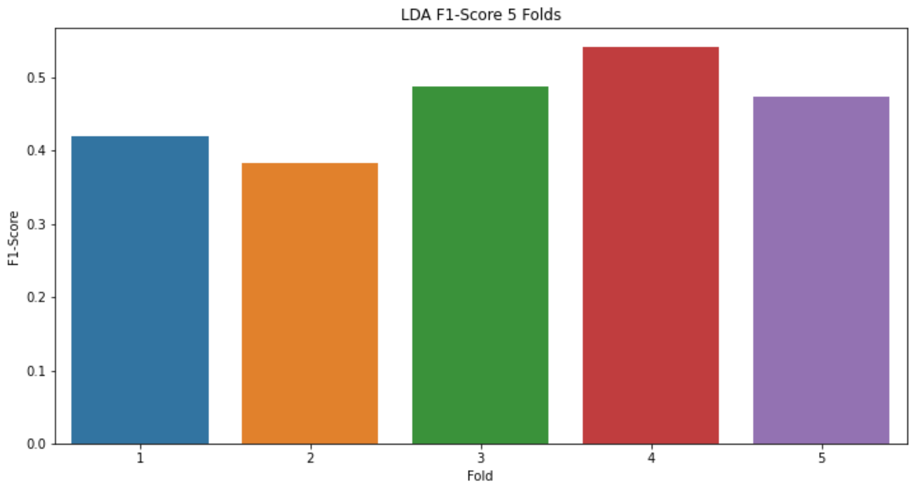


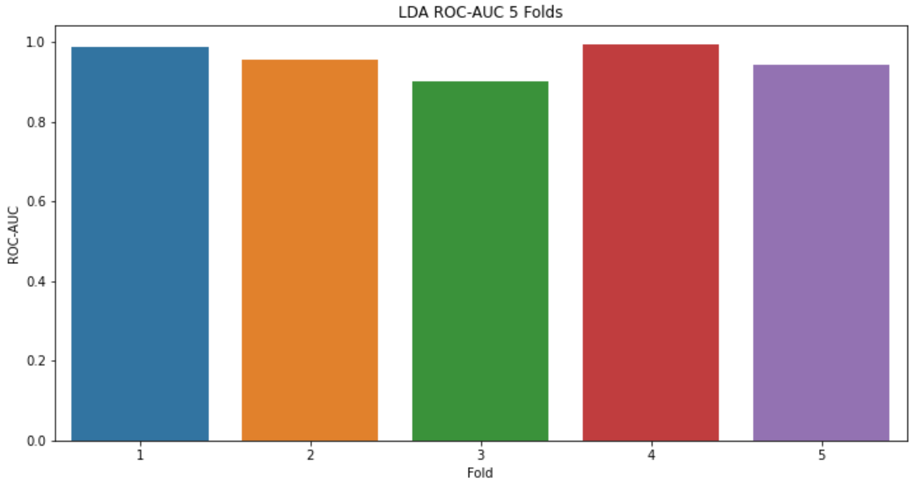
In multivariate statistics, a scree plot is a line plot of the eigenvalues of factors or principal components in an analysis. The scree plot is used to determine the number of factors to retain in an exploratory factor analysis (FA) or principal components to keep in a principal component analysis (PCA). The procedure of finding statistically significant factors or components using a scree plot is also known as a scree test. There are 10 data points with eigenvalue greater than 1, so the optimum factor size is found as 10. Loadings close to -1 or 1 indicate that the factor strongly influences the variable. Loadings close to 0 indicate that the factor has a weak influence on the variable. Some variables may have high loadings on multiple factors. Unrotated factor loadings are often difficult to interpret.

4) Discriminant Analysis

Discriminant analysis is a versatile statistical method often used by market researchers to classify observations into two or more groups or categories. In other words, discriminant analysis is used to assign objects to one group among a number of known groups.

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 951981





The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

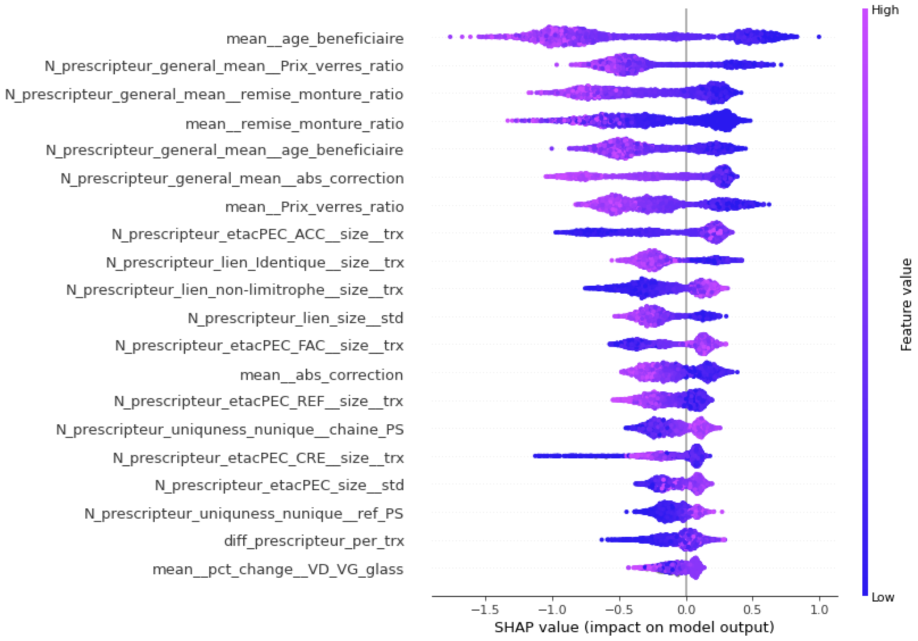
5) Fraud Detection using Multiple Regression

Fraud detection is a set of activities undertaken to prevent money or property from being obtained through false pretences.

Fraud detection is applied to many industries such as banking or insurance. In banking, fraud may include forging checks or using stolen credit cards. Other forms of fraud may involve exaggerating losses or causing an accident with the sole intent for the payout. Health-care fraud occurs when a health-care professional knowingly, willfully, and intentionally makes a false statement or claim. Making false statements or documentation to obtain program benefits, such as Medicare reimbursement when a provider would otherwise not be entitled to payment, is fraudulent.

Fraud typically involves multiple repeated methods, making searching for patterns a general focus for fraud detection. For example, data analysts can prevent insurance fraud by making algorithms to detect patterns and anomalies.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

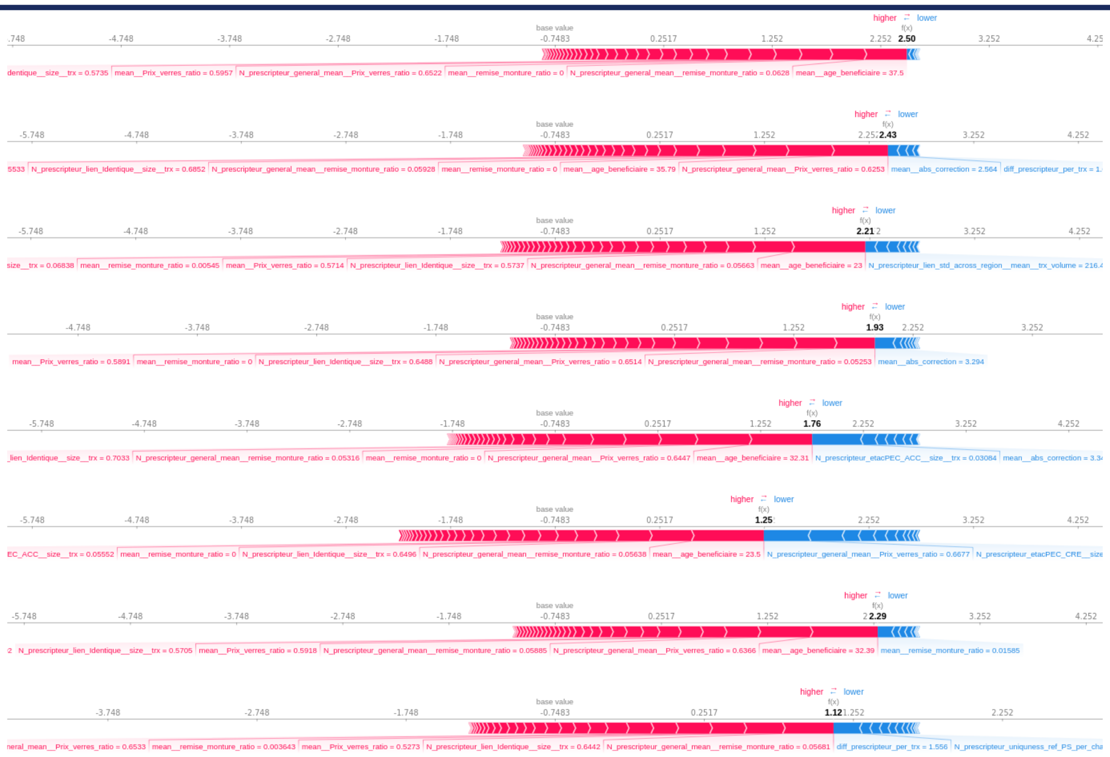


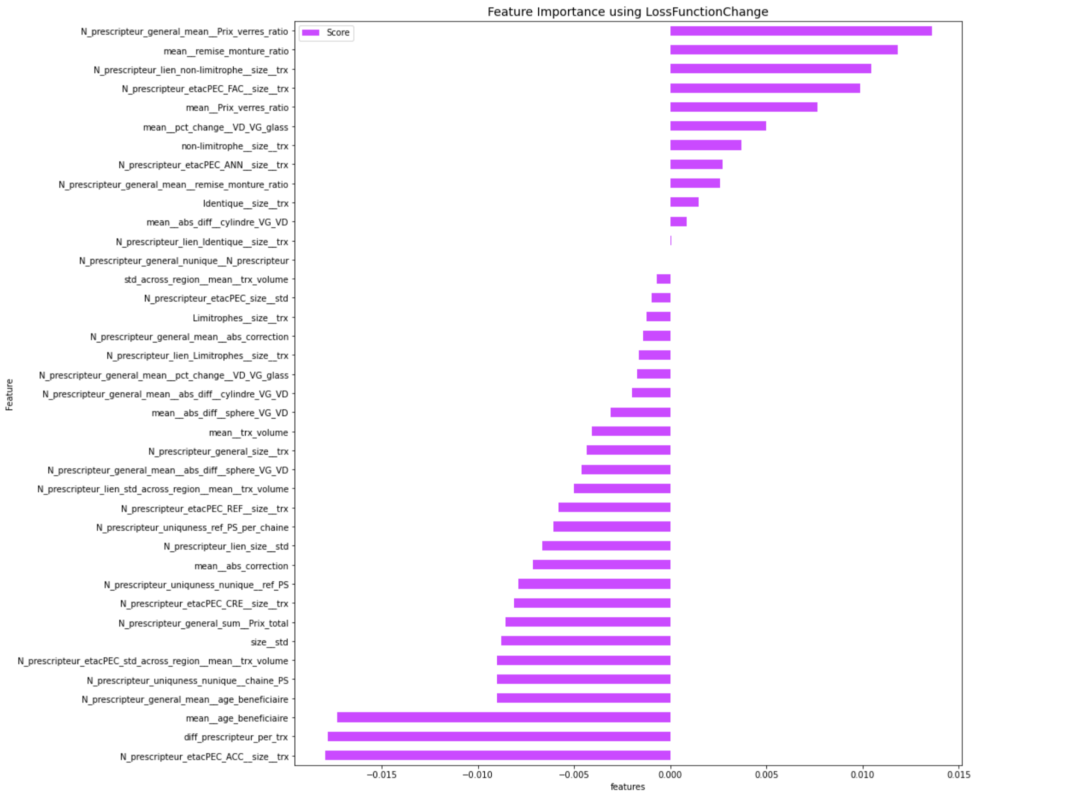
In feature extraction, we create new features from the “optical_care_transaction.csv” file. As you can see in the above image, we generate opticians based features but also prescripteur based features which have at least one time with this optician.

- mean_age_beneficiaire: Mean of age beneficiaire within the same optician transaction.
- N_prescripteur_general_mean__Prix_verres_ratio: Mean of Prix Verres within all transactions whose doctor wrote this.
- N_prescripteur_general_mean__remise_monture_ratio: Mean of Remise Monture within all transactions whose doctor wrote this.
- N_prescripteur_general_mean__abs_correction: Mean of Abs_Correction within all transactions whose doctor wrote this.

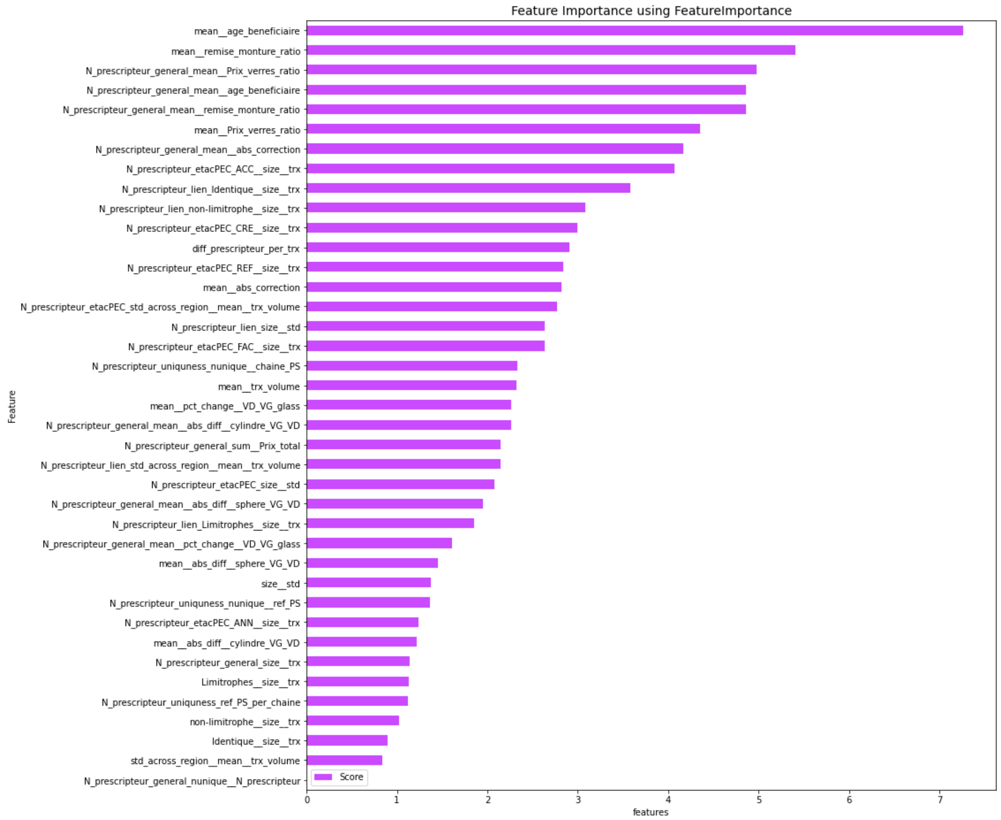
The Shapley value is the average marginal contribution of a feature value across all possible coalitions.The computation time increases exponentially with the number of features. One solution to keep the computation time manageable is to compute contributions for only a few samples of the possible coalitions.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 951981





This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981



Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable.

There are many types and sources of feature importance scores, although popular examples include statistical correlation scores, coefficients calculated as part of linear models, decision trees, and permutation importance scores.

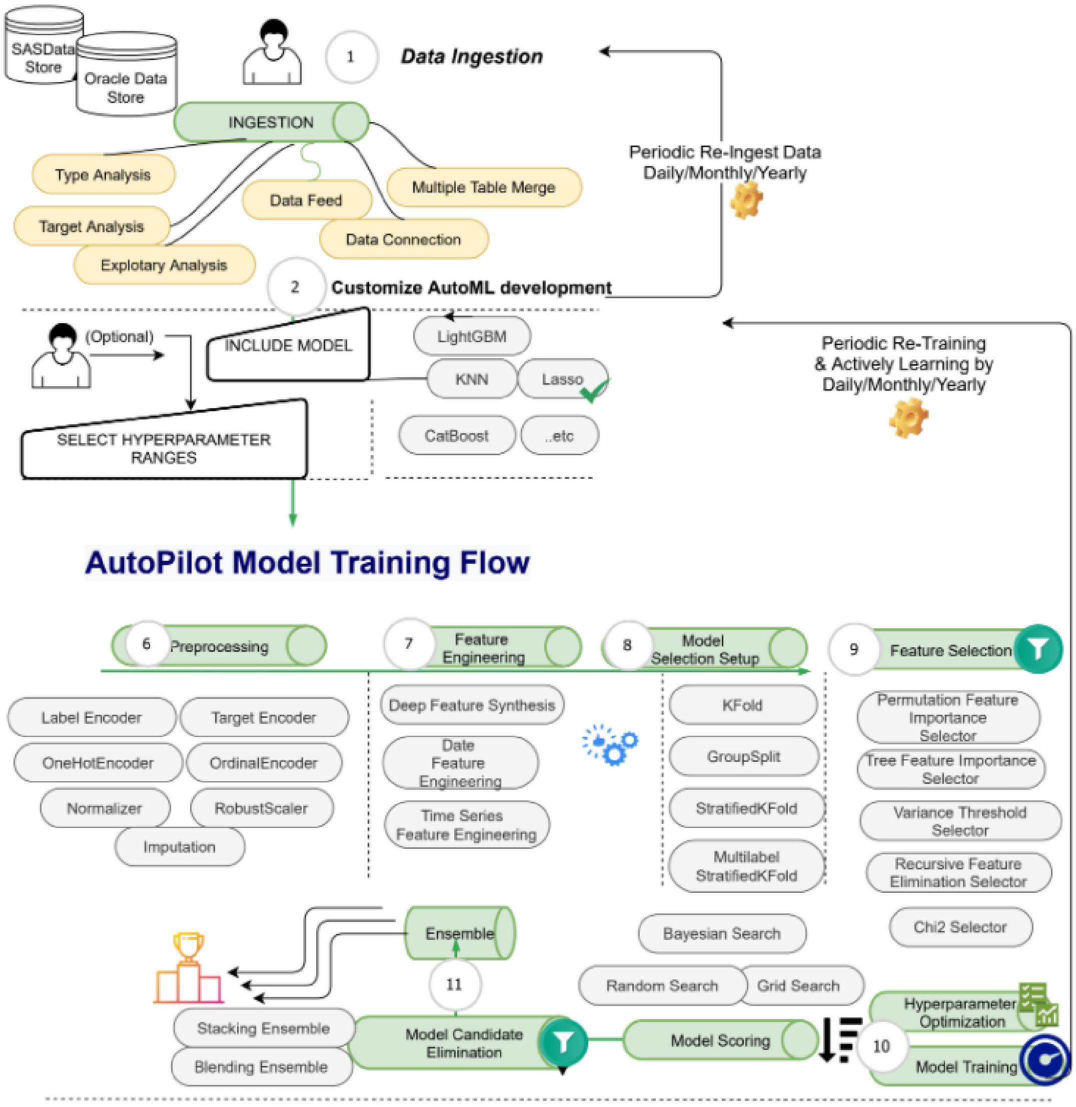
Feature importance scores play an important role in a predictive modelling project, including providing insight into the data, insight into the model, and the basis for dimensionality reduction and feature selection that can improve the efficiency and effectiveness of a predictive model on the problem.

In fraud detection model evaluation, the mean age of customers and the features extracted based on prescriptions has the highest importances.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

ADAPTIVE LEARNING FRAUD DETECTION MODELING

Continues learning and auto retraining for fraud models is one of the crucial point for the new clients, new opticians and transactions data generated. Here is the system architecture below has designed for Almersy's Fraud modelling adaptively learning solution in B2Metric Auto-ML Hunter framework.



End-to-End ML Modeling Flow Pipeline Setup with B2Metric ML Studio

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

3. **SCALABILITY AND FLEXIBILITY OF THE SOLUTION:** Discuss whether the solution can truly cope with humongous and increasing datasets and how flexible it is to adapt to other related domains

As for scalability and flexibility of the solution we have several IT design implementations in our solution set. For instance, B2Metric is running in EMEA's leading Telco giant Turk Telecom and we run more than 120 classification and regression ML models in parallel. We implement B2Metric ML Studio the architecture below. Also we implemented Apache Spark based scalibility on top of Kubernetes based instances.

Another work carried out for the MVP project's scalability during the period was the work carried out to ensure system integration and to launch the Hunter Almerys platform by installing it for the healthcare insurance customers. Within the scope of these studies, Kubernetes, Docker and Rancher environments were prepared.

Docker: Docker is a technology that provides virtualization thanks to hundreds or even thousands of isolated and independent containers on the same operating system.

Kubernetes: Kubernetes is a container clustering tool that allows you to manage your existing containerized applications, supported by the Cloud Native Computing Foundation, developed in the GO language by Google, with operations such as automatically deploying, increasing or decreasing their number.

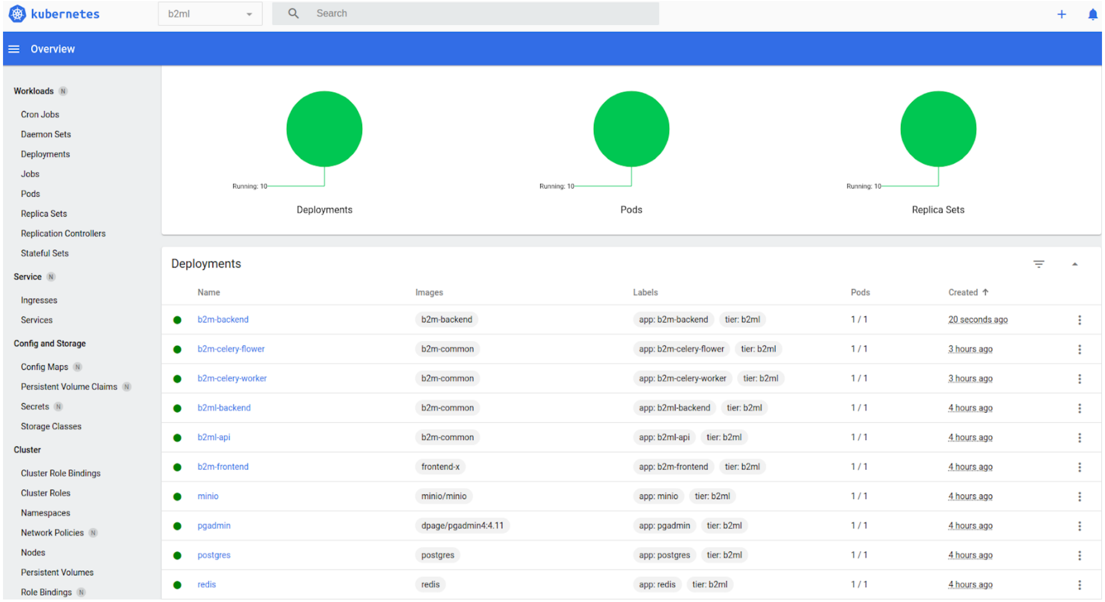
Rancher: With Rancher, Kubernetes is installed and managed, which takes us out of the complexity of the Kubernetes structure. Rancher, a Kubernetes management tool, also provides great convenience in log tracking and system monitoring stages.

Kubernetes Studies within the B2Metric Hunter Almerys Project

It has gained importance to follow the innovations in the application development processes with the developing technology and tools. Administrative difficulties brought by the acceleration of processes with container and microservice structures are organised with different tools and it is aimed to accelerate the workflow and follow-up.

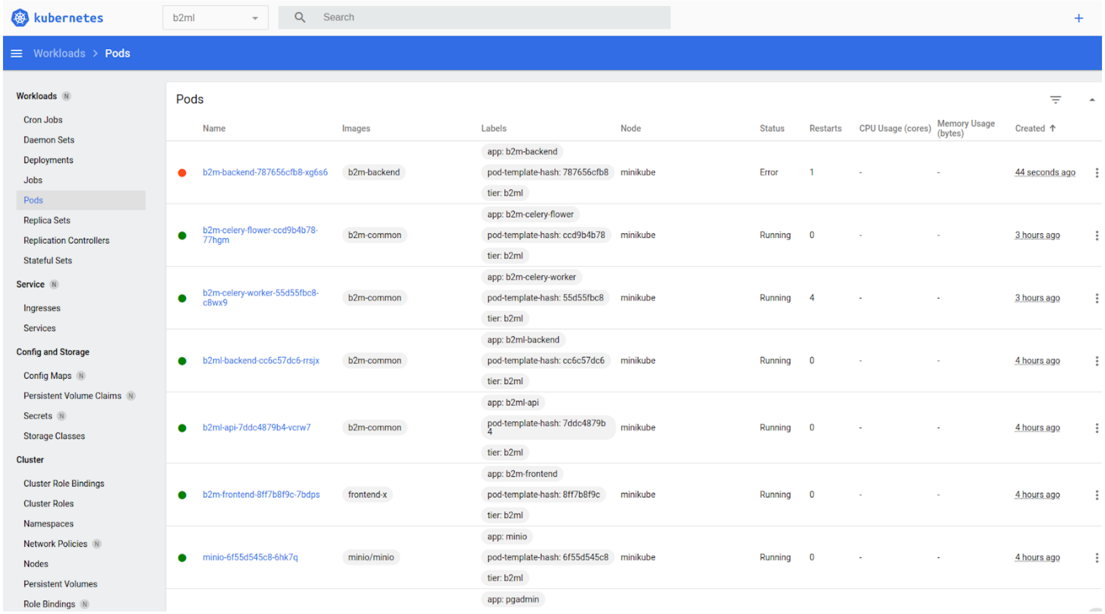
In this context, the application is migrated to the docker container structure and studies are carried out for performance improvements. For the management of the dockerized application, processes such as making it compatible with kubernetes - rancher - openshift applications, development of CI/CD processes and log tracking, performance and security improvements are continued.

Here is the screenshots below our Fraud Detection platform Rancher Kubernetes Instances after deployment to AWS EC2 Frankfurt instances.



Deployments

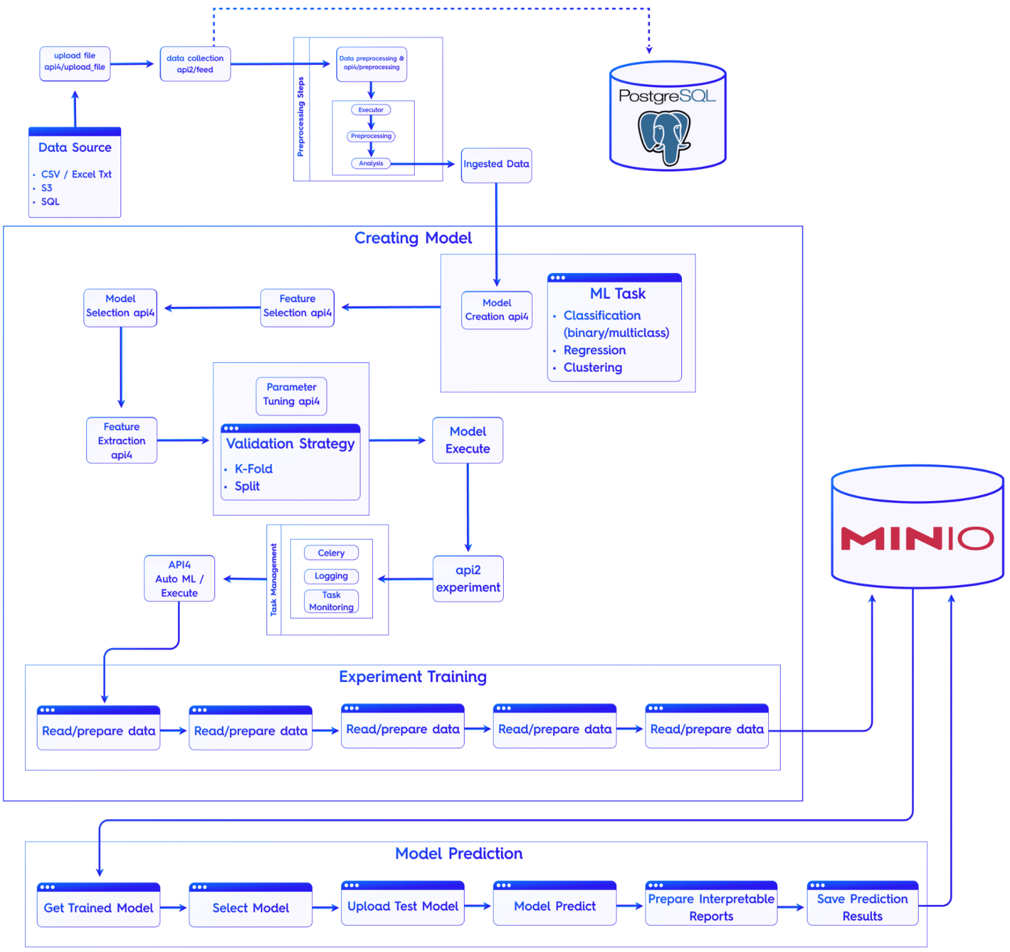
	Name	Images	Labels	Pods	Created ↑
●	b2m-backend	b2m-backend	app: b2m-backend tier: b2ml	1 / 1	20 seconds ago
●	b2m-celery-flower	b2m-common	app: b2m-celery-flower tier: b2ml	1 / 1	3 hours ago
●	b2m-celery-worker	b2m-common	app: b2m-celery-worker tier: b2ml	1 / 1	3 hours ago
●	b2ml-backend	b2m-common	app: b2ml-backend tier: b2ml	1 / 1	4 hours ago
●	b2ml-api	b2m-common	app: b2ml-api tier: b2ml	1 / 1	4 hours ago
●	b2m-frontend	frontend-x	app: b2m-frontend tier: b2ml	1 / 1	4 hours ago
●	minio	minio/minio	app: minio tier: b2ml	1 / 1	4 hours ago
●	pgadmin	dpage/pgadmin4:4.11	app: pgadmin tier: b2ml	1 / 1	4 hours ago
●	postgres	postgres	app: postgres tier: b2ml	1 / 1	4 hours ago
●	redis	redis	app: redis tier: b2ml	1 / 1	4 hours ago



Pods

	Name	Images	Labels	Node	Status	Restarts	CPU Usage (cores)	Memory Usage (bytes)	Created ↑
●	b2m-backend-787656cfb8-xg6s6	b2m-backend	app: b2m-backend pod-template-hash: 787656cfb8 tier: b2ml	minikube	Error	1	-	-	44 seconds ago
●	b2m-celery-flower-ccd9b4b78-77hgm	b2m-common	app: b2m-celery-flower pod-template-hash: ccd9b4b78 tier: b2ml	minikube	Running	0	-	-	3 hours ago
●	b2m-celery-worker-55d55fbc8-c8wx9	b2m-common	app: b2m-celery-worker pod-template-hash: 55d55fbc8 tier: b2ml	minikube	Running	4	-	-	3 hours ago
●	b2ml-backend-cc6c57dc6-rrsjx	b2m-common	app: b2ml-backend pod-template-hash: cc6c57dc6 tier: b2ml	minikube	Running	0	-	-	4 hours ago
●	b2ml-api-7ddc4879b4-vcrw7	b2m-common	app: b2ml-api pod-template-hash: 7ddc4879b4 tier: b2ml	minikube	Running	0	-	-	4 hours ago
●	b2m-frontend-8ff7b8f9c-7bdps	frontend-x	app: b2m-frontend pod-template-hash: 8ff7b8f9c tier: b2ml	minikube	Running	0	-	-	4 hours ago
●	minio-6f55d545c8-6hk7q	minio/minio	app: minio pod-template-hash: 6f55d545c8 tier: b2ml	minikube	Running	0	-	-	4 hours ago
			app: pgadmin						

B2Metric Hunter Auto-ML Fraud Detection Software Architecture design



4. **DATA GOVERNANCE AND LEGAL COMPLIANCE:** Describe the security level of the proposed solution, i.e. how authentication, authorization policies, encryption or other approaches are used to keep data secure. Explain how you will be compliant with the current data legislations concerning security and privacy (e.g. GDPR).

At B2Metric, maintaining user data privacy and security is embedded in our culture. Building data security is a continuous process that shapes the foundation of our development processes and helps to ensure the outstanding performance of our industry-leading technologies. More than 50 leading enterprise brands trust B2Metric with their data. B2metric is preparing to become a GDPR-compliant and ISO/IEC 27001 ISMS certified company, data privacy and security lie at the core of our technology — and our culture: We will announce that to complete the SOC 2 Type 1 audit and are committed to performing further SOC 2 examinations in the future. Also, B2Metric system can be installable with Dockerize and Kubernetes system as an on-premises based solution. Near by, B2Metric Auto-ML has LDAP,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

JWT authentication based login policies.

5. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Describe the quality process planned for the final product. Technologically, which are the potential risks in all the phases of the project (design of the solution, development, testing, deployment…) and indicate mitigation plans to still fulfil the challenge/Theme Challenges and data provider requirements.

As B2Metric, we apply software quality assurance which is process which works parallel to development of our AI based AutoML systems and Fraud platform software. It focuses on improving the process of development of software so that problems can be prevented before they become a major issue. Software Quality Assurance is a kind of Umbrella activity that is applied throughout the software process.
Software Quality Assurance has:

- A quality management approach
- Formal technical reviews
- Multi testing strategy
- Effective software engineering and data scince technology
- Measurement, reporting and A/B, unit testing mechanism

These are the possible risks that can be happen during the projects but we have already find suitable ways to handle these problems within our team.

Have been solved poor communication with a customer

The lack of efficient communication between the parties carries the most severe risk for a fraud detection product. Hopefully, it is possible to prevent undesirable outcomes by asking questions. A request to clarify a specific moment helps to save valuable resources and meet the deadlines. We have already started to contact Almerys via email and online meetings. Also we have a French speaker sales engineer in our team that will remove communication barriers.

Have been solved lack of data problem

If the data it is impossible to precisely estimate the minimum amount of data required for this project's Fraud prevention AI models. Obviously, the very nature of any AI project will influence significantly the amount of data you will need. For this projects we have got raw transaction data from Almerys' team and there is no problem about lack of data risk. Also, our Auto-ML platform's adaptive learning module tests data maturity periodically. That's why, if we scale the project with the large set of healthcare datasets, the B2ML Studio Hunter Auto-ML

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951981

platform identify and generates alert about the dataset maturity.

Model Performance

If you plan on getting a product in production, you need more. A small dataset might be good enough for a proof of concept, but in production, you'll need way more data.

Will solve the risk of frequently changing requirements

Too frequent changes in the requirements can result in a resource gap or exhaustion. It can affect both financial and human factors. Moreover, it puts product quality and meeting deadlines at risk. Incorrect prioritisation Sometimes customers focus on the little things too much, letting significant aspects shift to the background. As a result, a team has to pay too much attention to the secondary features while neglecting the main functionality. It is a good idea to define product highlights early. Still, the core functionality should be a priority. As B2Metric we have finalized more than 60 AI/ML modeling projects for 9 industries, enterprises big data. So, our team knows very well to identify risks and define the priorisation of the projects while planning phase.

Will solve the risk of Incorrect prioritisation

Sometimes customers focus on the little things too much, letting significant aspects shift to the background. As a result, a team has to pay too much attention to the secondary features while neglecting the main functionality. It is a good idea to define product highlights early. Still, the core functionality should be a priority. As B2Metric we have finalized more than 60 AI/ML modelling projects for 9 industries, enterprises big data. So, our team knows very well to identify risks of the projects while project planning phase.

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 951981