Activity 1 – Sentiment Analysis

In this activity, we will:

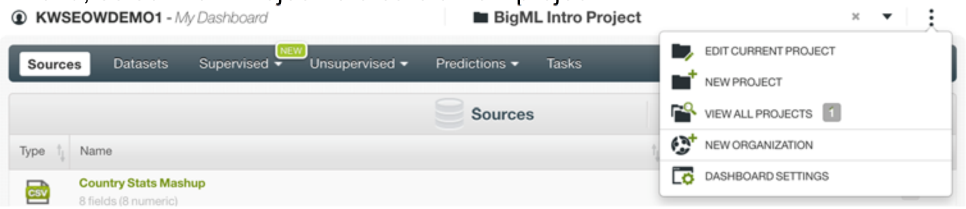
- ❑ Create an account on BigML.com
- ❑ Create Project
- ❑ Update data source and create a dataset
- ❑ Create models for sentiments analysis
- ❑ Evaluate the model
- ❑ Perform predictions

1. Setup account on BigML

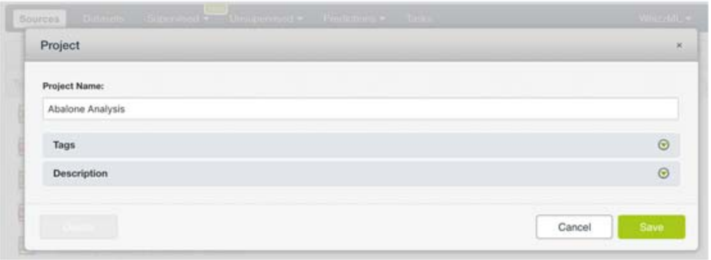
a) Sign up for an account on https://bigml.com. The sign up process is pretty straightforward without the need for credit card information.
* Free account allows for unlimited total tasks and storage, 16 MB dataset size, 2 parallel tasks and 1 user account with no support.

2. Create a new Project

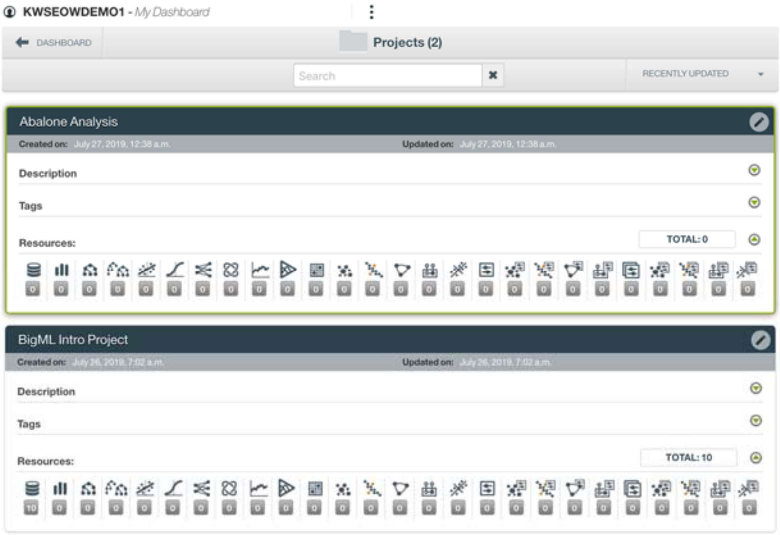
a) On the top right menu, select New Project to create a new project.



b) Enter a project name and click Save.


Project Name: Abalone Analysis

c) For new account, you should see your newly created project and the default **BigML intro Project** in your Dashboard view.

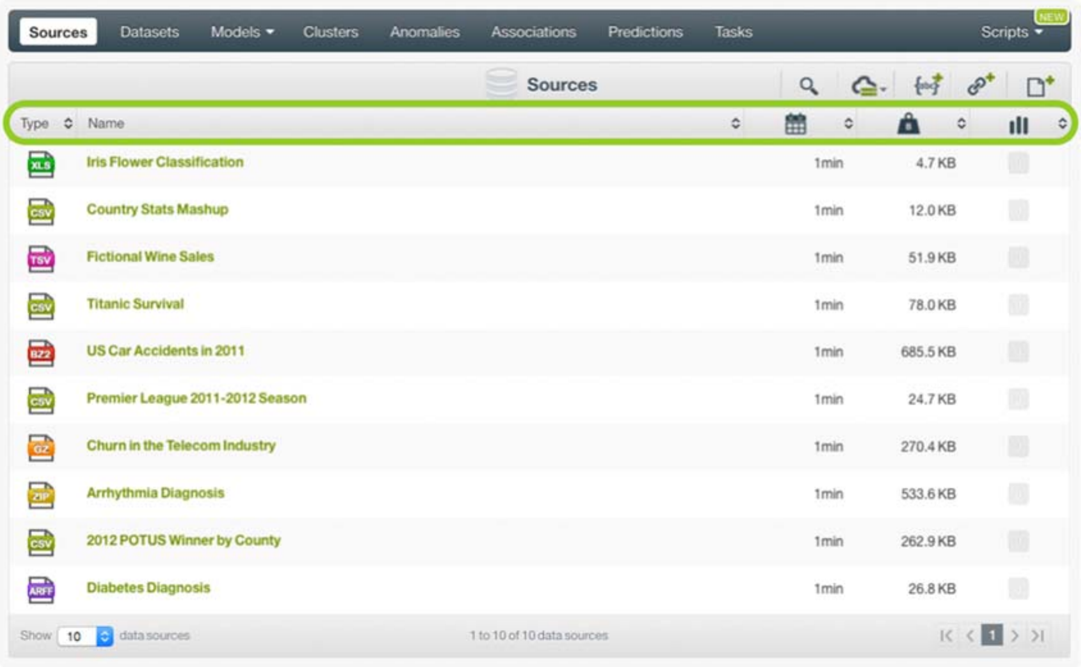

KWSEOWDEMO1 - My Dashboard
DASHBOARD
Projects (2)
Search
RECENTLY UPDATED
Abalone Analysis
Created on: July 27, 2019, 12:38 a.m.
Updated on: July 27, 2019, 12:38 a.m.
Description
Tags
Resources:
TOTAL: 0
BigML Intro Project
Created on: July 26, 2019, 7:02 a.m.
Updated on: July 26, 2019, 7:02 a.m.
Description
Tags
Resources:
TOTAL: 10
10

d) Click on the project name **you used in step 2b** in the dashboard to continue with the next step.

3. Uploading your data

a) The first step is to upload your data to your BigML account. BigML offers several ways to do it, you can drag and drop a local file, connect BigML to your cloud repository (e.g., S3 buckets) or copy and paste a URL. BigML automatically identifies the field types. Field types and other source parameters can alternatively be configured by clicking in the source configuration option.

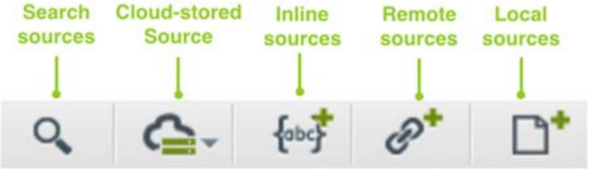
b) The first tab of the BigML Dashboard's main menu allows you to list all your available sources. You will see an empty list for your Abalone Analysis project. However, when you first create an account at BigML, you will find a list of promotional BigML sources in your **BigML intro Project**. In this source list view (see below), you will see for each source, the **Type**, **Name**, **Age** (time since the BigML source was created), **Size**, and **Number of Datasets** that have been created using the BigML source.



Type	Name			
XLS	Iris Flower Classification	1min	4.7 KB	
CSV	Country Stats Mashup	1min	12.0 KB	
TSV	Fictional Wine Sales	1min	51.9 KB	
CSV	Titanic Survival	1min	78.0 KB	
BZ2	US Car Accidents in 2011	1min	685.5 KB	
CSV	Premier League 2011-2012 Season	1min	24.7 KB	
GZ	Churn in the Telecom Industry	1min	270.4 KB	
ZIP	Arrhythmia Diagnosis	1min	533.6 KB	
CSV	2012 POTUS Winner by County	1min	262.9 KB	
ARFF	Diabetes Diagnosis	1min	26.8 KB	

Show 10 data sources 1 to 10 of 10 data sources 1

c) On the top corner of the source list view, you can see the menu options show below.



These menu options perform the following operations (from right to left):

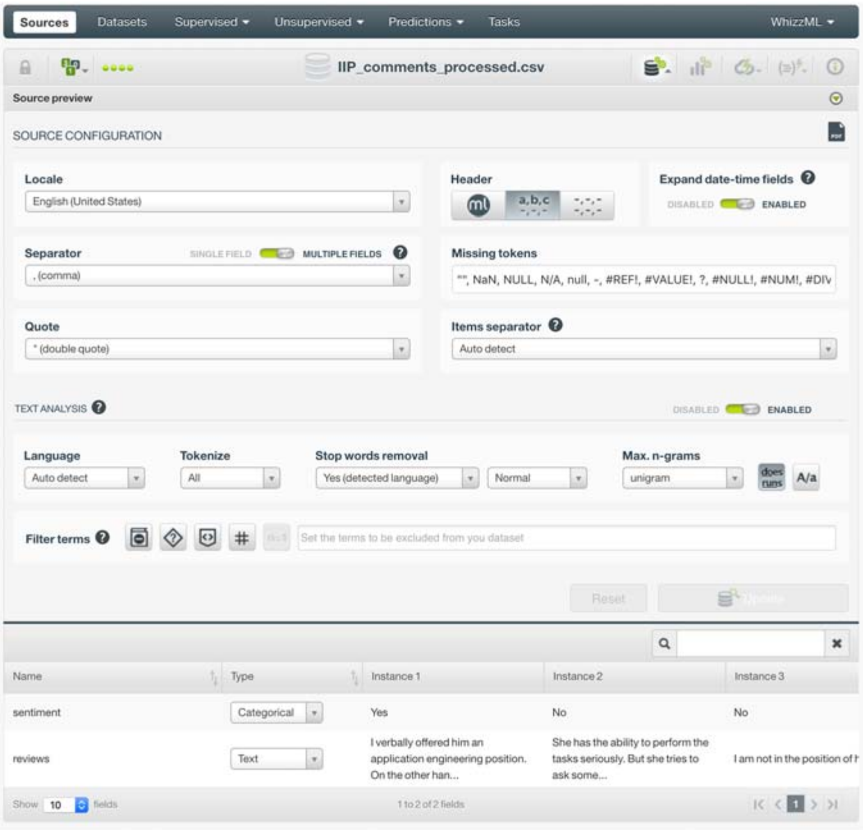
1) **Create a source from a local source** opens a file dialog that helps you browse files in your local drives.
2) **Create a source from a URL** opens a modal window that helps you input the URL of that BigML will use to automatically download a remote source.
3) **Create a inline source** opens an editor where you can directly input or paste data into it.
4) **Cloud Storage DropDown** helps you browse through previously configured cloud storage providers.
5) **Search** searches your sources by name.

d) Download a copy of the dataset from your instructor from the github link provided by your instructor. The dataset we will be using is IIP_comments_processed.csv. The file consist of two columns, ***sentiment*** and ***reviews***.

e) Click on **Create a source from a local source,** in the file selection dialog box, select the dataset csv file and click **Open**. The csv file will be uploaded to BigML and when completed, you should see the csv file listed on your **Sources** tab.

f) In BigML you can configure text parameters like the tokenization strategy, the case sensitivity, stemming or stop words using the Text Analysis options at the source configuration level in the Sources with the BigML Dashboard. These text configuration options define the vocabulary for all BigML models, including topic models. However, topic models are the only models that allow you to change this configuration at the topic model creation time. You can build several topic models using different text configurations without going back to the source and having to create different datasets.

Among the text configuration options for topic models you can find parameters like the language, the tokenization, whether to include or exclude stop words, the maximum size for the n-grams to be considered in your model vocabulary, the stemming, and the case sensitivity. You can also define the terms that you want to completely exclude from your model..



IIP_comments_processed.csv

SOURCE CONFIGURATION

Locale: English (United States)

Header

Expand date-time fields: DISABLED ENABLED

Separator: , (comma) — SINGLE FIELD MULTIPLE FIELDS

Missing tokens: "", NaN, NULL, N/A, null, -, #REF!, #VALUE!, ?, #NULL!, #NUM!, #DIV

Quote: " (double quote)

Items separator: Auto detect

TEXT ANALYSIS — DISABLED ENABLED

Language: Auto detect

Tokenize: All

Stop words removal: Yes (detected language) Normal

Max. n-grams: unigram

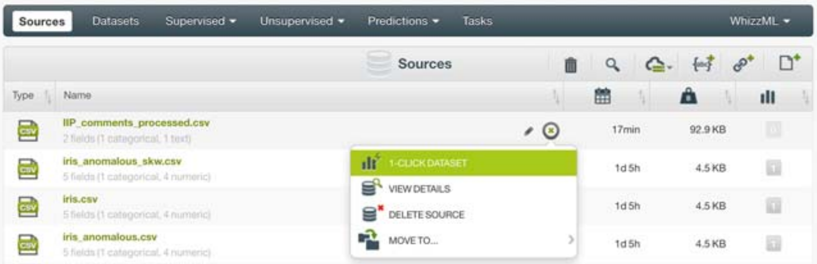
Filter terms: Set the terms to be excluded from you dataset

Name	Type	Instance 1	Instance 2	Instance 3
sentiment	Categorical	Yes	No	No
reviews	Text	I verbally offered him an application engineering position. On the other han...	She has the ability to perform the tasks seriously. But she tries to ask some...	I am not in the position of h

Show 10 fields 1 to 2 of 2 fields

4. Create a Dataset

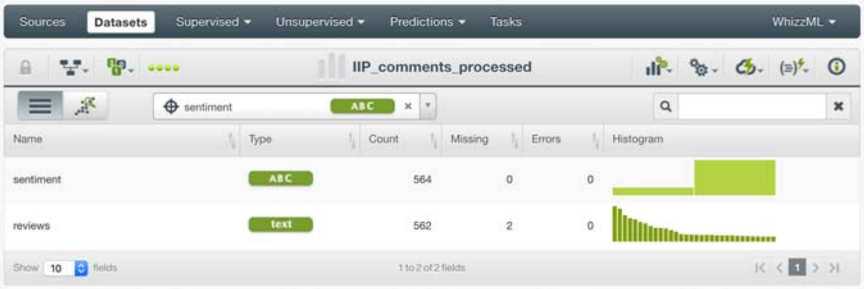
a) From your source view, use the **1-click dataset** option to create a dataset, a structured version of your data ready to be used by a Machine Learning algorithm.



Sources

Type	Name		
CSV	IIP_comments_processed.csv 2 fields (1 categorical, 1 text)	17min	92.9 KB
CSV	iris_anomalous_skw.csv 5 fields (1 categorical, 4 numeric)	1d 5h	4.5 KB
CSV	iris.csv 5 fields (1 categorical, 4 numeric)	1d 5h	4.5 KB
CSV	iris_anomalous.csv 5 fields (1 categorical, 4 numeric)	1d 5h	4.5 KB

b) In the **dataset** tab, click on the newly created dataset, you will be able to see a summary of your field values, univariate statistics, and the field histograms to analyse your data distributions. This view is really useful to see any errors or irregularities in your data. You can also filter the dataset by several criteria and create new fields using different pre-defined operations as needed. However, for our simple reviews dataset, you will not be able to see too much.

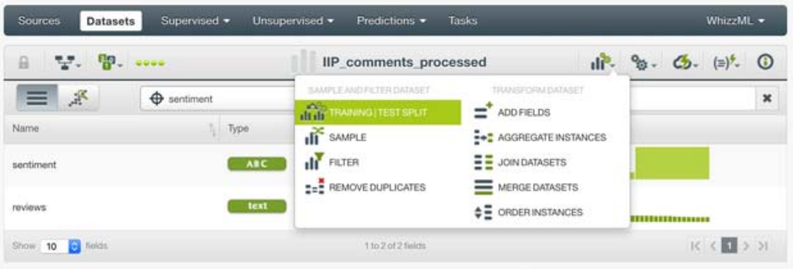


IIP_comments_processed

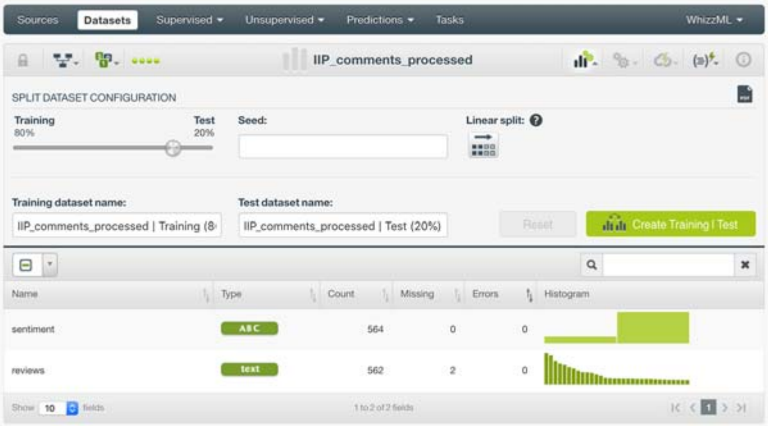
Name	Type	Count	Missing	Errors	Histogram
sentiment	ABC	564	0	0	
reviews	text	562	2	0	

Show 10 fields 1 to 2 of 2 fields

c) Once your data is clean and free of errors you can split your dataset into two different subsets: one for training your model, and the other for testing. It is crucial to train and evaluate your model with different data to ensure it generalizes well against unseen data. You can easily split your dataset using the BigML 1-click option, which randomly sets aside 80% of the instances for training and 20% for testing.



d) Can you change the split ratio. But for our activity, we will just keep it to 80-20 which is the norm. Click on **Create Training | Test**



SPLIT DATASET CONFIGURATION

Training 80% Test 20%

Seed:

Linear split:

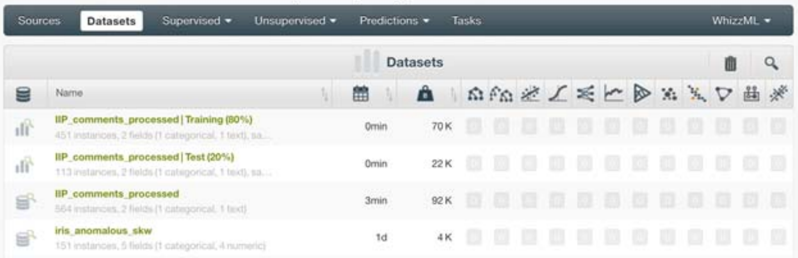
Training dataset name: IIP_comments_processed | Training (8

Test dataset name: IIP_comments_processed | Test (20%)

Name	Type	Count	Missing	Errors	Histogram
sentiment	ABC	564	0	0	
reviews	text	562	2	0	

Show 10 fields 1 to 2 of 2 fields

e) After the creation, return to the Datasets tab, you will see two new datasets created. One with “Training (80%)” appended to the dataset name and one “Test (20%)” appended.

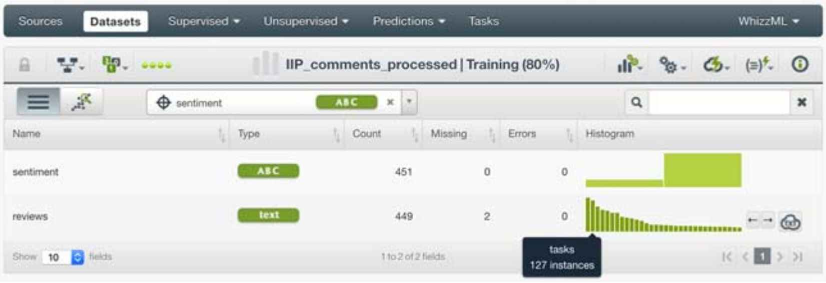


Name		
IIP_comments_processed | Training (80%) 451 instances, 2 fields (1 categorical, 1 text), sa...	0min	70 K
IIP_comments_processed | Test (20%) 113 instances, 2 fields (1 categorical, 1 text), sa...	0min	22 K
IIP_comments_processed 564 instances, 2 fields (1 categorical, 1 text)	3min	92 K
iris_anomalous_skw 151 instances, 5 fields (1 categorical, 4 numeric)	1d	4 K

5. Create a Model (Single Decision Tree)

a) To get a first sense of whether the text in the reviews have any power to predict the sentiment, we are going to build a single decision tree by selecting the “sentiment” as the objective field and using the reviews as input

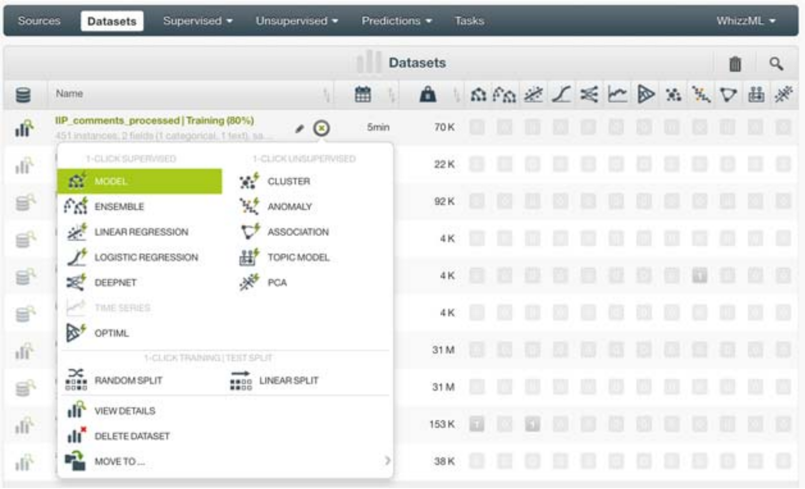
b) When you are done creating your dataset, you can see that the reviews dataset is composed of two fields: sentiment (Yes/positive or No/negative) and reviews (the review text). Not surprisingly, the words “learn”, “tasks” and “attitude” are the most frequent ones in the collection (see image below). Mouse over the histogram to see the summary.



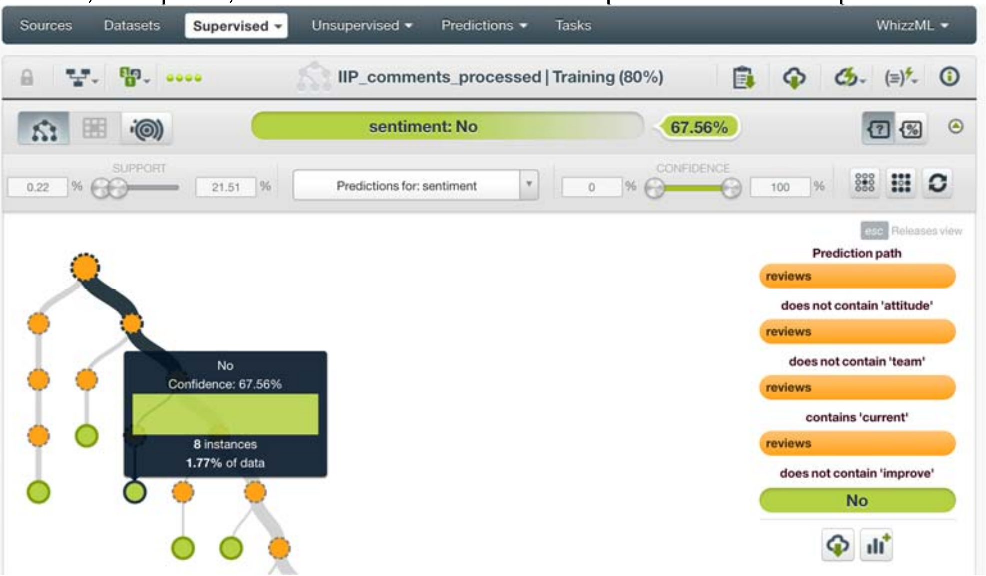
Name	Type	Count	Missing	Errors	Histogram
sentiment	ABC	451	0	0	
reviews	text	449	2	0	

Show 10 fields 1 to 2 of 2 fields

c) Perform a 1-click model by clicking on the dropdown menu icon and select **MODELS**.



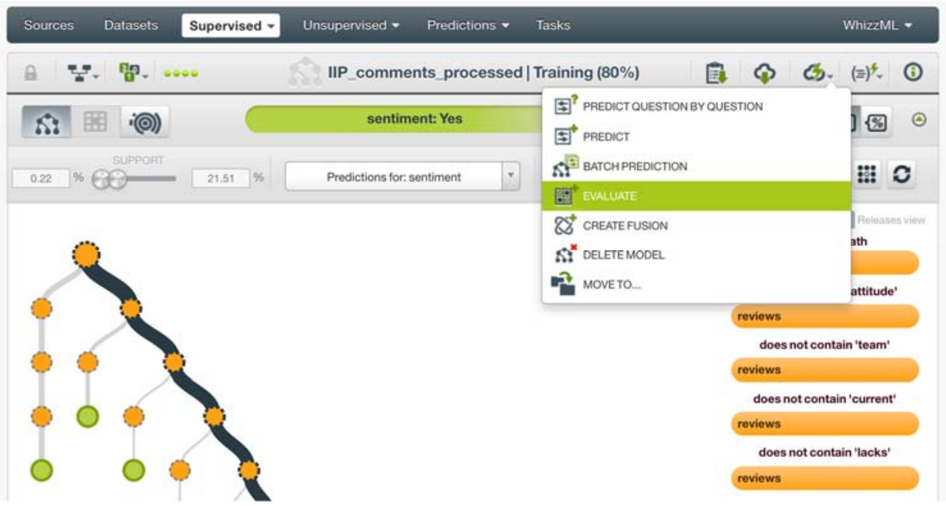
d) Once completed (it should take less then a minute), you can mouse over the root node and go down the tree until you reach the leaf nodes, you can see the different prediction paths. For example, in the image below you can see that if the review does not contains the terms “attitude”, “team”, “improve” and does contain “current” then it is a “negative” review with 67.56% of confidence. As expected, we can find words in the nodes such as “lacks”, “fast”, “learn”, “complete”, which are the terms that best split the data to best predict sentiment.



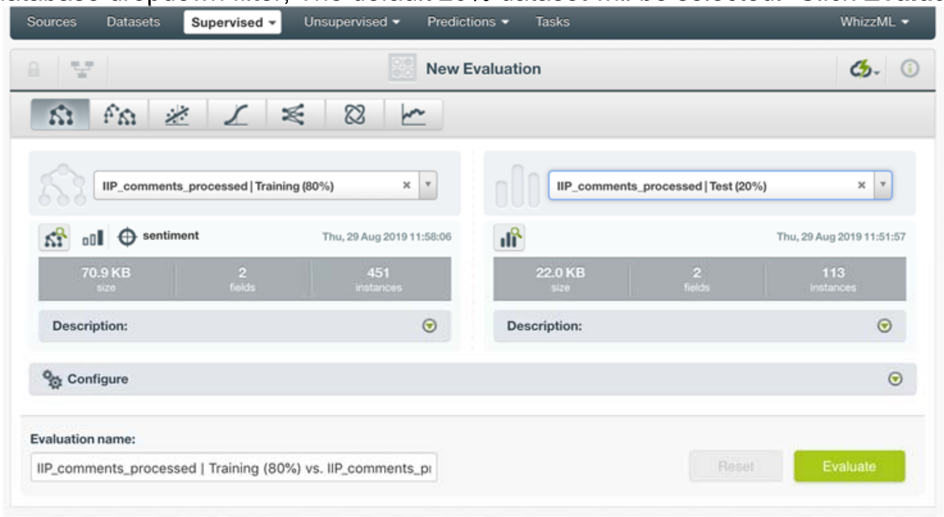
6. Evaluate the Model

Confidence values seem pretty high for most prediction paths in this tree, but to measure its predictive power we need to evaluate it by using data it has not seen before. We use the previously set aside 50% of the dataset, which contains the remaining 25,000 movie reviews that have not been used to train our model.

a) Navigate to the detail view of the model. Click on **Supervised**, select **MODELS** and click on the model you created. In the 1-click menu, select **EVALUATION**.



b) In the search database dropdown filter, The default 20% dataset will be selected. Click **Evaluate**.



IIP_comments_processed | Training (80%)

sentiment Thu, 29 Aug 2019 11:58:06

70.9 KB size	2 fields	451 instances

Description:

IIP_comments_processed | Test (20%)

Thu, 29 Aug 2019 11:51:57

22.0 KB size	2 fields	113 instances

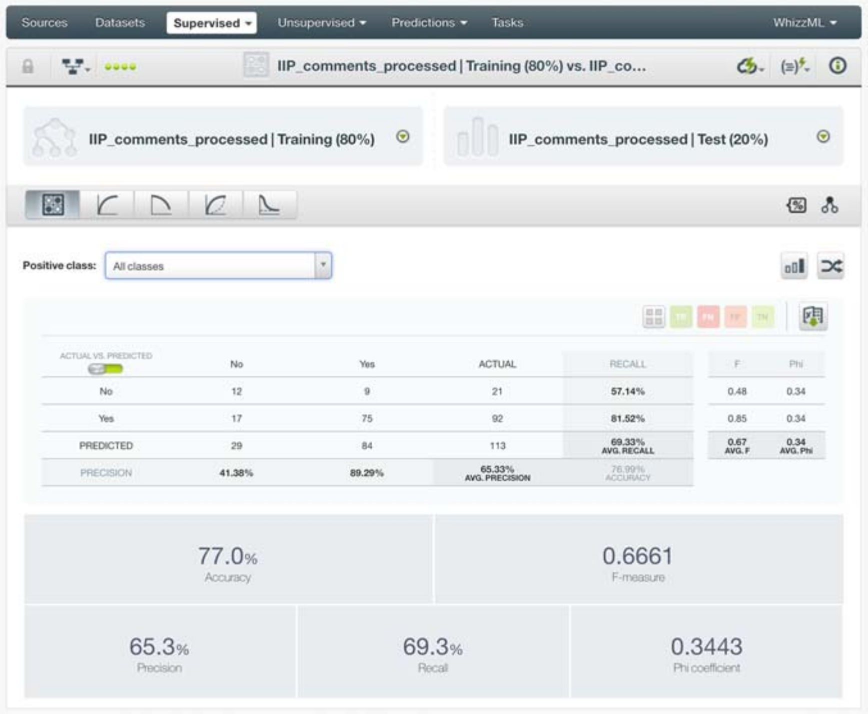
Description:

Configure

Evaluation name:

IIP_comments_processed | Training (80%) vs. IIP_comments_pı

c) The evaluation yields an overall accuracy of 77.0%, which is not that bad for a 1-click model. However, we can definitely improve on this! (do bear in mind that the dataset is imbalance too)



IIP_comments_processed | Training (80%)

IIP_comments_processed | Test (20%)

Positive class: All classes

ACTUAL VS. PREDICTED	No	Yes	ACTUAL	RECALL	F	Phi
No	12	9	21	57.14%	0.48	0.34
Yes	17	75	92	81.52%	0.85	0.34
PREDICTED	29	84	113	69.33% AVG. RECALL	0.67 AVG. F	0.34 AVG. Phi
PRECISION	41.38%	89.29%	65.33% AVG. PRECISION	76.99% ACCURACY		

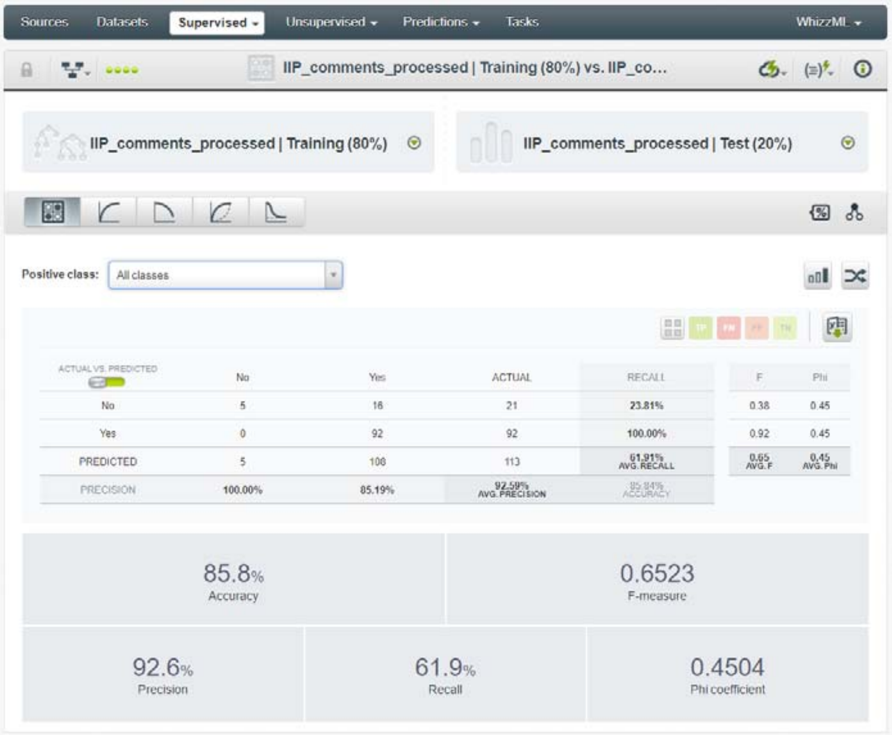
77.0% Accuracy	0.6661 F-measure

65.3% Precision	69.3% Recall	0.3443 Phi coefficient

7. Ensemble

Go ahead to try out **Ensemble 1 click model** and check the accuracy you can get. It may take a long time 1-2 hours to build the model.

On my system, the accuracy of the default ensemble model achieves average accuracy of 85.8%.


Sources
Datasets
Supervised
Unsupervised
Predictions
Tasks
WhizzML
IIP_comments_processed | Training (80%) vs. IIP_co...
IIP_comments_processed | Training (80%)
IIP_comments_processed | Test (20%)

Positive class: All classes

ACTUAL VS. PREDICTED	No	Yes	ACTUAL	RECALL	F	Phi
No	5	16	21	23.81%	0.38	0.45
Yes	0	92	92	100.00%	0.92	0.45
PREDICTED	5	108	113	61.91% AVG. RECALL	0.65 AVG. F	0.45 AVG. Phi
PRECISION	100.00%	85.19%	92.59% AVG. PRECISION	85.84% ACCURACY		

85.8% Accuracy

0.6523 F-measure

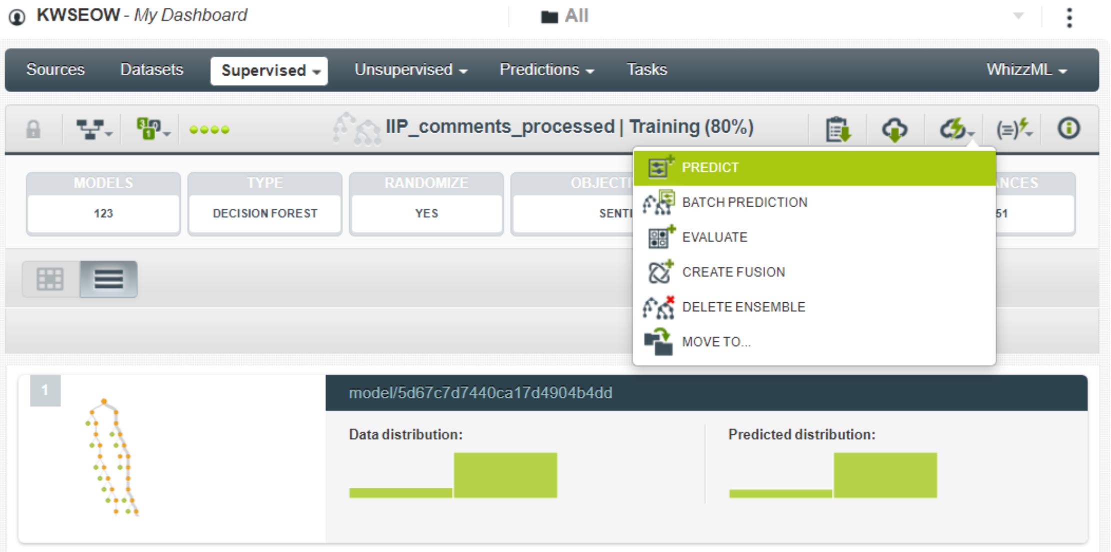
92.6% Precision

61.9% Recall

0.4504 Phi coefficient

8. Make Predictions

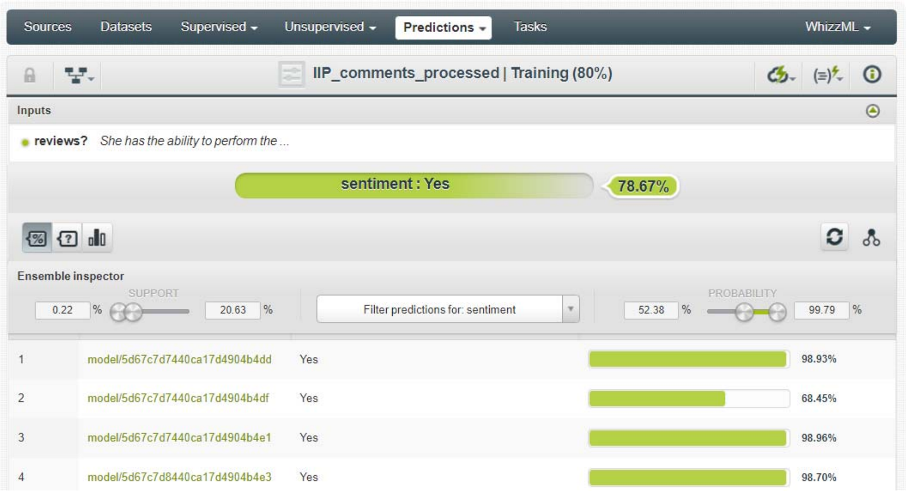
a) Once you are happy with the model, you can perform a prediction. Click on **Predict**.



MODELS 123 | TYPE DECISION FOREST | RANDOMIZE YES | OBJECTI… SENTI… | …NCES 51

model/5d67c7d7440ca17d4904b4dd

b) A form containing all your input fields will be displayed and you will be able to set the values for a new instance. At the top of the view, you will see the objective field prediction changing as you change your input field values, in our case, the reviews.



IIP_comments_processed | Training (80%)

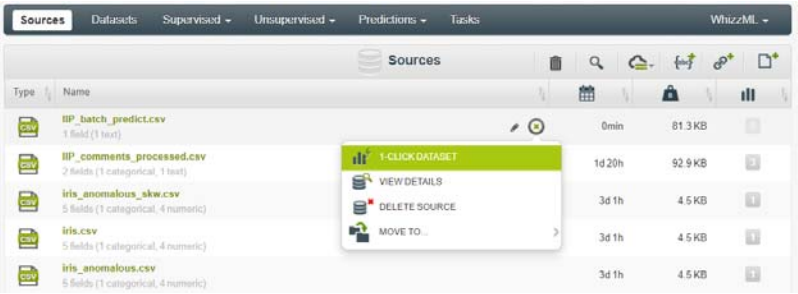
reviews? *She has the ability to perform the ...*

sentiment : Yes 78.67%

0.22 % – 20.63 % | Filter predictions for: sentiment | 52.38 % – 99.79 %

1	model/5d67c7d7440ca17d4904b4dd	Yes	98.93%
2	model/5d67c7d7440ca17d4904b4df	Yes	68.45%
3	model/5d67c7d7440ca17d4904b4e1	Yes	98.96%
4	model/5d67c7d8440ca17d4904b4e3	Yes	98.70%

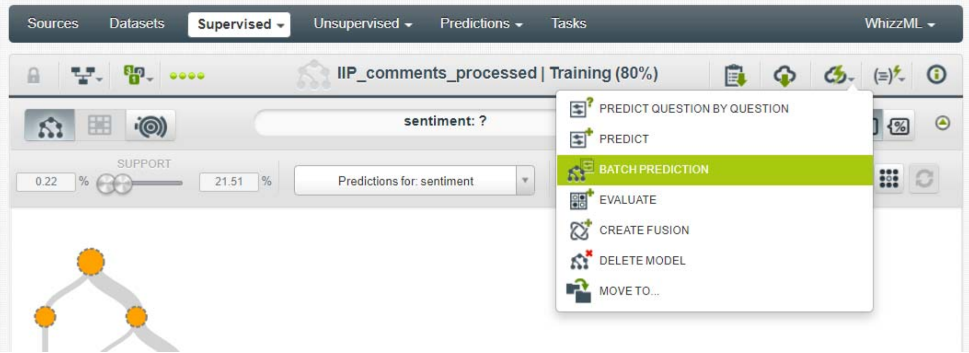
c) **Batch predictions** - Use the Batch Prediction option in the 1-click menu and select the dataset containing the instances for which you want to know the objective field value. However, before you do this, create a new source using the IIP_batch_predict.csv and then create a new dataset from this source



Name		
IIP_batch_predict.csv 1 field (1 text)	0min	81.3 KB
IIP_comments_processed.csv 2 fields (1 categorical, 1 text)	1d 20h	92.9 KB
iris_anomalous_skw.csv 5 fields (1 categorical, 4 numeric)	3d 1h	4.5 KB
iris.csv 5 fields (1 categorical, 4 numeric)	3d 1h	4.5 KB
iris_anomalous.csv 5 fields (1 categorical, 4 numeric)	3d 1h	4.5 KB

d) You can configure several parameters of your batch prediction such as the option to include both confidence interval and prediction interval in the batch prediction output dataset and file. When your batch prediction finishes you will be able to download the CSV file and see the output dataset.

Select the model you want to use for the batch prediction. In our case, click on **Supervised**, **Models**, **Batch Predictions**.

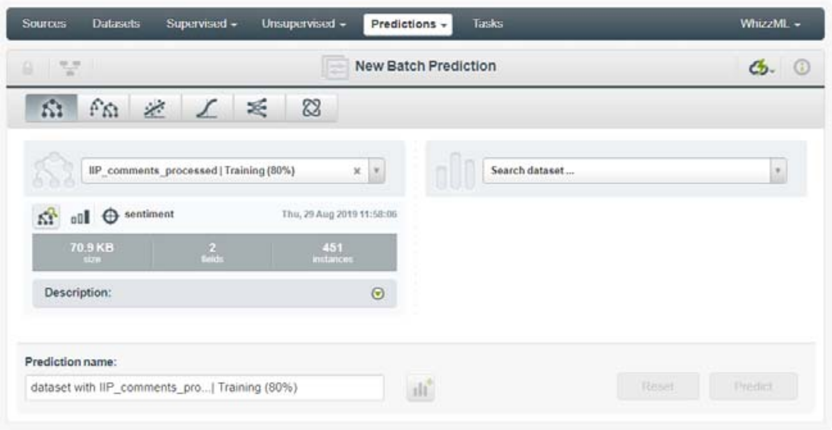


IIP_comments_processed | Training (80%)

sentiment: ?

0.22 % – 21.51 % | Predictions for: sentiment

e) In the **Search Dataset…** field, search and select your dataset for batch prediction.



New Batch Prediction

IIP_comments_processed | Training (80%)

Search dataset ...

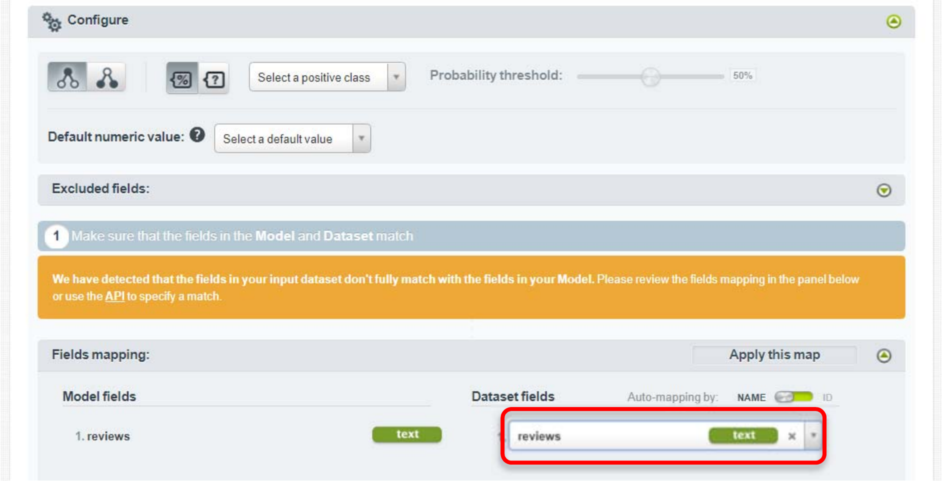
sentiment Thu, 29 Aug 2019 11:58:06

70.9 KB size	2 fields	451 instances

Description:

Prediction name: dataset with IIP_comments_pro...| Training (80%)

f) Check that the model and dataset fields matched. Select **reviews** in the Dataset fields.



Probability threshold: 50%

Default numeric value: Select a default value

Excluded fields:

1 Make sure that the fields in the **Model** and **Dataset** match

We have detected that the fields in your input dataset don't fully match with the fields in your Model. Please review the fields mapping in the panel below or use the API to specify a match.

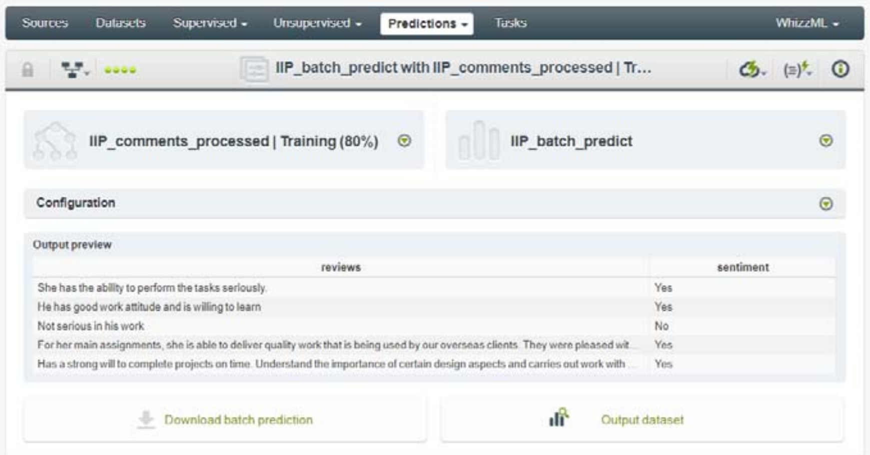
Fields mapping:

Model fields: 1. reviews text

Dataset fields (Auto-mapping by: NAME ID): reviews text

g) Click on **Predict** to start the prediction. Usually it will just take a moment to complete.

h) Once the prediction is done, you can check out the result on the **Output preview** section or **download batch prediction** if you want to keep a copy of all the prediction.



IIP_batch_predict with IIP_comments_processed | Tr...

IIP_comments_processed | Training (80%)

IIP_batch_predict

Configuration

Output preview

reviews	sentiment
She has the ability to perform the tasks seriously.	Yes
He has good work attitude and is willing to learn	Yes
Not serious in his work	No
For her main assignments, she is able to deliver quality work that is being used by our overseas clients. They were pleased wit...	Yes
Has a strong will to complete projects on time. Understand the importance of certain design aspects and carries out work with ...	Yes

Activity wrap-up:

We learn how to:

- ❑ Create an account on BigML.com
- ❑ Create Project
- ❑ Update data source and create a dataset
- ❑ Create models for sentiments analysis
- ❑ Evaluate the model
- ❑ Perform predictions