Anomaly Detection with the BigML Dashboard

The BigML Team

Version 1.2

MACHINE LEARNING MADE BEAUTIFULLY SIMPLE

info@bigml.com

Last updated January 29, 2024

About this Document

This document provides a comprehensive description of how to solve anomaly detection tasks using the BigML Dashboard. Learn how to use the BigML Dashboard to configure, visualize, and interpret this unsupervised model and use it to calculate anomaly scores for single instances, and whole datasets as well.

This document assumes that you are familiar with:

- Sources with the BigML Dashboard. The BigML Team. June 2016. [7]
- Datasets with the BigML Dashboard. The BigML Team. June 2016. [6]

To learn how to use the BigML Dashboard to build supervised predictive models read:

- Classification and Regression with the BigML Dashboard. The BigML Team. June 2016. [4]
- Time Series with the BigML Dashboard. The BigML Team. July 2017. [8]

To learn how to use the BigML Dashboard to build other unsupervised models read:

- Cluster Analysis with the BigML Dashboard. The BigML Team. June 2016. [5]
- Anomaly Detection with the BigML Dashboard. The BigML Team. June 2016. [2]
- Association Discovery with the BigML Dashboard. The BigML Team. June 2016. [3]
- Topic Modeling with the BigML Dashboard. The BigML Team. November 2016. [9]

Contents

CHAPTER 1

Introduction

There are problems that require identifying the instances within a dataset that do not conform to a regular pattern, e.g., detecting any kind of fraud, or discovering errors in your data. BigML anomaly detector (called anomaly in BigML) is an unsupervised learning method that is capable to detect anomalous instances in unlabeled datasets. This means that you do not need to collect a training dataset knowing in advanced which instances are anomalous and which are normal. The algorithm can find suspicious patterns in your data given a set of input fields.

BigML anomaly is an optimized implementation of the Isolation Forest algorithm, a highly scalable method that can efficiently deal with high-dimensional datasets. Learn more about Isolation Forests in Chapter 2.

This chapter provides a comprehensive description of BigML anomalies, including how they can be created with 1-click (Chapter 3), all the configuration options (Chapter 4), and the visualization provided by BigML (Chapter 5). For each instance, BigML computes an anomaly score, which can take values between 0% and 100%, and the field importances, an indicator of each field contribution to the anomaly score (Section 2.1). Once your anomaly detector has been created, you can use it to score new instances one by one or in batch (Chapter 6). You can even download the anomaly score to score new instances locally (see Section 7.1). You can also create, configure, retrieve, list, update, delete, and use your anomaly detector for making scoring predictions using the BigML API and bindings (Section 7.2 and Section 7.3).

The fifth tab of the main menu of the BigML Dashboard allows you to list all your available anomalies. The anomaly list view (Figure 1.1), shows the dataset used to create each anomaly, the **Name**, **Top N** (the top number of anomalies explained in Section 4.1), **Age** (time elapsed since it was created), **Size**, and number of **Scores** and **Batch Scores** that have been created using that anomaly. The SEARCH menu option in the top right menu of the anomaly list view allows you to **search** your anomalies by name.

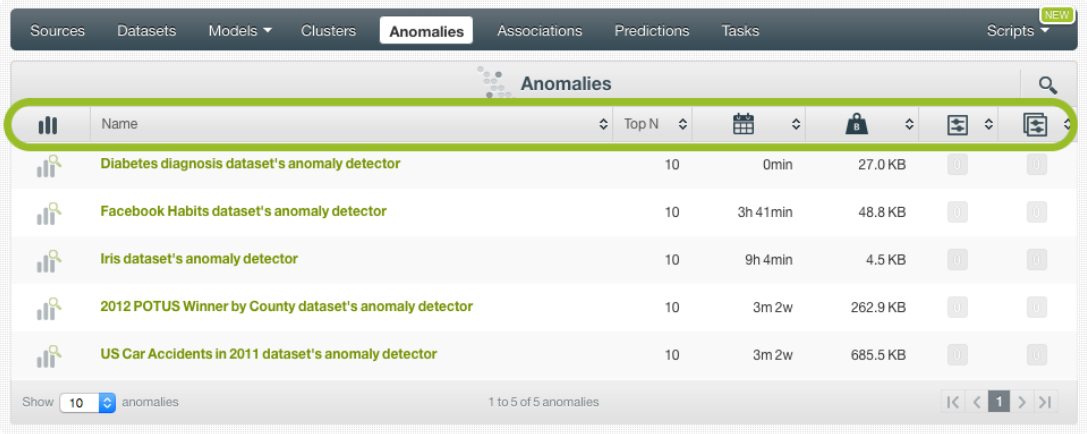

Figure 1.1: Anomaly list view

When you first create an account at BigML, or every time that you start a new project, your list of anomalies will be empty. (See Figure 1.2.)

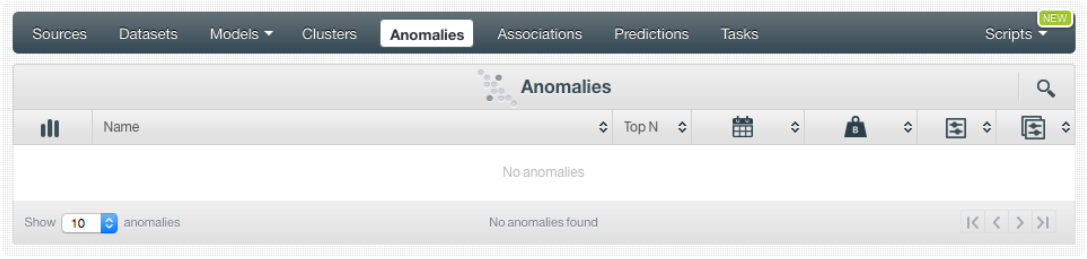

Figure 1.2: Anomaly empty list view in the BigML Dashboard

Finally, in Figure 1.3 you can see the icon used to represent an anomaly in BigML.

Figure 1.3: Anomalies icon

CHAPTER 2

Understanding Anomalies

This chapter describes internal details about the BigML anomalies, providing the foundations to understand the configuration options to create an anomaly detector.

Anomaly detection tasks try to find data points in a dataset following patterns that significantly differ from the rest of the instances. To achieve this, BigML uses a state-of-the-art algorithm called **Isolation Forest**, explained in the following section (Section 2.1). An **anomaly score** is calculated for each of the anomalous instances along with an indicator of each input field contribution, known as **field importance**.

An advantage of this method, is that BigML anomalies can support categorical and numeric fields as well as missing values as **input data** (explained in Section 2.2).

At the end of this chapter, you can find an example illustrating how to **interpret anomalies** in BigML (Section 2.3).

2.1 Isolation Forest

BigML anomaly detector is an optimized implementation of the Isolation Forest [1] algorithm to help users detect anomalies in their datasets. The basic idea is that anomalous instances are more susceptible to be **isolated** than normal instances when using a decision tree approach. Therefore, BigML builds an ensemble that deliberately overfits each single tree to isolate each instance from the rest of the data points. Each tree is built by selecting a random feature and a random split, then the space is recursively partitioned randomly until single instances are isolated. Anomalous instances should take less partitions to isolate than normal data points. Figure 2.1 illustrates how anomalous instances can be isolated with less splits than normal instances. In other words, the closer to the tree root in a single tree, the more anomalous the instance.

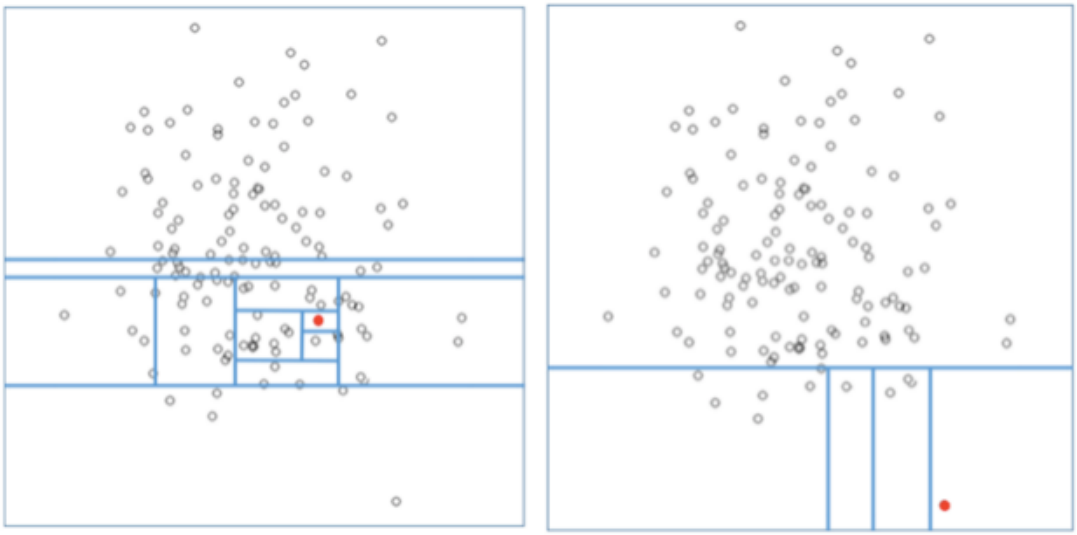

Figure 2.1: Graphic representation example of a normal data point (left) versus an anomalous data point (right)

When all instances have been isolated, BigML automatically calculates an **anomaly score** by averaging the **number of splits** needed to isolate an instance across trees in the ensemble. Lower number of splits will result in higher scores. Then these averages are normalized to get a final score that can take values between 0% and 100%. This score measures how anomalous an instance is, e.g., the red data point on the left in Figure 2.1 took 10 partitions to isolate, while the one on the right took only 4, so the one on the right will have a higher anomaly score.

BigML also calculates the input **field importances** for each anomalous instance, that can be defined as the contribution of each field to the anomaly score. Field importances are calculated by finding the per-field sums of the instances partitioned by each split during an evaluation of an Isolation Forest. BigML normalizes these sums, and this yields a percentage per field ranging from 0% and 100% that measures its relative contribution to the anomaly score.

No distance metric is needed to detect anomalous instances in BigML. Empirical comparisons between Isolation Forest and other distance-based methods have demonstrated that detecting anomalies based on isolation techniques **perform significantly better,** especially for high-dimensional datasets.

Isolation Forests have several **advantages**:

- It is a **highly scalable** method that can deal with large and high-dimensional datasets.
- No distance metric is required, which makes anomaly detection much more **efficient** in terms of computational costs.
- There is no need for data rescaling since it does not calculate distances.
- It can handle **missing** data and **categorical** fields. (See Section 2.2.)
- It is very **robust to noise**, i.e. can handle irrelevant or redundant fields since it uses an ensemble of decision trees.
- The **contribution** of each field to the anomaly can be easily computed, as opposed to a black-box model.
- It is almost a **parameter-free** method contributing to ease of use and reduction in performance tuning efforts.

2.2 Input Data for Anomalies

BigML anomalies support **categorical** and **numeric** fields as input fields. If the dataset contains **text** and **items** fields, they will be automatically included in the model as ID fields, but they will not be used

to **compute** the anomalies. (See Section 4.4.)

Moreover, anomalies include **missing values** as valid values in the computation by default. Consequently, instances with missing values may appear among the top anomalies if anomalous patterns can be learned from missing data. By using the BigML API[1] you can replace missing values so they are not considered by the model.

2.3 Interpreting BigML Anomalies

BigML displays a list of the top anomalous instances ranked by score. Usually a **score of 60%** or higher is a good rule of thumb for a given instance to be considered anomalous. BigML also provides the **field importances** for each top anomaly.

Figure 2.2 shows an anomaly view example created from a dataset containing some diabetes patient data. The first instance can be considered **anomalous** since its score is **higher than 60%**. If you mouse over the **field importance** histogram under the orange score bar, you can see that "Diabetes pedigree" is the field that contributes the most to the anomaly score, more than 25%.

If you further **inspect the field values** for this instance in the data inspector to the right, you find that this patient has very high values for "Diabetes pedigree", an indicator of diabetes history among the family members, as well as for "Glucose" and "Pregnancies", three fields that tend to be positively correlated with diabetes. But for this patient, "Diabetes" is false, so the algorithm rightfully points that this pattern represents an anomaly. Of course, this may be simply due to a data entry error, or it could be genuinely a personal anomaly. Regardless, it is a data point unlike the majority of data points in this dataset.

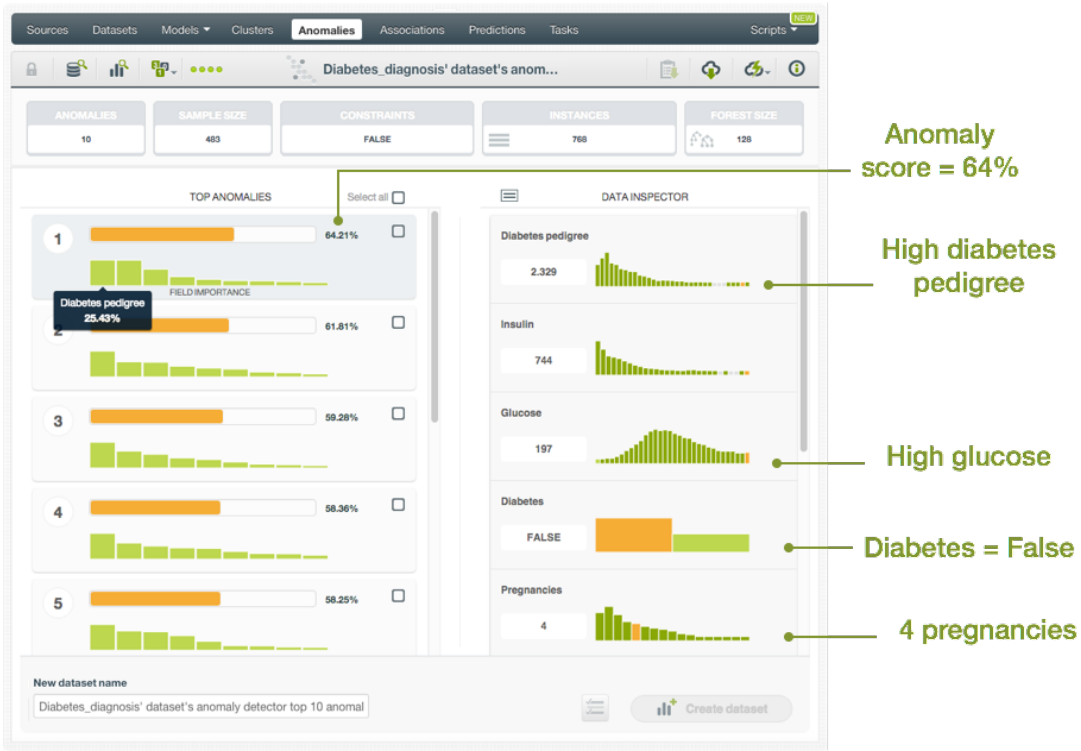

Figure 2.2: Anomaly example

Note: the 60% threshold is no longer valid if the parameter Constraints is enabled since scores tend to be inflated. (See Section 4.3.)

[1] https://bigml.com/api/anomalies#an_create_default_numeric_value

2.4 Anomalies with Images

Image obviously is one of the most important categories among all data, and its presence is ever increasing. It is estimated that more than 85% of all Internet traffic today are visual data. Research also indicates that 90% of the information transmitted to human brain is visual. Therefore it's very important to support and develop machine learning with images.

BigML extracts image features at the source level. Image features are sets of numeric fields for each image. They can capture parts or patterns of an image, such as edges, colors and textures. BigML also supports image features extracted by pre-trained CNNs which capture more sophisticated features. Depending on different machine learning use cases and goals, all these image features can be effective in anomaly detection, as well as other unsupervised and supervised models.

For information about the image features, please refer to section Image Analysis of the Sources with the BigML Dashboard[2][7].

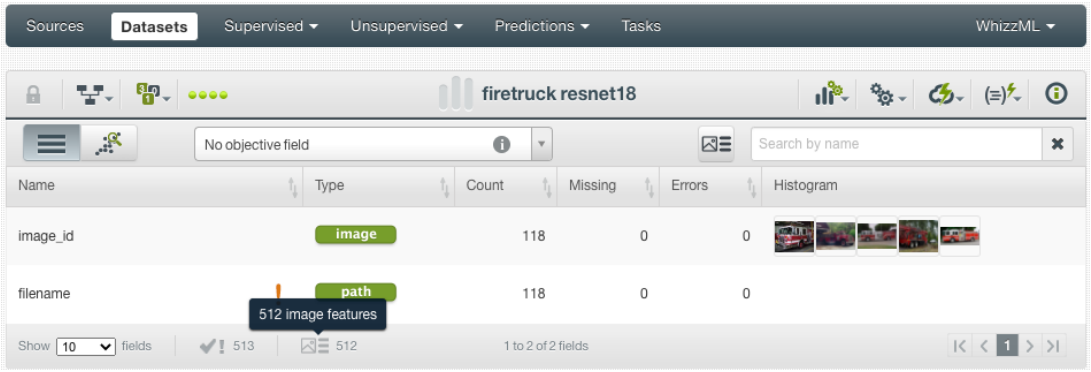

Figure 2.3: A dataset with images and image features

As shown in Figure 2.3, the example dataset has an image field *image_id*. It also has image features extracted from the images referenced by *image_id*. Image feature fields are hidden by default to reduce clutter. To show them, click on the icon "Click to show image features", which is next to the "Search by name" box. In Figure 2.4, the example dataset has 512 image feature fields, extracted by a pre-trained CNN, *ResNet-18*.

[2] https://static.bigml.com/pdf/BigML_Sources.pdf

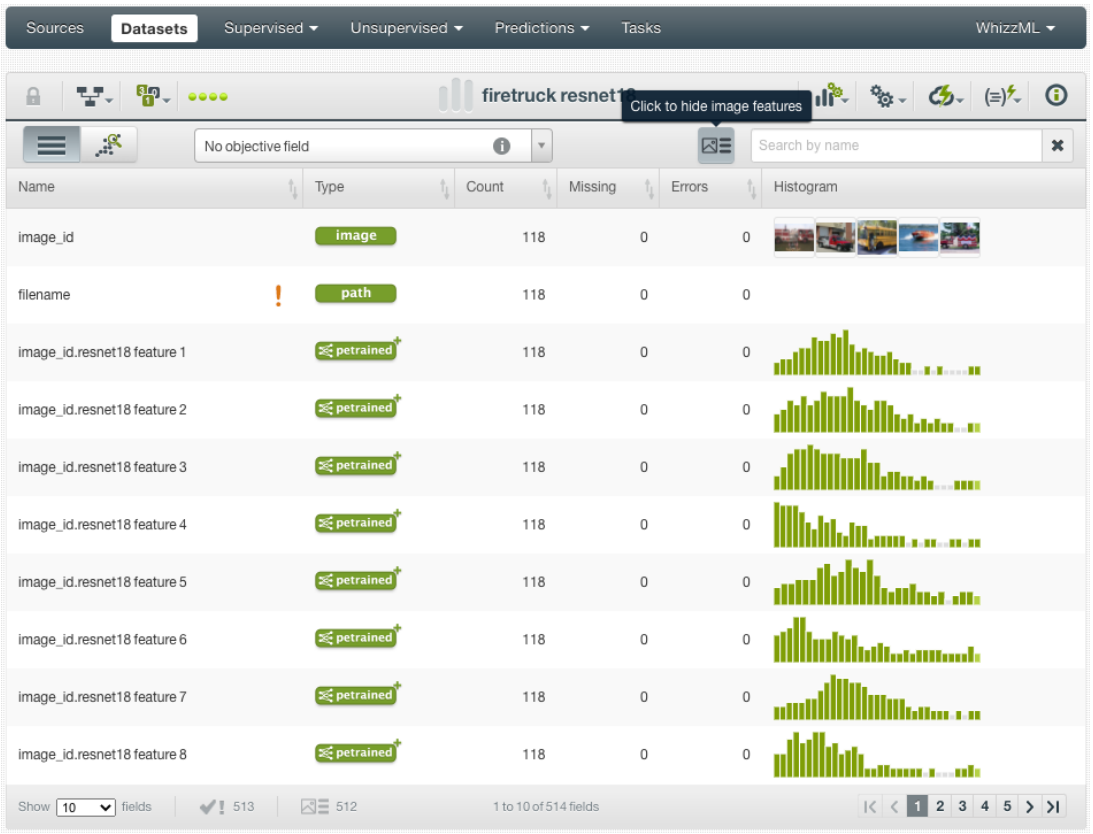

Figure 2.4: A dataset with image feature fields shown

From image datasets like this, anomalies can be created and configured using the steps described in the following chapters.

CHAPTER 3

Creating Anomalies with 1-Click

To create an anomaly in BigML you have two options: you can use the **1-click option** which uses the default values for all available configuration options, or you can tune the parameters in advance using the **configuration option** explained in Chapter 4.

You can find the 1-CLICK ANOMALY option in the **1-click action menu** from the dataset view. (See Figure 3.1.)

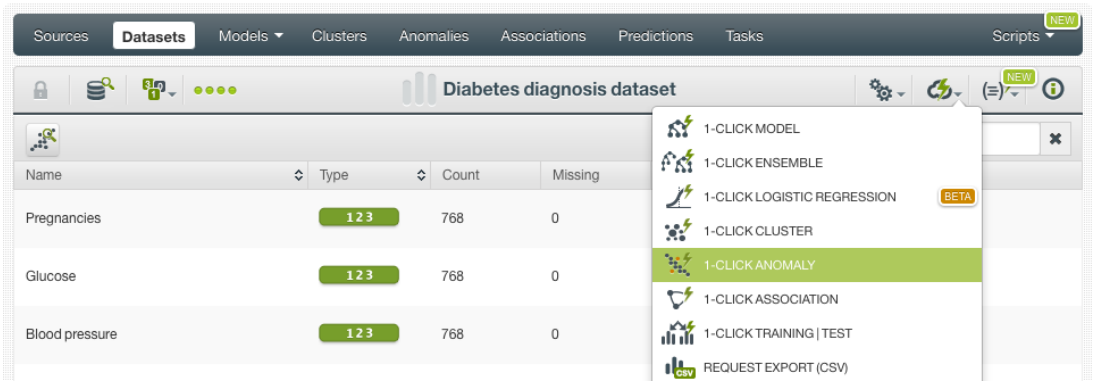

Figure 3.1: Create anomaly detector from 1-click action menu

Alternatively, you can use the 1-CLICK ANOMALY option in the **pop up menu** from the dataset list view. (See Figure 3.2.)

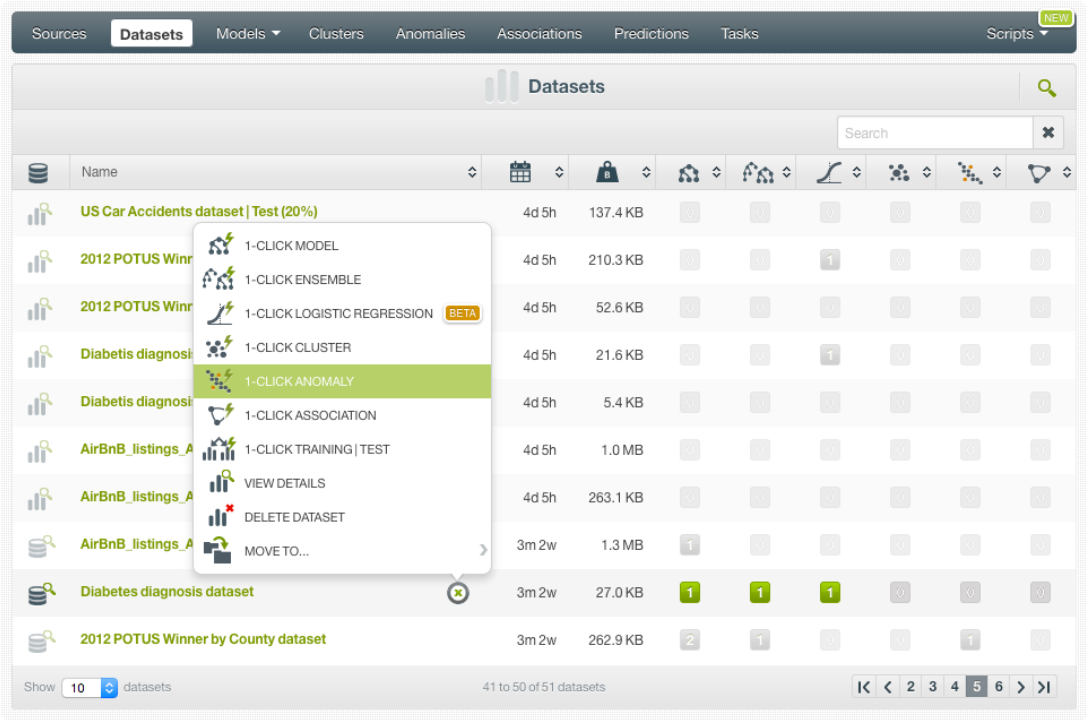

Figure 3.2: Create anomaly from pop up menu

Either option builds an anomaly detector using the default values for the available configuration options explained in the following section. (See Chapter 4.)

CHAPTER 4

Anomaly Configuration Options

You can configure a number of parameters that affect the way BigML creates anomalies. See the following sections for a detailed explanation.

To display the configuration panel to see all options, click the CONFIGURE ANOMALY menu option in the **configuration menu** from the dataset view. (See Figure 4.1.)

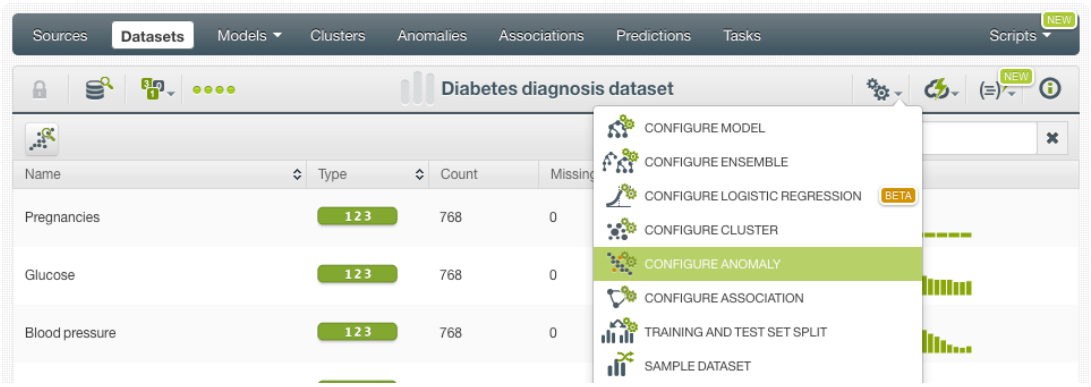

Figure 4.1: Configure anomalies

4.1 Number of Anomalies

You can specify the top number of anomalous instances to be displayed in the anomaly view. (See Figure 4.2.) You can request up to 1,024 anomalies. By default, you get the top 10 anomalous instances.

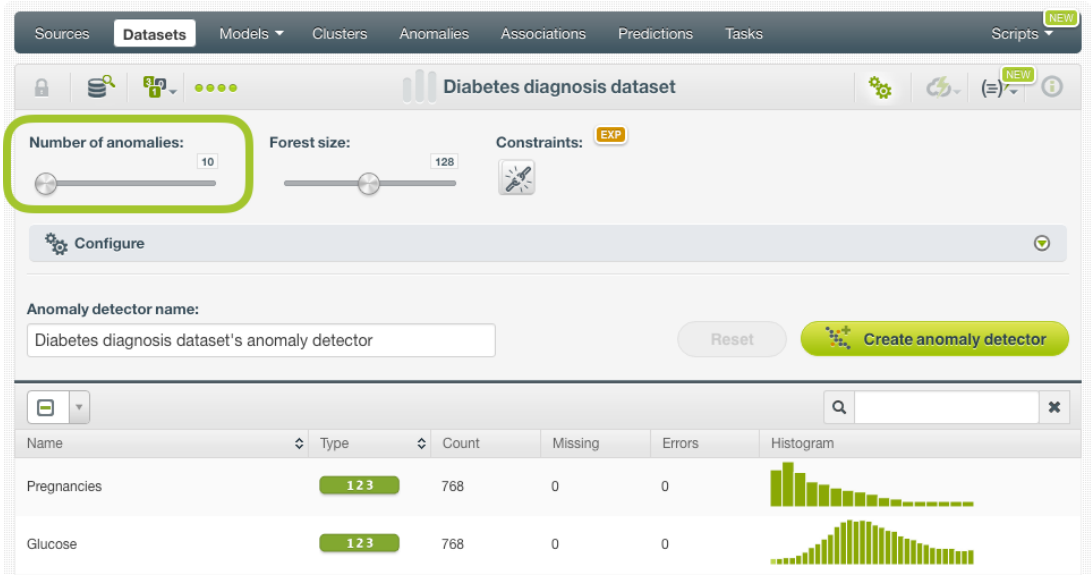

Figure 4.2: Top number of anomalies

4.2 Forest Size

As explained in Chapter 2, BigML anomalies use the Isolation Forest algorithm which is an ensemble of decision trees to detect anomalies. The **Forest size** parameter allows you to configure the number of decision trees composing the ensemble. This must be a number up to 256 (1,000 if you are using the BigML API). Higher numbers tend to give better results although they take longer to process. By default the number of models is 128.

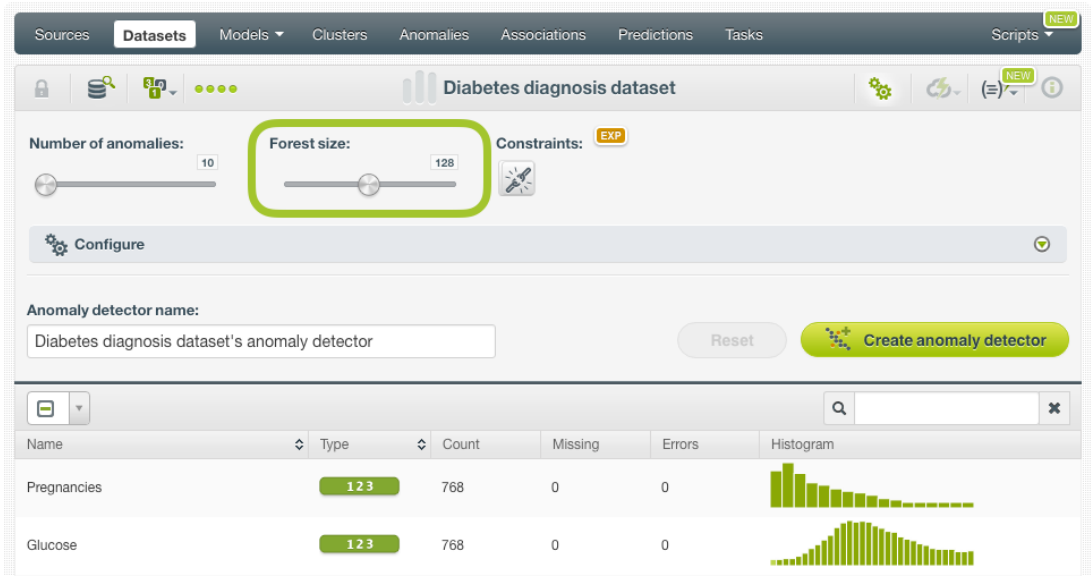

Figure 4.3: Forest size configuration

4.3 Constraints

Constraints parameter is an experimental option that makes Isolation Forest trees more sensitive to anomalous data. Constraints add more predicates to a node split when building the trees for the Isolation Forest, so an instance gets isolated earlier, thus anomaly scores are higher.

For example, in a normal situation, if constraints are disabled, each tree split yields two branches with

one predicate each:

- `Monthly-salary > $2,000`
- `Monthly-salary =< $2,000`

By contrast, if constraints are enabled, each branch will have extra predicates picked randomly:

- `Monthly-salary > $2,000 AND Occupation=employed`
- `Monthly-salary =< $2,000 AND Occupation=student OR Occupation=employed`

If one instance has a `Monthly-salary=4,000` and `Occupation=student`, in the first case, it meets the rule `Monthly-salary > $2,000`, so at least another split is needed in order to isolate the instance. However, in the second case, it will not meet either branch rule, so it will be isolated faster than in the first case, hence its anomaly score will be higher.

This option tends to inflate the anomaly scores and it is more costly in terms of computational costs, but it can also make the trees more effective at flagging anomalous data, especially with categorical data.

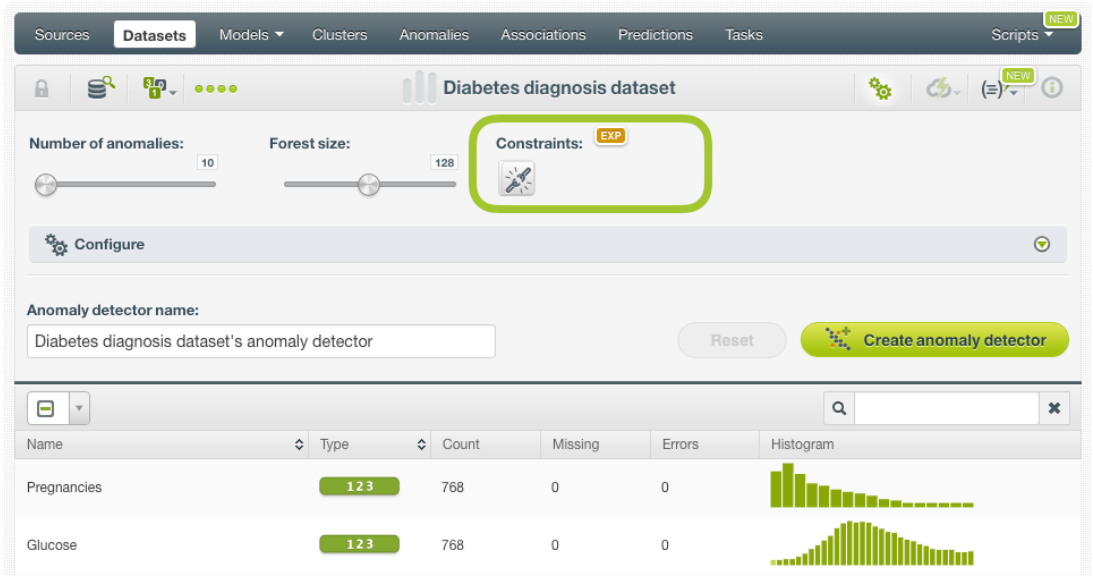

Figure 4.4: Anomalies constraints

4.4 ID Fields

This option allows you to include fields in your anomaly view to see their values in the DATA INSPECTOR (see Chapter 5) but they will not be used to compute the anomaly score. Select fields one by one by typing the first characters of the field name in the selector shown in Figure 4.5. A list of fields found in the dataset matching those characters will appear so you can select one of them. Text and items fields are always treated as ID fields since anomalies cannot include them to compute the anomaly score (Section 2.2).

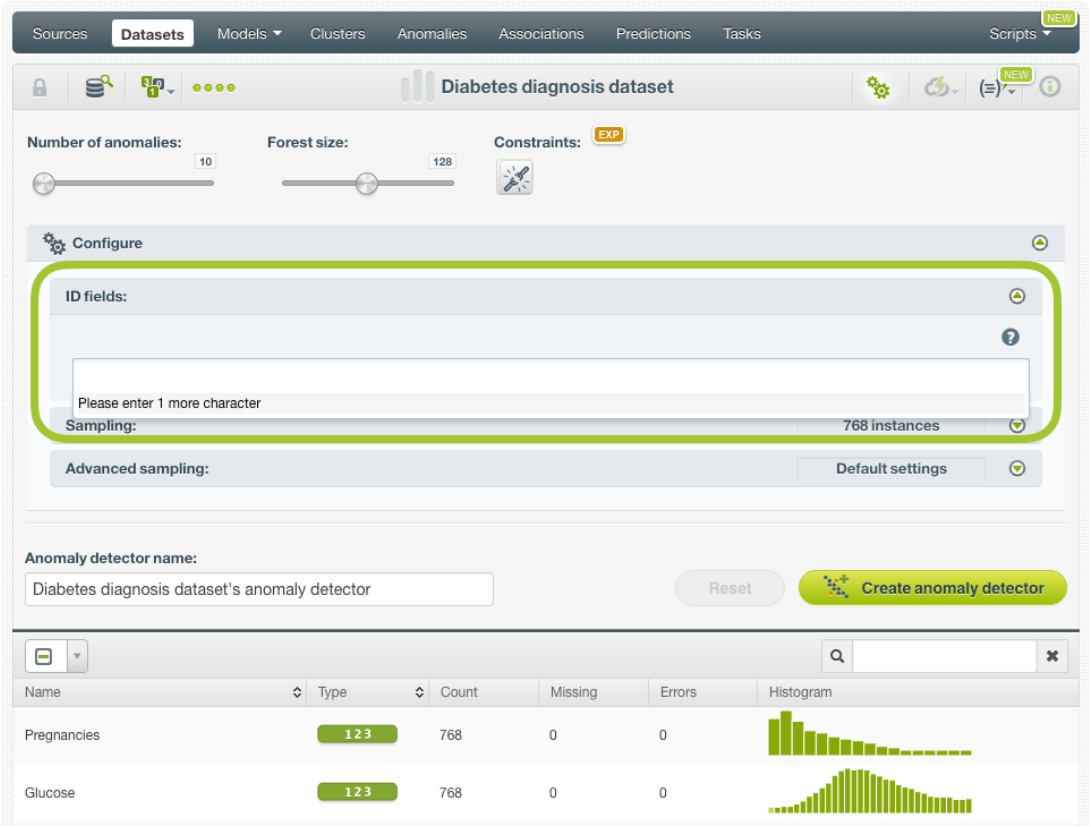

Figure 4.5: Anomalies ID fields

Note: fields marked as non-preferred in your dataset are not eligible as ID fields. To include a non-preferred field as an ID field, first set that field as preferred. (See the corresponding section Updating fields of the Datasets document[1].)

4.5 Sampling Options

Sometimes you do not need all the data contained in your testing dataset to generate your anomalies. If you have a very large dataset, sampling may be a good way of getting faster results. (See Figure 4.6.) You can configure the sampling options explained in the following sections.

4.5.1 Rate

The rate is the proportion of instances to include in your sample. Set any value between 0% and 100%. It defaults to 100%.

4.5.2 Range

Specifies a subset of instances from which to sample, e.g., choose from instance 1 until 200. The **Rate** you set will be computed over the **Range** configured.

4.5.3 Sampling

By default, BigML selects your instances for the sample by using a random number generator, which means two samples from the same dataset will likely be different even when using the same rates and row ranges. If you choose deterministic sampling, the random-number generator will always use the

[1] https://static.bigml.com/pdf/BigML_Sources_and_Datasets.pdf

same seed, thus producing repeatable results. This lets you work with identical samples from the same dataset.

4.5.4 Replacement

Sampling with replacement allows a single instance to be selected multiple times. Sampling without replacement ensures that each instance cannot be selected more than once. By default BigML generates samples without replacement.

4.5.5 Out of Bag

This argument will create a sample containing only out-of-bag instances for the currently defined rate. If an instance is not selected as part of a sample, it is considered out of bag. Thus, the final total percentage of instances for your sample will be 100% minus the rate configured for your sample (when replacement is false). This can be useful for splitting a dataset into training and testing subsets. It is only electable when a sample rate is less than 100%.

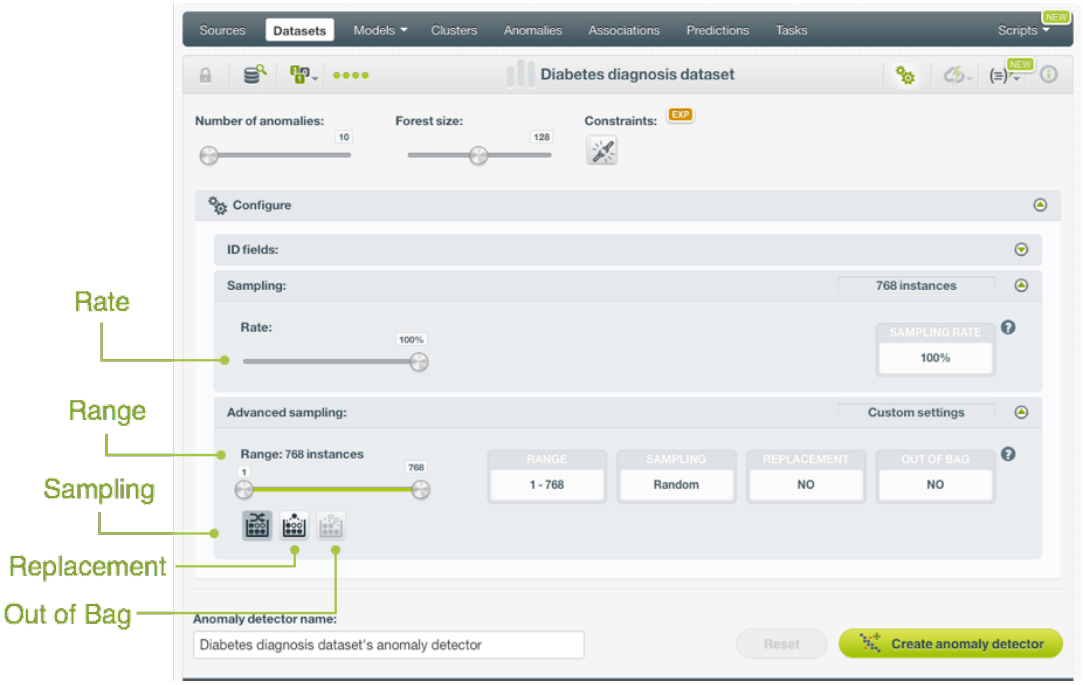

Figure 4.6: Sampling options for anomalies

4.6 Creating Anomolies with Configured Options

After finishing the configuration of your options, you can change the default anomoly name in the editable text box. Then you can click on the Create anomoly detector button to create the new anomoly, or reset the configuration by clicking on the Reset button.

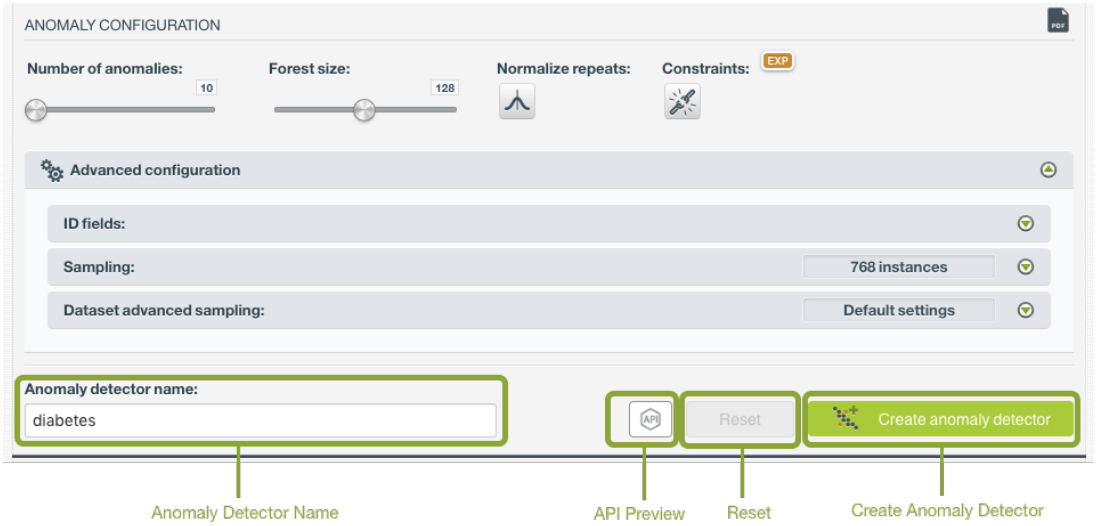

Figure 4.7: Create anomoly after configuration

4.7 API Request Preview

The API Request Preview button is in the middle on the bottom of the configuration panel, next to the Reset button (See (Figure 4.7)). This is to show how to create the anomoly programmatically: the endpoint of the REST API call and the JSON that specifies the arguments configured in the panel. Please see (Figure 4.8) below:

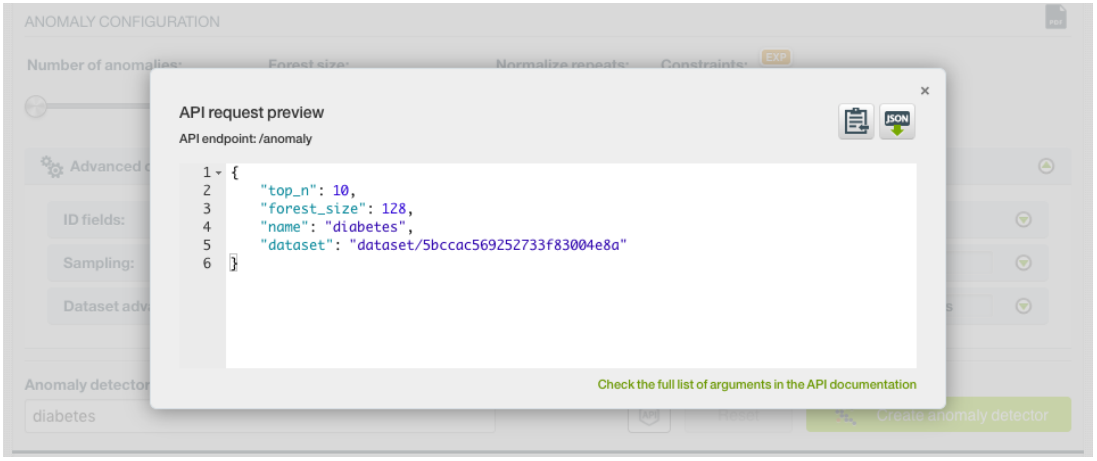

Figure 4.8: Anomoly API request preview

There are options on the upper right to either export the JSON or copy it to clipboard. On the bottom there is a link to the API documentation for anomolies, in case you need to check any of the possible values or want to extend your knowledge in the use of the API to automate your workflows.

Please note: when a default value for an argument is used in the chosen configuration, the argument won't appear in the generated JSON. Because during API calls, default values are used when arguments are missing, there is no need to send them in the creation request.

CHAPTER 5

Visualizing Anomalies

BigML anomaly view is composed of two main blocks: TOP ANOMALIES (on the left) and DATA INSPECTOR (on the right). (See Figure 5.1.)

Figure 5.1: Anomaly view

- TOP ANOMALIES is a list containing the top anomalous instances found in the dataset ranked by their anomaly scores. The list contains the top 10 anomalous instances by default, unless you configured the number of anomalies in advanced (see Section 4.1). For each anomalous instance you get the following information:

  - **Anomaly score**: it is always a number between 0% and 100%. Higher values indicate more anomalous instances. Usually a **score of 60%** or higher is a solid basis for a given instance to be considered anomalous. Learn more about anomaly score in Section 2.1.

    Note: the 60% threshold is no longer valid if the parameter Constraints is enabled since scores tend to be inflated. (See Section 4.3.)

  - **Field importances**: you can see a histogram indicating the contribution of the input fields to the anomaly score. Each field importance can range from 0% to 100%. Learn more about field importances in Section 2.1.

- When you mouse over an instance from the TOP ANOMALIES list, you can see the values per field in the DATA INSPECTOR on the right. The fields in the DATA INSPECTOR are ordered by importance, so fields with contributing more to the anomaly score for that instance will appear at the top. At the end of the list, you will find the fields selected as **ID fields** and the text and items fields, which are not used to compute the anomaly score (see Section 4.4). Apart from the instance values for each field, you can also see the field histogram and statistics. (You can find an explanation of fields statistics in the section Understanding Datasets of the Datasets document[1].)

Figure 5.2: Anomaly visualization

By clicking on the icon in the top left of the DATA INSPECTOR, you can also see and copy the instance values in CSV and JSON format. (See Figure 5.4.)

[1] https://static.bigml.com/pdf/BigML_Sources_and_Datasets.pdf

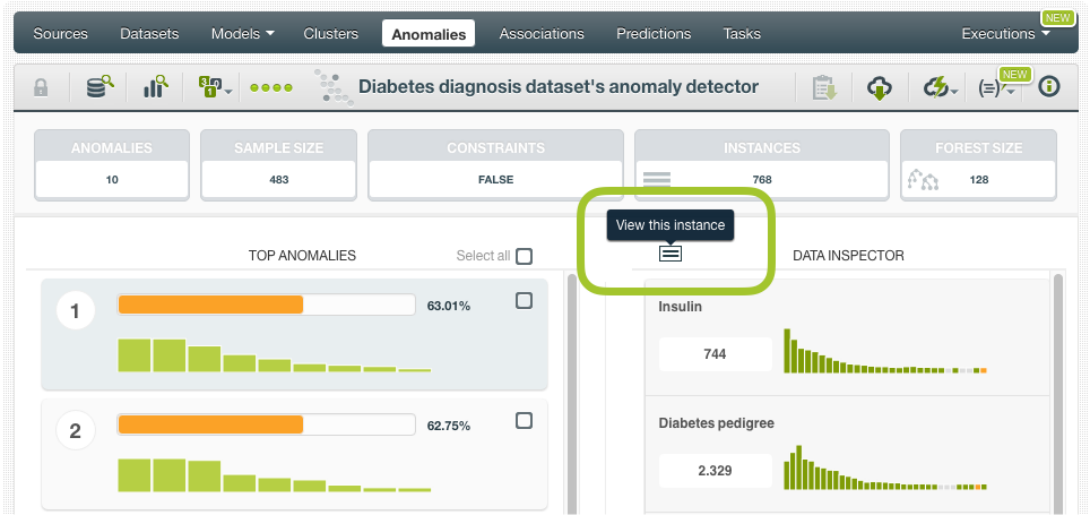

Figure 5.3: Click to see anomalous instances values

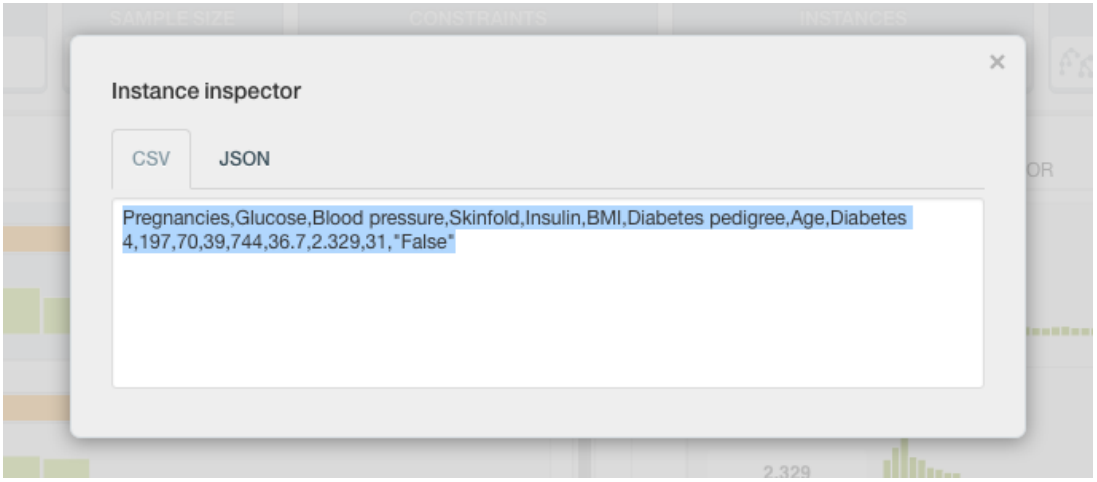

Figure 5.4: Export anomalous instances values

5.1 Anomaly Visualization with Images

When anomalies have images, their visualization is the same as what is described earlier in this chapter. Additionally, there are image previews in the DATA INSPECTOR.

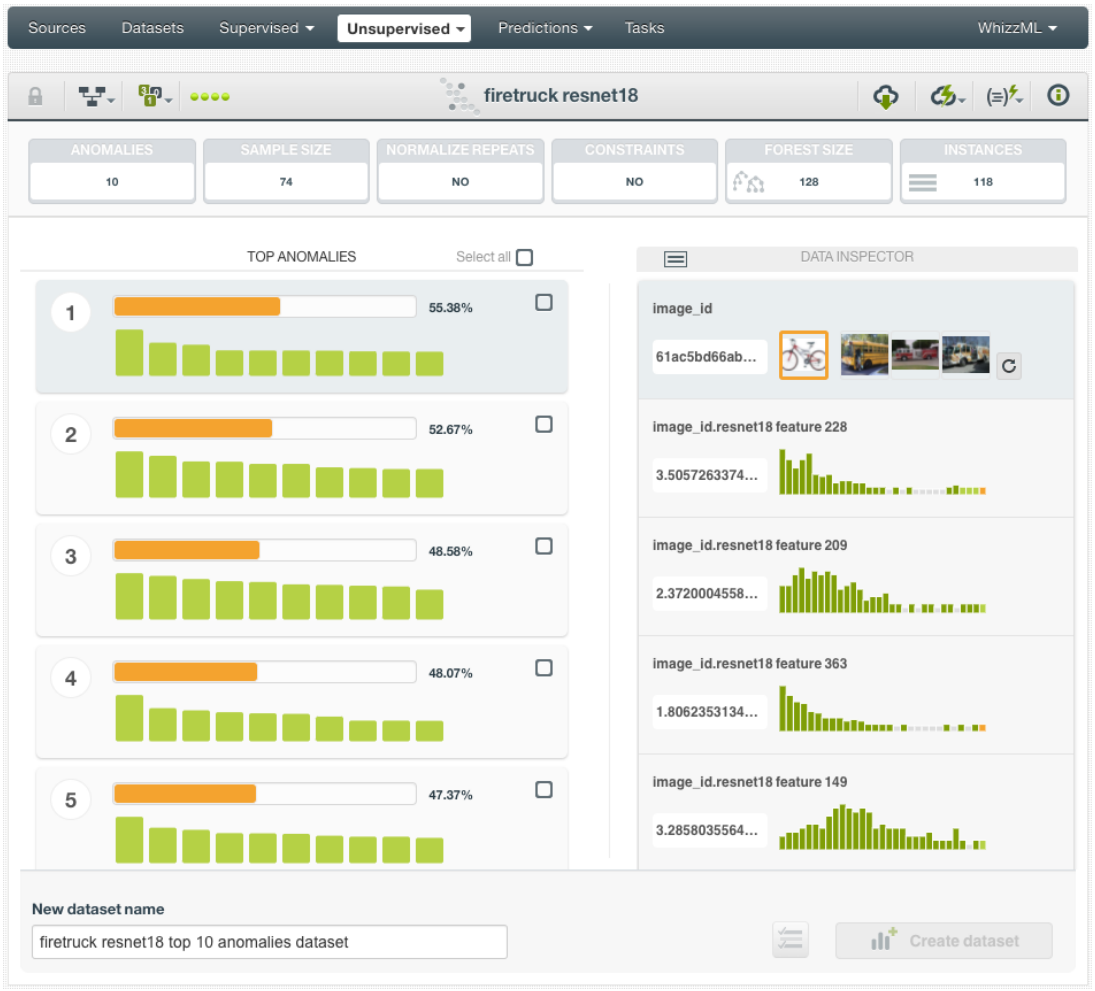

Figure 5.5: Cluster view with images

As shown in Figure 5.5, the DATA INSPECTOR not only includes field values and field histograms, it also presents a list of thumbnail images as the value of the image field, *image_id*. The thumbnail images serve as previews of the images and can be changed by using the reloading icon next to them. Clicking on a thumbnail will bring up the close-up view of the image.

5.2 Create a Dataset From Anomalies

In BigML, you can easily remove the anomalous instances and create a new clean dataset; or you can create a new dataset just including the anomalous instances to analyze them further. The following sections explain both options.

5.2.1 Remove Anomalous Instances

Create a new dataset **removing** the anomalous instances from the original dataset used to create the anomalies. The new dataset will contain all the input fields used to compute the anomaly score and the ID fields. Read more about ID fields in Section 4.4.

First, select the top anomalies you want to remove. Then click the icon next to the green button to remove the selected instances and finally click Create dataset. (See Figure 5.6.)

Figure 5.6: Create a dataset removing anomalies

5.2.2 Include Only Anomalous Instances

Create a new dataset **including** just the selected instances of your top anomalies. The new dataset will contain all the input **fields** used to compute the anomaly score and the ID fields. Read more about ID fields in Section 4.4.

First, select the top anomalies you want to include. Then ensure the icon next to the green button to remove anomalies is not clicked as shown in Figure 5.7. Finally, click `Create dataset` button.

Figure 5.7: Create a dataset including only anomalies

CHAPTER 6

Anomaly Predictions: Anomaly Scores

6.1 Introduction

Besides finding out the anomalous instances in a dataset, you can also use your anomaly detector to score new data that the model has not yet seen. Predictions for anomalies are referred to as **anomaly scores** in BigML, since they aim to quantify the level of anomalousness for new data instances. Anomaly scoring is possible either for **single instances**, i.e., one by one, or for **multiple instances** simultaneously, i.e., in batch. Each score comes with a field importance measure to indicate the relative contribution of each field in the anomaly score.

The predictions tab in the main menu of the BigML Dashboard is where all of your saved predictions are listed (Figure 6.1). In the scores list view, you can see the icon for the **Anomaly Detector** used for each score, the **Name** of the score, the **Anomaly Score**, and the **Age** (time since the score was created). You can also search your scores by name clicking in the search menu option on the top right menu.

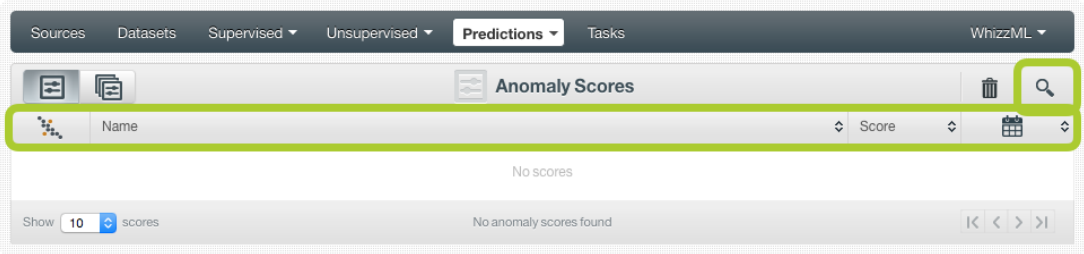

Figure 6.1: Predictions list view

By default, when you first create an account at BigML, or every time that you start a new project, your list view for predictions will be empty. (See Figure 6.2.)

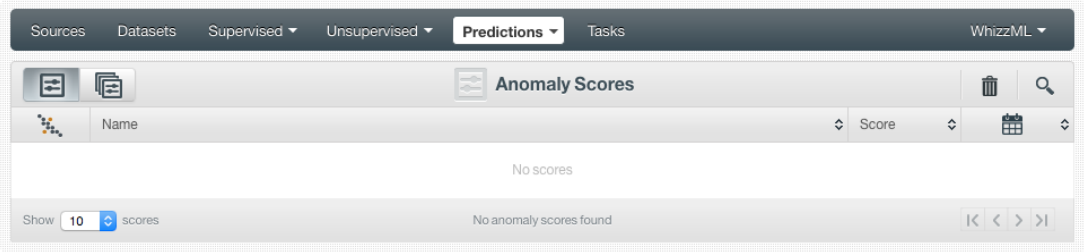

Figure 6.2: Empty Dashboard scores view

Anomaly scores are saved under the ANOMALY DETECTION otpion in the menu (see Figure 6.3.)

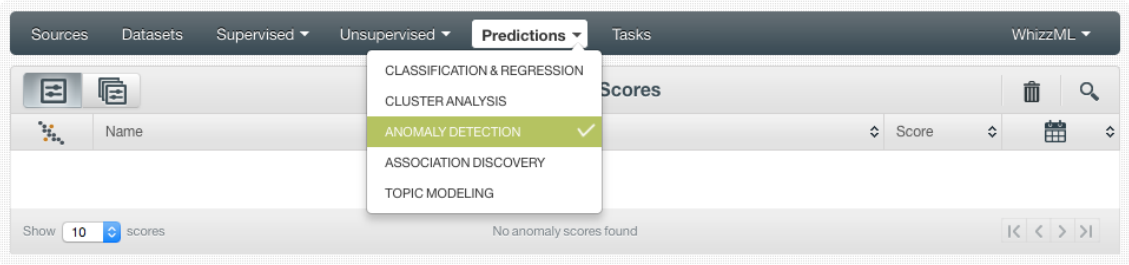

Figure 6.3: Menu options of the scores list view

From this view, you can select to view the list of your **single anomaly scores** or your **batch anomaly scores** by clicking on the corresponding icons (see Figure 6.4 and Figure 6.5.)

Figure 6.4: Single scores icon

Figure 6.5: Batch anomaly scores icon

6.2 Creating Anomaly Scores

BigML provides two different ways to predict scores for new instances using your anomaly detector:

- ANOMALY SCORE: to score single instances.
- BATCH ANOMALY SCORE: to score multiple instances simultaneously.

6.2.1 Anomaly Score

To score new single instances BigML provides a form containing the fields used by the anomaly detector so you can easily configure the input data and get an immediate response.

Follow these steps to create your anomaly score:

1. Click the ANOMALY SCORE option in the **1-click action menu**. (See Figure 6.6.)

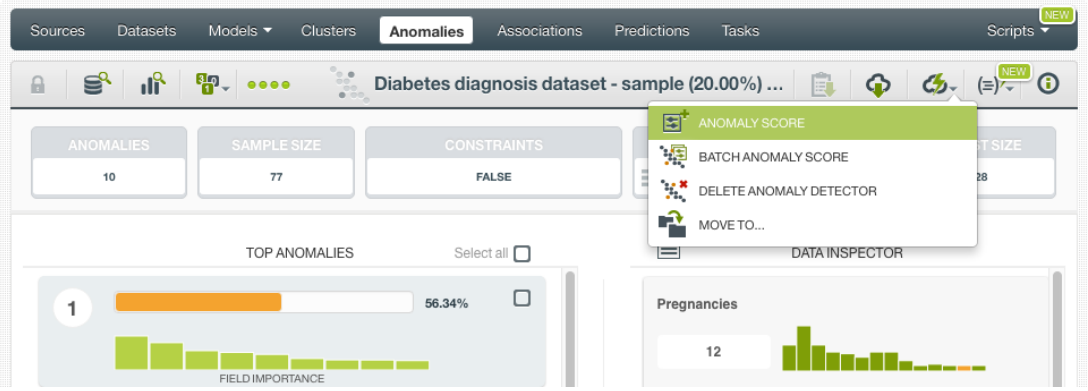

Figure 6.6: Predict option from anomaly 1-click menu

Alternatively, click ANOMALY SCORE in the **pop up menu** from the anomaly list view as shown in Figure 6.7.

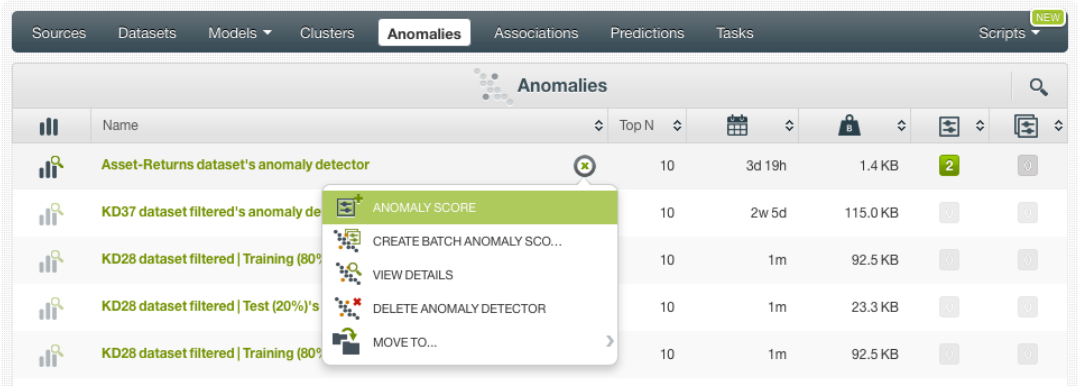

Figure 6.7: Predict option from anomaly pop up menu

2. You will be redirected to the **prediction form**, where you will find all the input fields used by the anomaly to compute the anomaly score. (See Figure 6.8.)
3. **Select** the input fields and **set their values**. Depending on the field type you will need to input the values in different ways. (See Figure 6.8.)
   - Numeric fields: move the slider or input a specific value in the box.
   - Categorical fields: select one class from the selector.

   Note: text and items fields are not supported to create anomalies (see Section 2.2).
4. Click Score to get the **anomaly score** on top of the form. (See Figure 6.8.)
5. The score is saved automatically so you can find it afterwards in the prediction list view. (See Figure 6.1.)

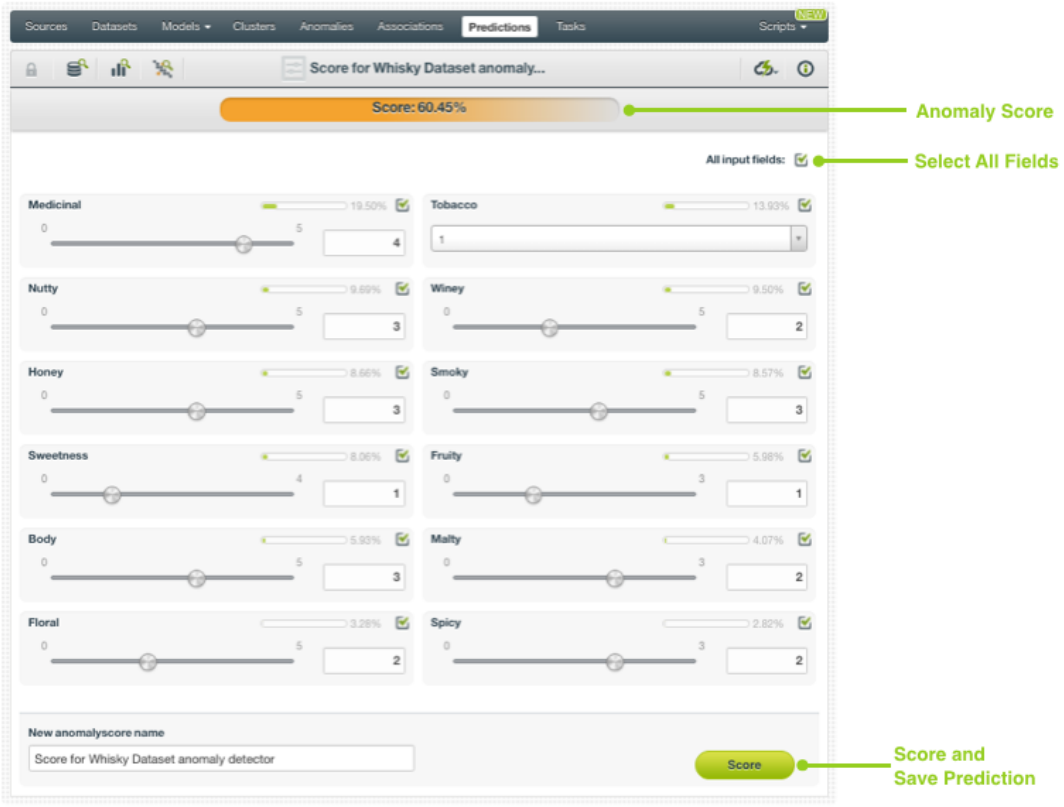

Figure 6.8: Single anomaly score prediction

Note: single anomaly scores are only available for anomalies with less than 100 fields from the BigML Dashboard. If you want to perform single anomaly scores for anomalies with higher number of fields you can use the BigML API[1].

6.2.1.1 Anomaly Score with Images

BigML anomalies can be trained from images using extracted image features (Section 2.4). Because image features are automatically generated numeric fields, creating anomaly scores with images is the same as creating other anomaly socres. The only thing different is input fields of images.

Note: When the input fields contain images, in order to create the anomaly score, BigML will extract image features automatically to match what were used in the dataset to train the anomaly.

[1] https://bigml.com/api/anomalyscores

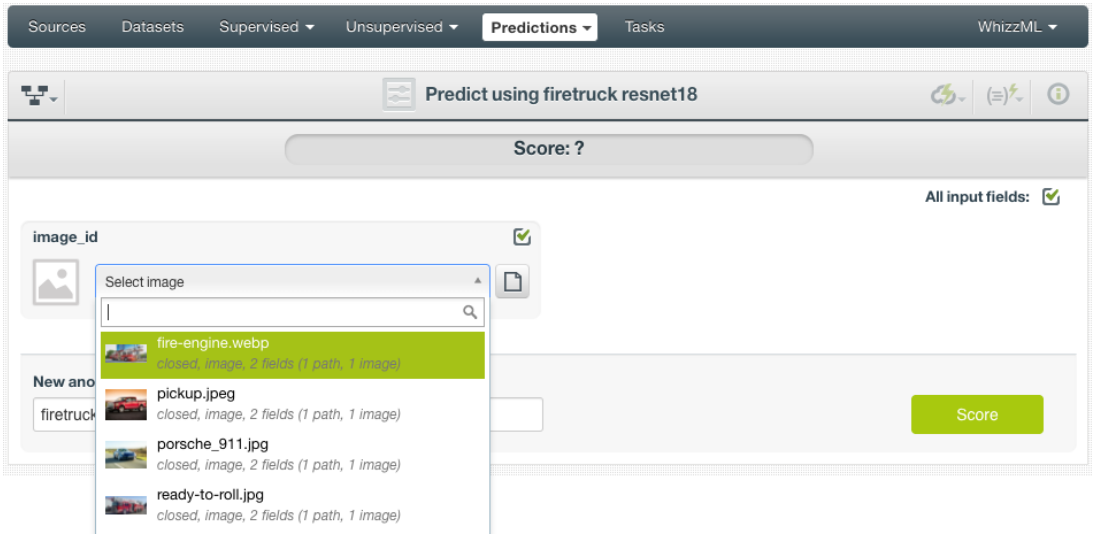

Figure 6.9: Select a single image source in the image input field

The anomaly in Figure 6.9, “firetruck resnet18”, was created from a dataset containing image features extracted from a pre-trained CNN, *ResNet-18*. Creating an anomaly score using the anomaly will be directed to the **prediction form** which presents all input fields used by the cluster. One of them is the image field. Because this is a single anomaly score, which is a single prediction, an image is input by using a single image source. Clicking on the input field box, single image sources available will be in the dropdown list. There is also a search box which can be used to locate specific ones.

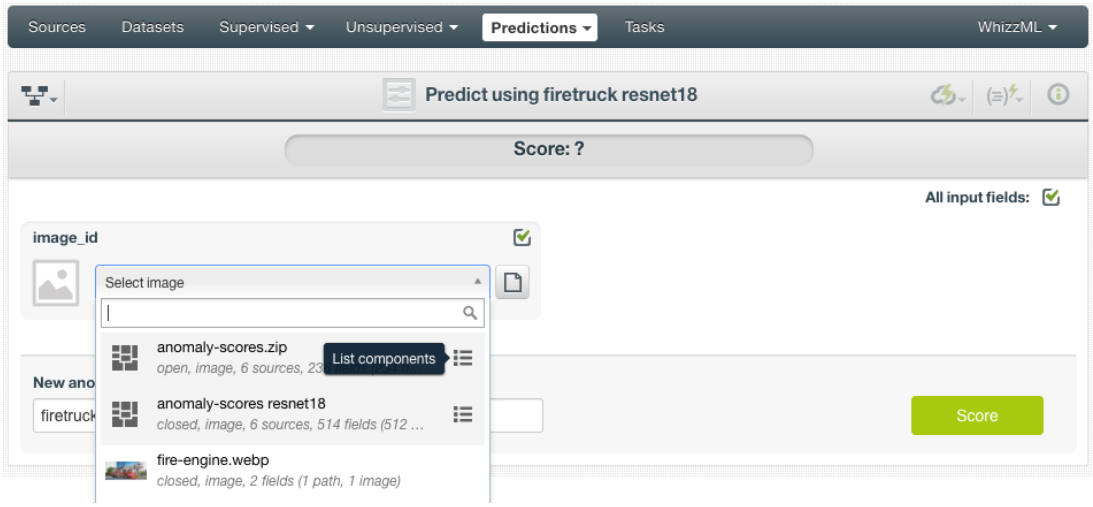

Figure 6.10: List the components of a composite source

Oftentimes single image sources were used for creating a composite source, they become component sources of the composite source. Or an image was uploaded as a part of an archive file (zip/tar) which created a composite source. In those cases, the composite source will be shown in the dropdown list, along with an icon “List components”. In the example in Figure 6.10, `anomaly-scores.zip` is a composite source, click on the icon to show its component sources.

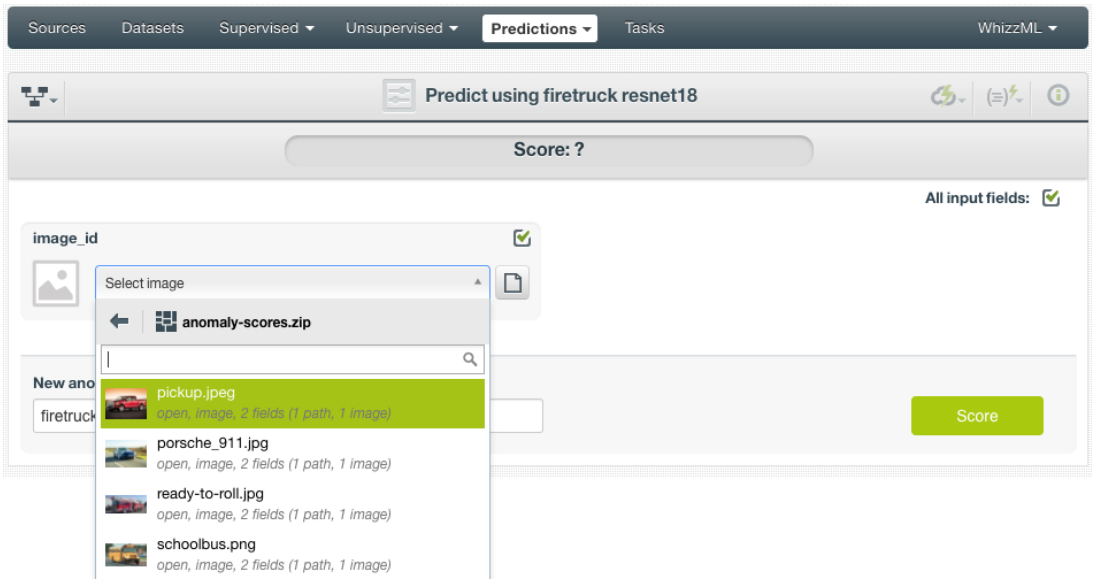

Figure 6.11: Select a component of a composite source

After the component sources of the composite are listed, scroll the dropdown list to find the desired one, then click to select it, as shown in Figure 6.11. There is also a search box to locate specific component sources.

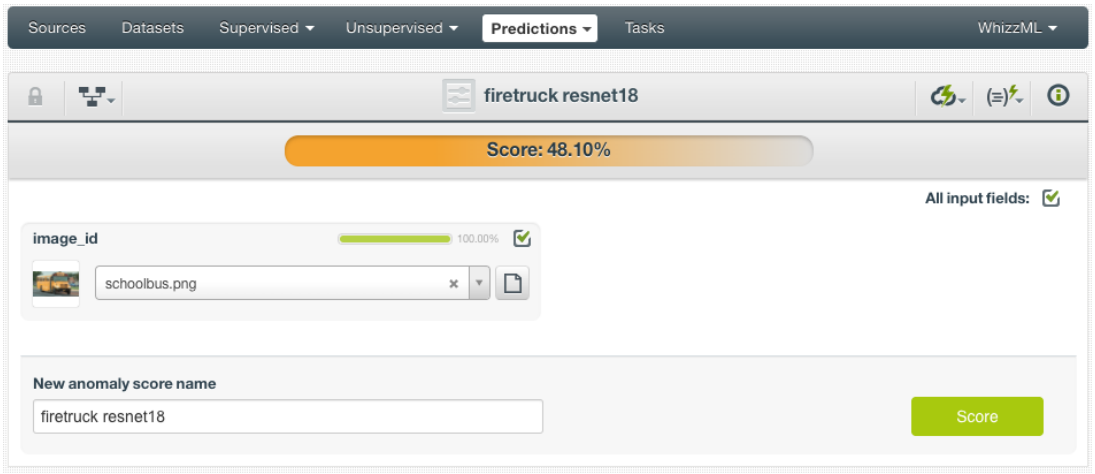

Figure 6.12: An anomaly score with images

After a new anomaly score is created, as shown in Figure 6.12, the score is at the top of the form. The anomaly interface is the same as ones created by non-image anomalies. Everything described earlier in this section (Subsection 6.2.1) applies.

6.2.2 Batch Anomaly Scores

BigML batch anomaly scores allow you to make predictions for multiple instances simultaneously. All you need is the anomaly detector you want to use and a dataset containing the instances for which you want to obtain the scores. BigML will create a score for each instance.

Follow these steps to create a batch anomaly score:

1. Click BATCH ANOMALY SCORE option in the anomaly **1-click action menu**. (See Figure 6.13.)

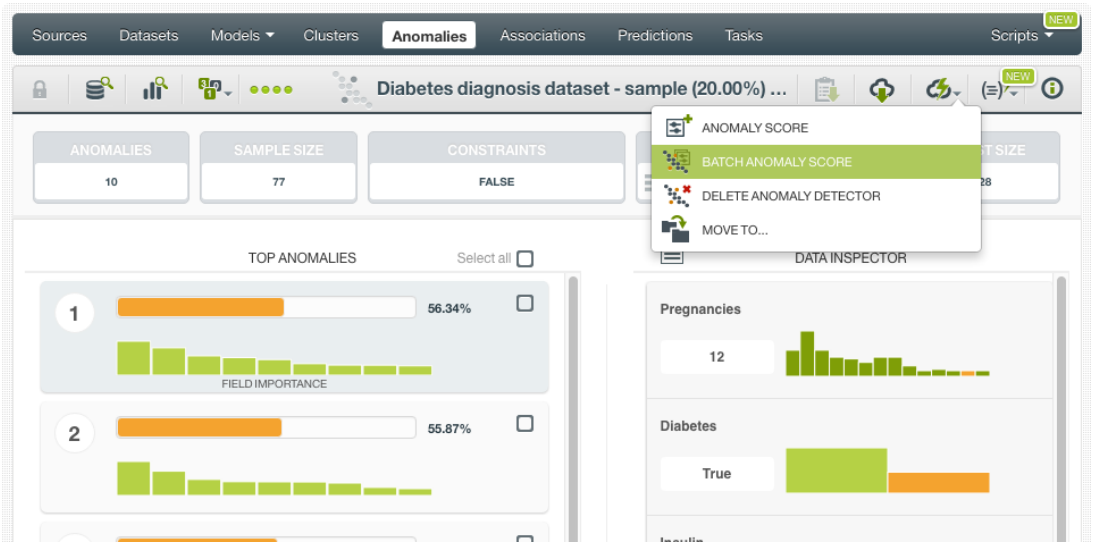

Figure 6.13: Batch anomaly score from 1-click action menu

Alternatively, click CREATE BATCH ANOMALY SCORE in the **pop up menu** from the anomaly list view as shown in Figure 6.14.

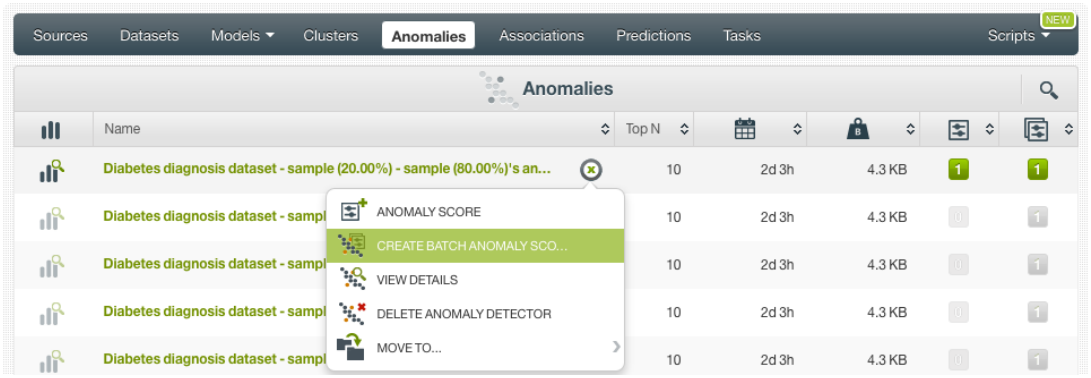

Figure 6.14: Batch anomaly score from pop up menu

2. **Select the dataset** containing all the instances you want to predict. (See Figure 6.15.) The instances should contain the input values for the fields used by the anomaly detector. BigML batch anomaly scores can handle missing data in your dataset as explained in Section 2.2. From this view you can also select another anomaly from the anomaly selector.

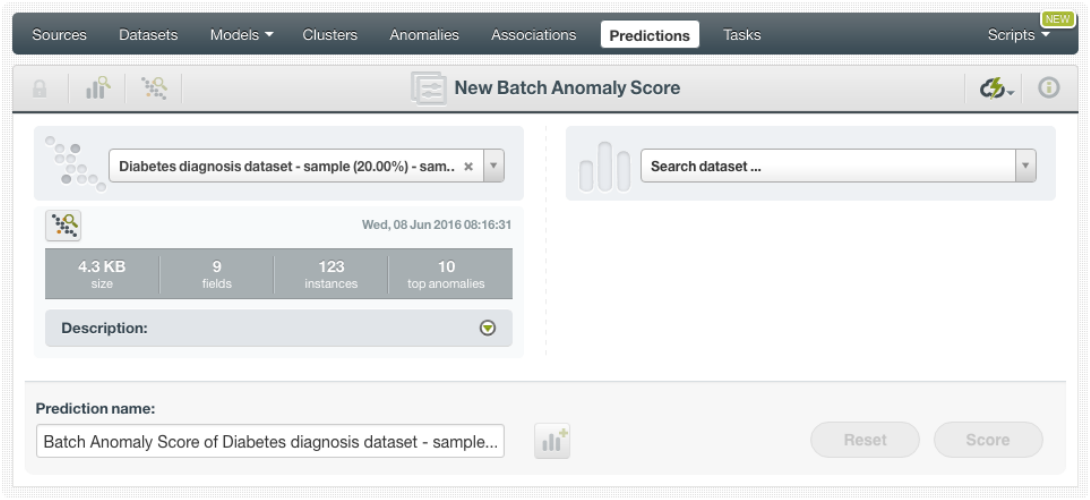

Figure 6.15: Select dataset for batch anomaly scores

3. After you select the anomaly detector and the dataset, the batch anomaly score **configuration options** will appear along with a **preview of the prediction file**. (See Figure 6.16.) The default format is a CSV file including all your dataset fields and adding an extra column for the anomaly scores. You can configure this file using the output settings explained in Section 6.3.

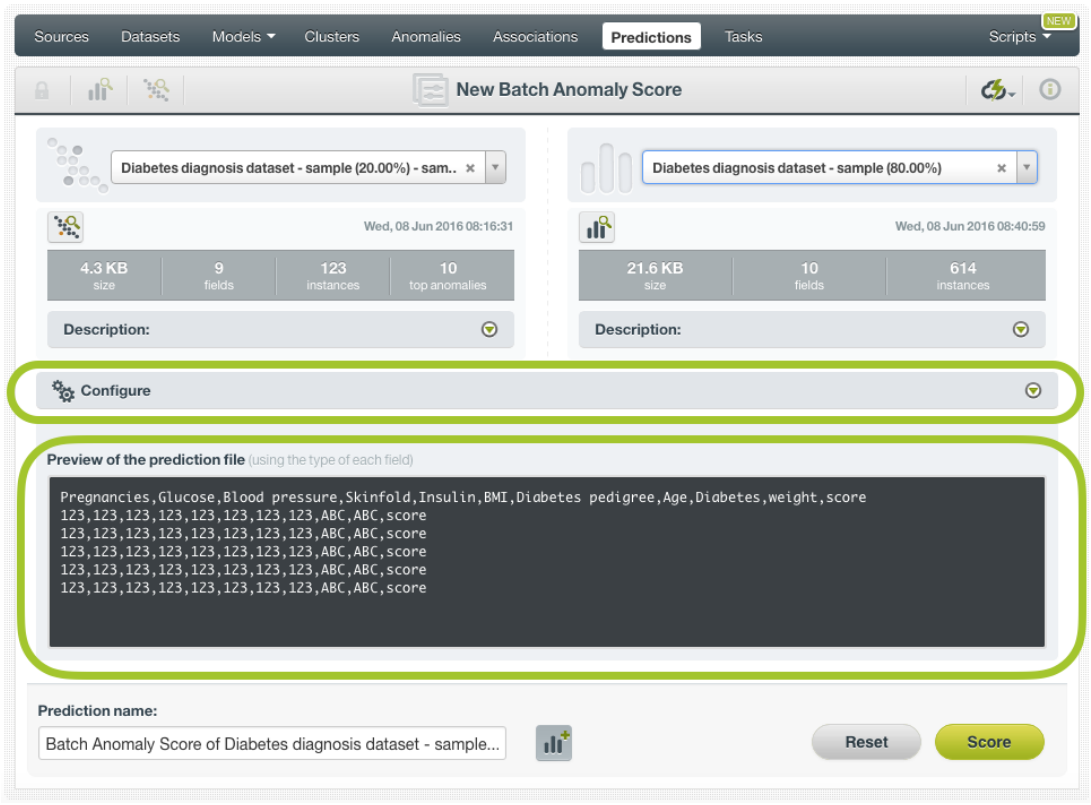

Figure 6.16: Configuration options displayed and output preview

4. By default, BigML generates an output dataset with your batch anomaly scores that you can later find in your datasets section in the BigML Dashboard. This dataset can be helpful to analyze your results afterwards. This option is active by default, but you can deactivate it by clicking in the icon shown in Figure 6.17.

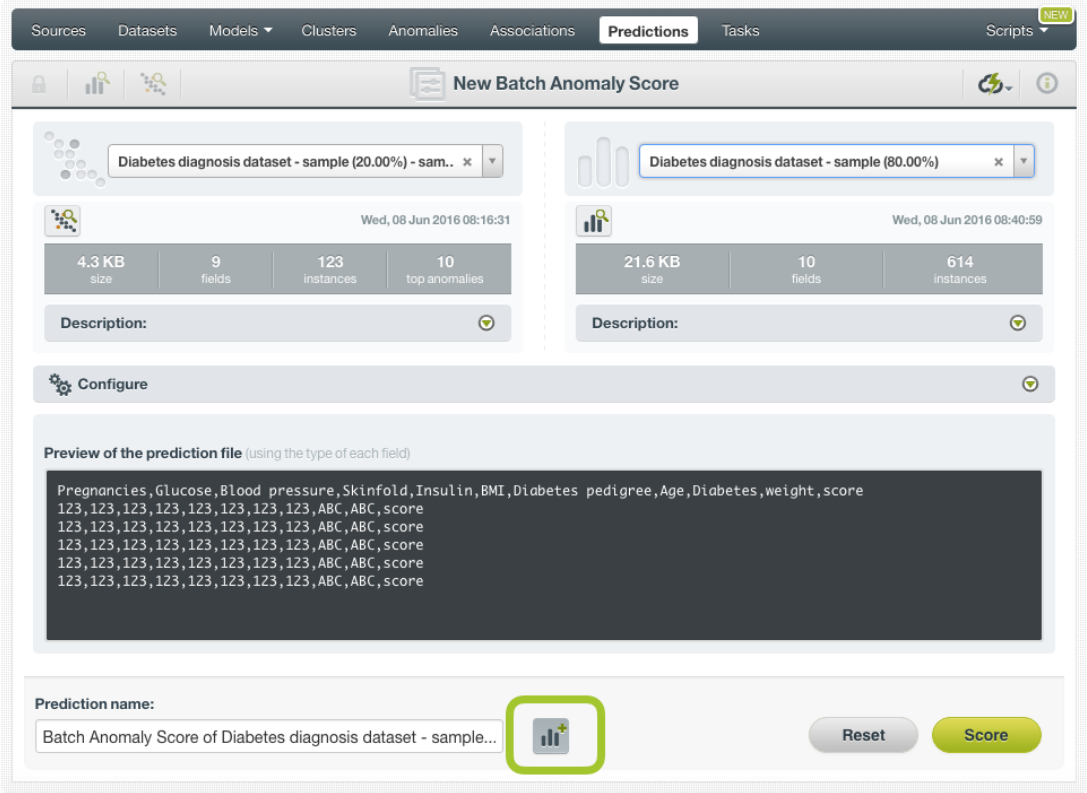

Figure 6.17: Create dataset from batch prediction

5. Finally click on the Score button to generate your batch anomaly score.

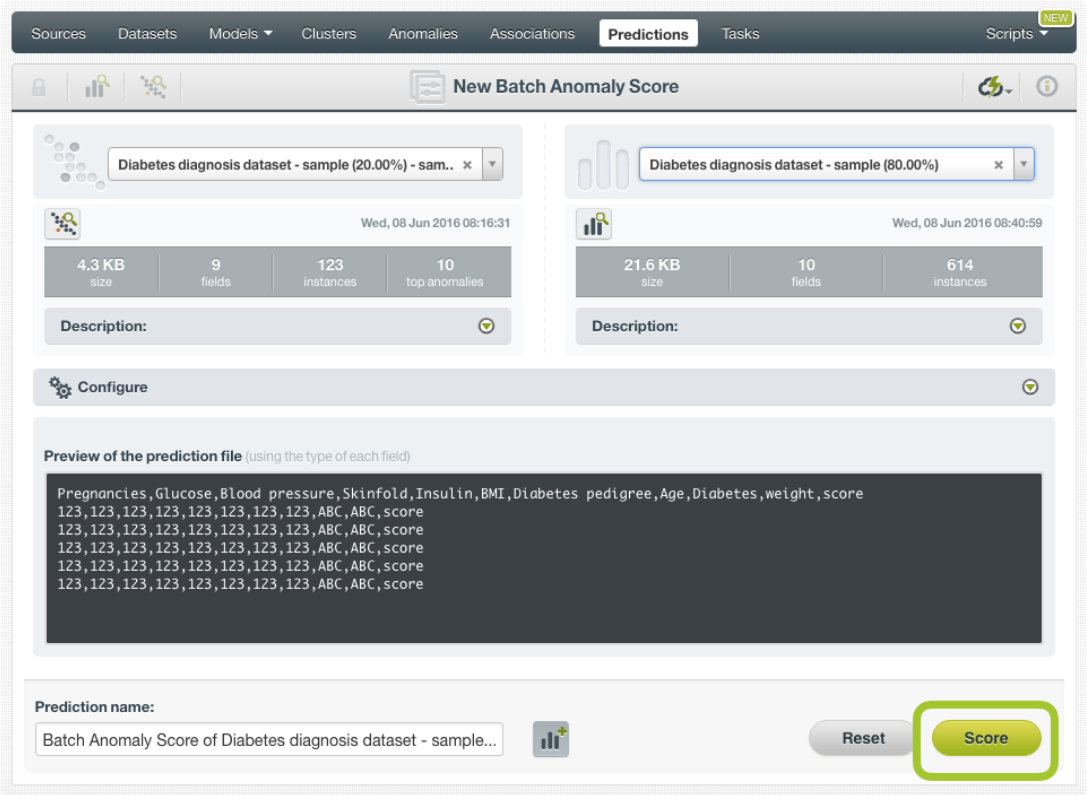

Figure 6.18: Click score

6. When the batch anomaly score is created, you will be able to **download the batch score** containing all your dataset instances along with a score for each one of them. If you did not disable the option to create a new dataset, you will also be able to access the **output dataset** from the batch anomaly score view. (See Figure 6.19.)

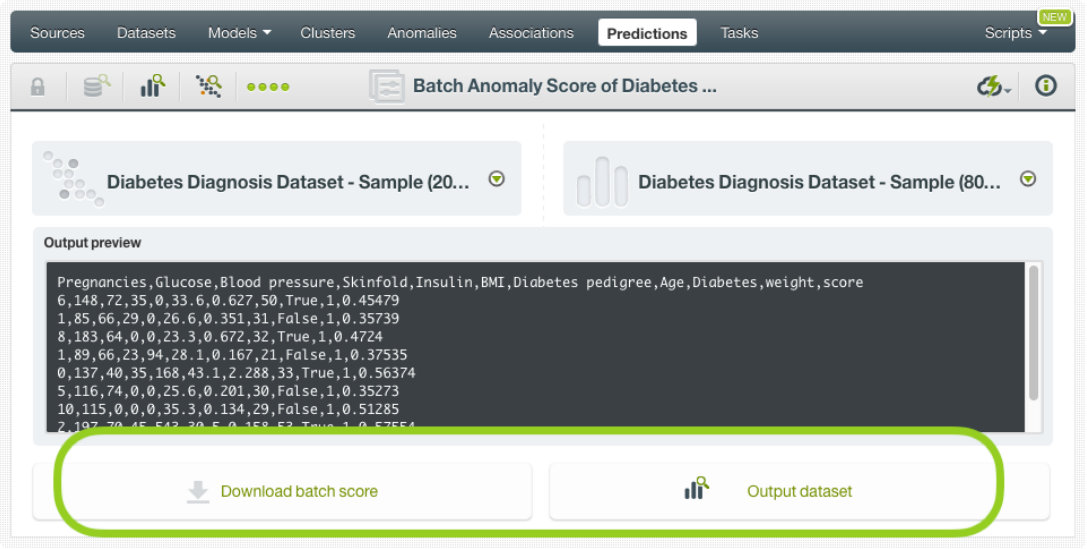

Figure 6.19: Download batch score and access output dataset

6.2.2.1 Batch Anomaly Scores with Images

BigML anomalies can be trained from images using extracted image features (Section 2.4). The input of a batch anomaly score is a dataset. So when creating a batch anomaly score with images, the dataset has to have the same image features used to train the anomaly. The image features are in the dataset used to create the anomaly.

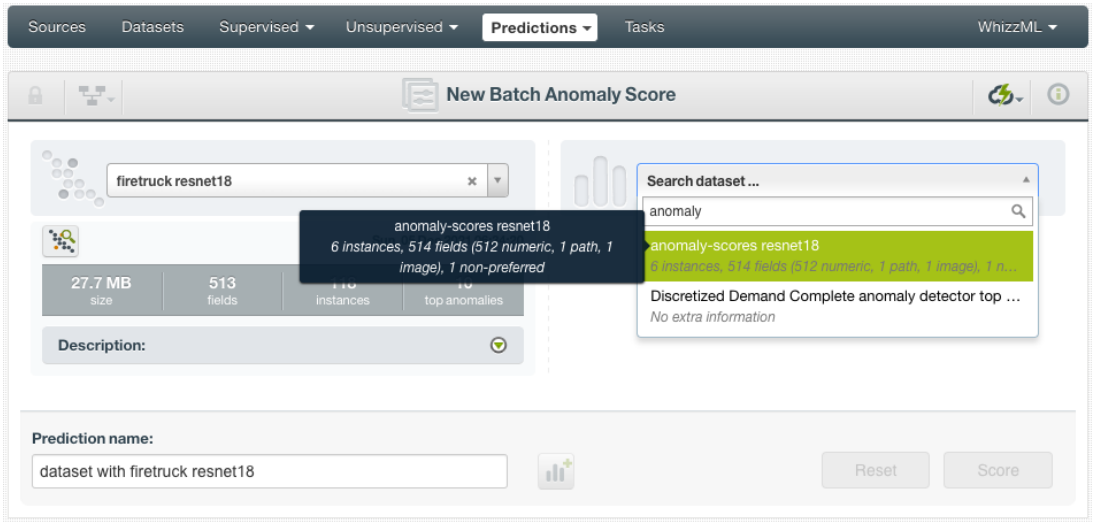

Figure 6.20: Batch anomaly score using an image dataset

As shown in Figure 6.20, the input for the batch anomaly score is selected as `anomaly-scores resnet18`, which is a dataset consisting of six images and contains image features extracted from a pre-trained CNN, *ResNet-18*.

Image features are configured at the source level. For more information about the image features and how to configure them, please refer to section Image Analysis of the Sources with the BigML Dash-

board[2][7].

For the rest of batch anomaly scores with images, including batch anomaly score configuration options and output datasets, everything stated earlier in current section (Subsection 6.2.2) applies.

6.3 Configuring Anomaly Scores

BigML provides several options to configure your anomaly scores, such as defining the automatic **field mapping** performed by BigML and the **output file settings**. See the following for an explanation of both options.

6.3.1 Field Mapping

You can specify which fields in the anomaly match with which fields in the dataset containing the instances you want to score. BigML automatically matches fields by **name**, but you can set an automatic match by **field ID** by clicking in the green switcher shown in Figure 6.21. You can also **manually** search for fields or remove them if you do not want to consider them during the scoring.

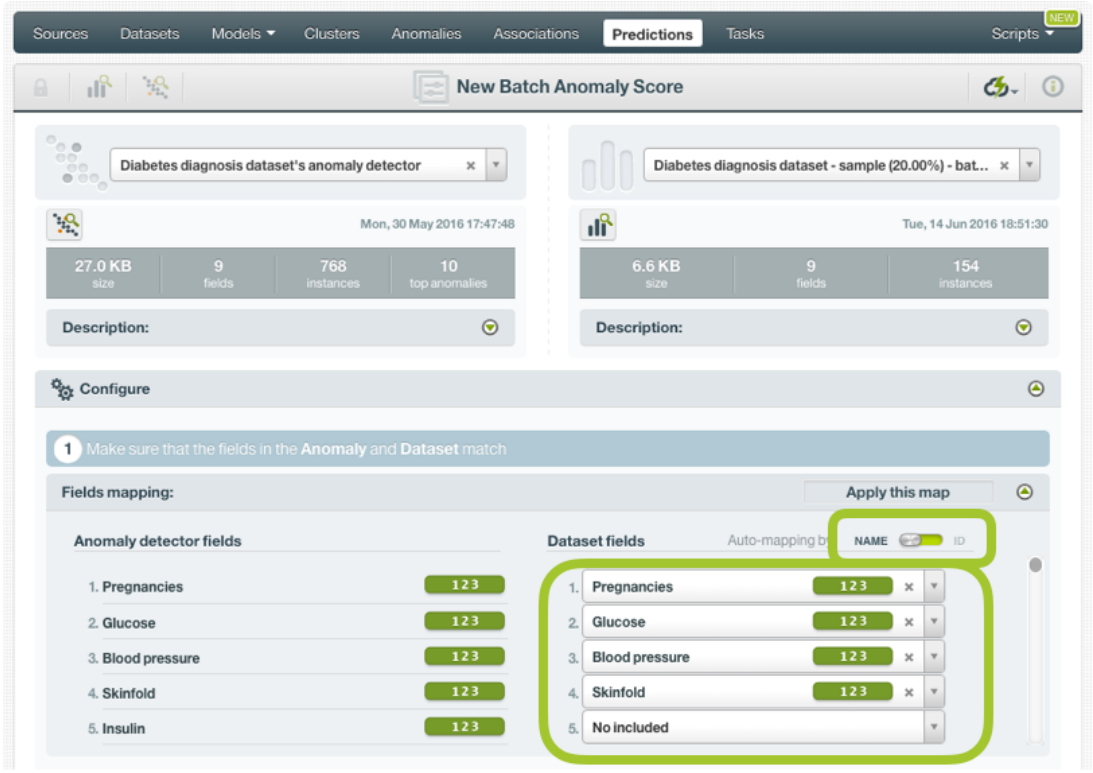

Figure 6.21: Field mapping for batch scores

Note: the field mapping from the BigML Dashboard has a limit of 200 fields, for batch scores with higher number of fields you can use BigML API[3].

6.3.2 Output Settings

Batch anomaly scores return a CSV file containing all your instances and their scores by default. You can tune the following settings to customize your output file:

- **Separator:** this option allows you to choose the best separator for your output file columns. The default separator is comma. You can also select semicolon, tab or space.

[2] https://static.bigml.com/pdf/BigML_Sources.pdf
[3] https://bigml.com/api/batchanomalyscores#ba_batch_anomaly_score_arguments

- **New line**: this option allows you to set the new line character to use as the line break in the generated csv file: “LF”, “CRLF”.
- **Output fields:** you have an option to include or exclude all your dataset fields from your output file. You can also select the fields you want to include or exclude one by one from the preview shown in Figure 6.22.

  Note: a maximum of 100 fields are displayed in the preview, but all your dataset fields are included in the output file by default unless you exclude them.
- **Headers:** this option includes or excludes a first row in the output file (and in the output dataset) with the names of each column. By default, BigML includes the headers.
- **Score column name:** you can customize the name for your scores column.
- **Field importances:** you can include field importances in the output file in addition to the anomaly score. This is an indicator of each field contribution to the anomaly score. (See Section 2.1.)

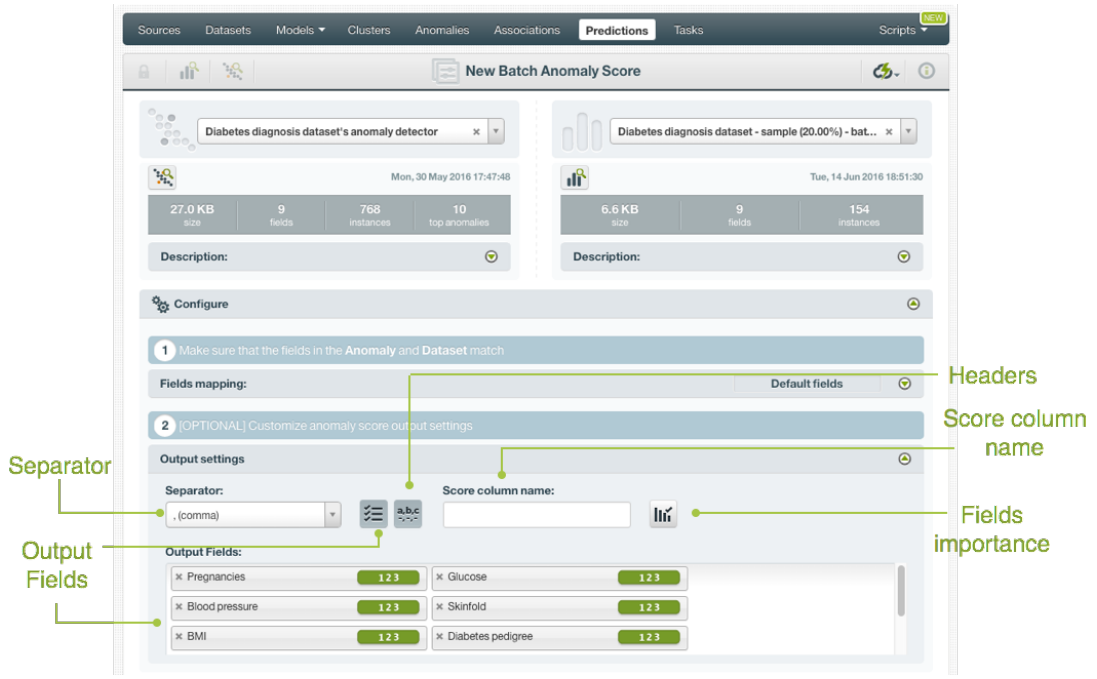

Figure 6.22: Output file settings for batch scores

6.4 Visualizing Anomaly Scores

The visualization of anomaly scores is different for single anomaly scores and batch anomaly scores. The following sections explain both of them.

6.4.1 Single Anomaly Scores

When scoring single instances, you can find the anomaly score at the top of the form. (See Figure 6.23.) Remember that instances scoring higher than 60% may be considered anomalous. You can change the value of the input fields at any time and click Score to have your score recalculated.

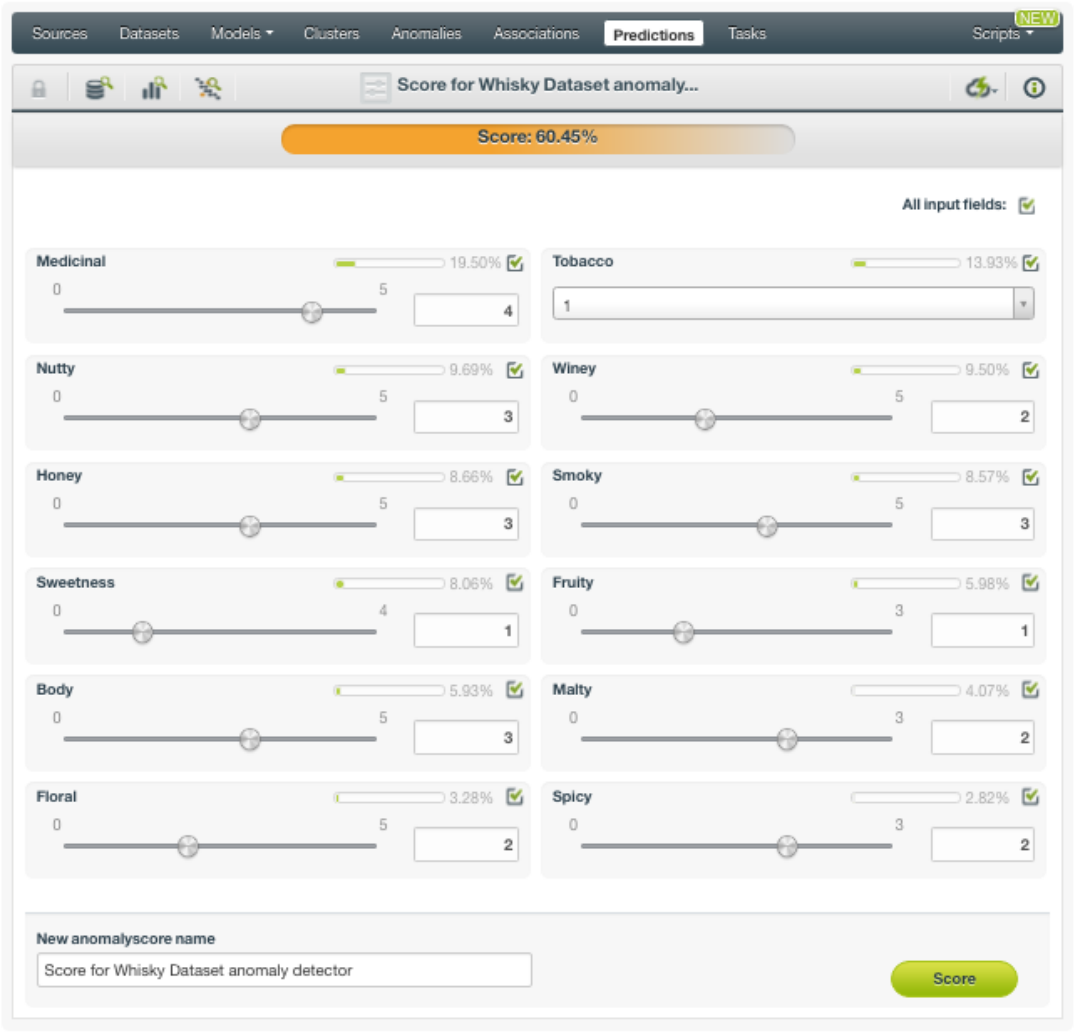

Figure 6.23: Single scores view

All your scores will be saved and listed in the score list view. (See Figure 6.1.)

6.4.2 Batch Anomaly Scores

For batch scores, you always get an **output file** and, optionally, an **output dataset**.

6.4.2.1 Output File

From the batch score view you can access the file view as shown in Figure 6.24.

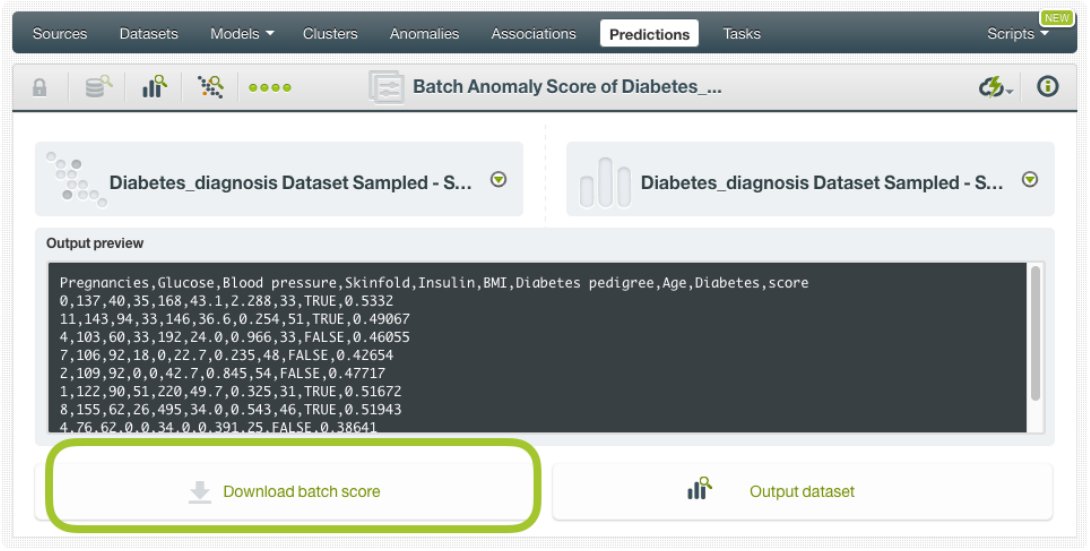

Figure 6.24: Download batch prediction output file

By default, it will be a CSV file containing all the dataset fields. The last column includes the predicted scores for each of the instances. You can customize the output file settings as explained in Subsection 6.3.2.

See an output CSV file example in Figure 6.25 where the last column contains the anomaly score for each instance.

```
Pregnancies,Glucose,Insulin,BMI,Diabetes pedigree,Age,Diabetes,score
0,137,168,43.1,2.288,33,TRUE,0.5332
11,143,146,36.6,0.254,51,TRUE,0.49067
4,103,192,24.0,0.966,33,FALSE,0.46055
7,106,0,22.7,0.235,48,FALSE,0.42654
2,109,0,42.7,0.845,54,FALSE,0.47717
1,122,220,49.7,0.325,31,TRUE,0.51672
8,155,495,34.0,0.543,46,TRUE,0.51943
4,76,0,34.0,0.391,25,FALSE,0.38641
1,118,94,33.3,0.261,23,FALSE,0.36354
```

Figure 6.25: An example of a batch prediction CSV file

6.4.2.2 Output Dataset

By default BigML automatically creates a dataset out of your batch anomaly score. You can disable this option by configuring your batch score as explained in Section 6.3. You can access your output dataset from the batch score view as shown in Figure 6.26.

Figure 6.26: Access batch prediction output dataset

In the output dataset you can find an additional **field** (named by default "score") containing the anomaly score for each one of your instances. (See Figure 6.27.) If you configured your batch score to include the field importance, you will be able to find an additional field for each field in your anomaly detector appended at the end of your output dataset.

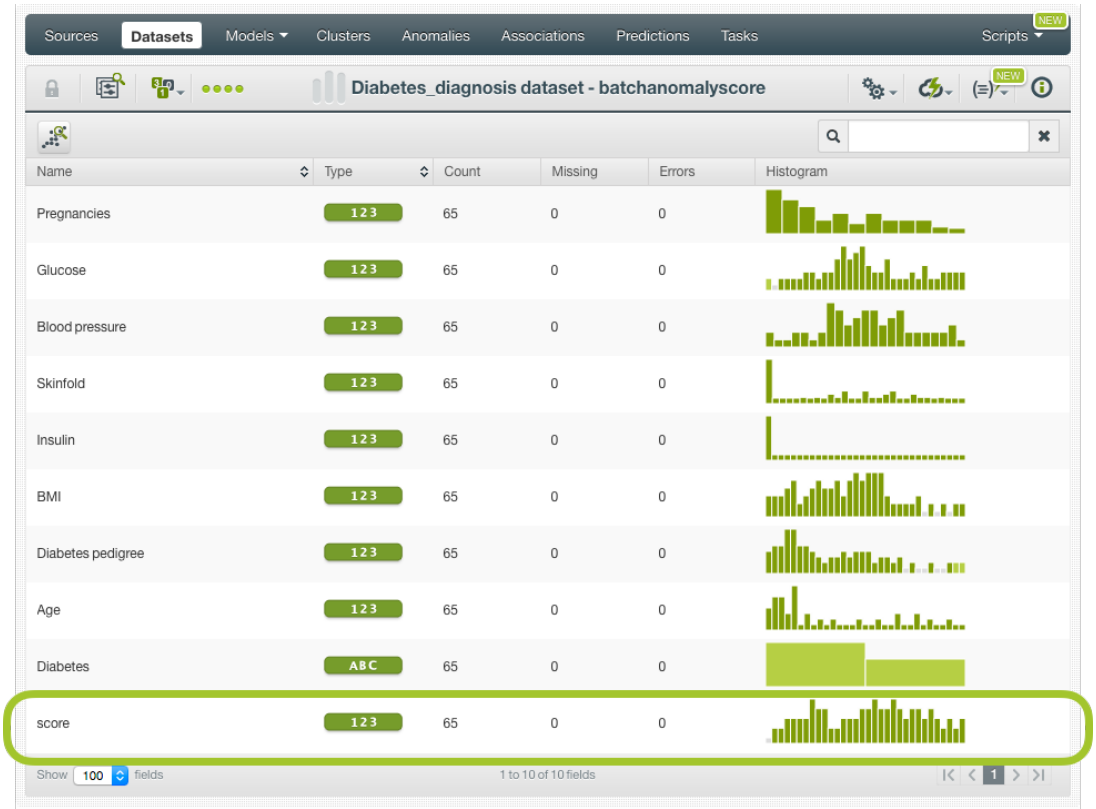

Figure 6.27: Batch scores output dataset

6.4.2.3 Batch Scores 1-Click Action Menu

From the batch score view you can perform the following actions shown in Figure 6.28

- BATCH SCORE AGAIN: this option will redirect you to the batch score creation view where you

will have the same anomaly and score dataset already selected. It is a quick way if you want to create the batch score again using a different configuration.

- BATCH SCORE WITH ANOTHER DATASET: this is an easy way to create a batch score using the same anomaly and a different dataset.
- BATCH SCORE USING ANOTHER ANOMALY: this is an easy way to create a batch score using the same dataset and a different anomaly.
- NEW BATCH SCORE: this will redirect you to the batch score creation view where you will be able to select a score dataset and a anomaly to create your batch score.

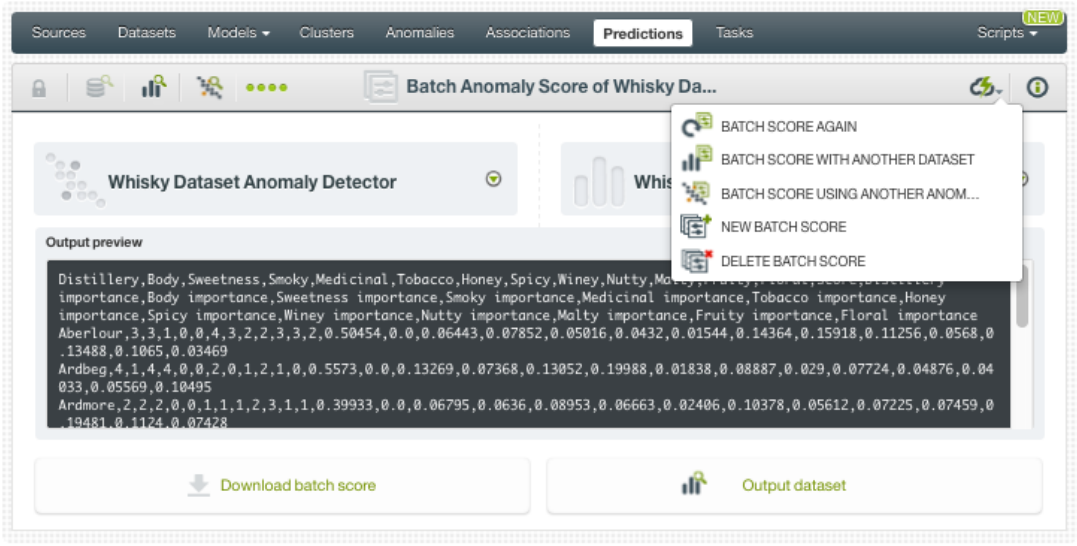

Figure 6.28: Batch score 1-click action menu

6.5 Consuming Anomaly Scores

You can fully used single and batch anomaly scores via the BigML API and bindings. The following subsections explain both tools.

6.5.1 Using Anomaly Scores Via the BigML API

You can perform all the scores actions explained in this document such as creating, configuring, retrieving, listing, updating, and deleting scores via the BigML API.

The example below shows how to create a batch anomaly score with the definition of the input data after the BIGML_AUTH environment variable that contains your authentication credentials is properly set:

```
curl "https://bigml.io/batchanomalyscore?$BIGML_AUTH" \
     -X POST \
     -H 'content-type: application/json' \
     -d '{"anomaly": "anomaly/5423625af0a5ea3eea000028",
          "dataset": "dataset/54222a14f0a5eaaab000000c"}'
```

For more information on using anomaly scores through the BigML API, please refer to anomaly scores REST API documentation[4].

[4] https://bigml.com/api/anomalyscores

6.5.2 Using Anomaly Scores Via the BigML bindings

You can also create, configure, retrieve, list, update, and delete single and batch anomaly scores via **BigML bindings** which are libraries aimed to make it easier to use the BigML API from your language of choice. BigML offers bindings in multiple languages including Python, Node.js Java, Swift and Objective-C. You can find below an example to create an anomaly score with the Python bindings.

```
from bigml.api import BigML
api = BigML()
prediction = api.create_anomaly_score("anomaly/50650bdf3c19201b64000020",
                                      {"salary": 20000, "age": 25})
```

For more information on BigML bindings, please refer to the bindings page[5].

6.6 Descriptive Information

Each anomaly score has an associated **name**, **description**, **category** and **tags**. Those options are editable through the MORE INFO menu on the top right of the anomaly view. (See Figure 6.29.)

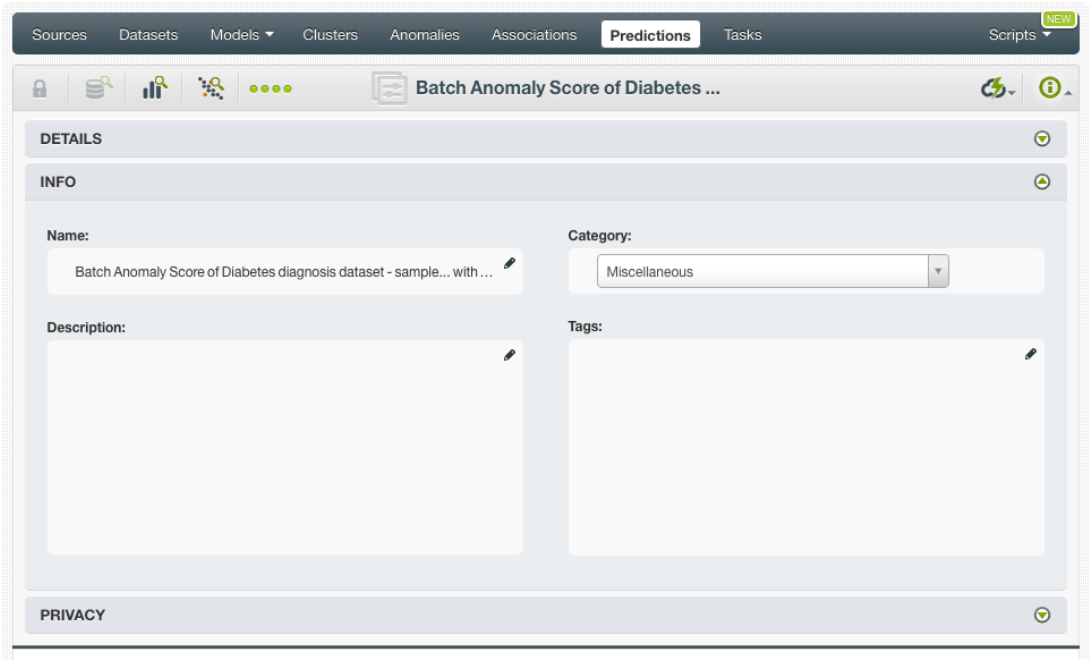

Figure 6.29: Edit scores metadata from More info panel

6.6.1 Scores Name

If you do not specify a **name** for your scores, BigML assigns a default name depending on the type of scores:

- **Single scores:** the name always follow the structure “Score for <objective field name>”.
- **Batch anomaly scores:** BigML combines your score dataset name and the anomaly name: “Batch score of <anomaly name> with <dataset name>”.

Scores names are displayed on the list view and also on the top bar of a score view. Score names are indexed to be used in searches. You can rename your scores at any time from the MORE INFO menu option.

[5] https://bigml.com/tools/bindings

The name of a score cannot be longer than **256** characters. More than one score can have the same name even within the same project, but they will always have different identifiers.

6.6.2 Description

Each anomaly score also has a **description** that it is very useful for documenting your Machine Learning projects. Single and batch scores take the description from the anomaly detector used to create them.

Descriptions can be written using plain text and also markdown[6]. BigML provides a simple markdown editor that accepts a subset of markdown syntax. (See Figure 6.30.)

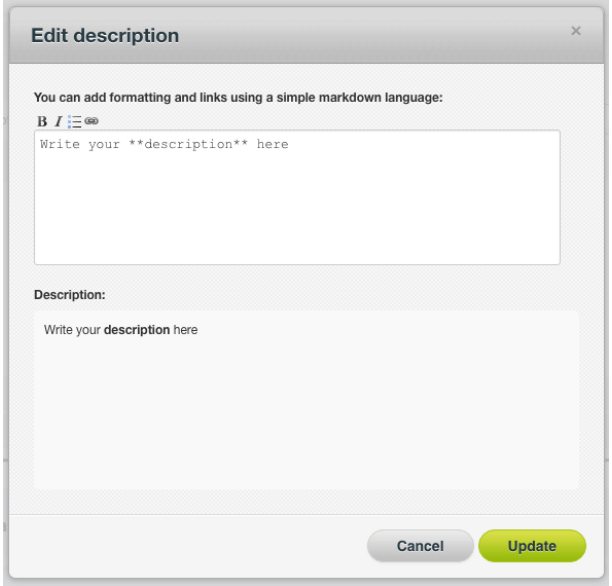

Figure 6.30: Markdown editor for scores descriptions

Descriptions cannot be longer than **8192** characters and can use almost any char.

6.6.3 Category

Each score has associated a **category**. Categories are useful to classify scores according to the domain which your data comes from. This is useful when you use BigML to solve problems across industries or multiple customers. By default, single and batch scores take the category from the anomaly detector used to create them.

A score category must be one of the categories listed on Table 6.1.

[6] https://en.wikipedia.org/wiki/Markdown

Table 6.1: Categories used to classify scores by BigML

Category
Aerospace and Defense
Automotive, Engineering and Manufacturing
Banking and Finance
Chemical and Pharmaceutical
Consumer and Retail
Demographics and Surveys
Energy, Oil and Gas
Fraud and Crime
Healthcare
Higher Education and Scientific Research
Human Resources and Psychology
Insurance
Law and Order
Media, Marketing and Advertising
Miscellaneous
Physical, Earth and Life Sciences
Professional Services
Public Sector and Nonprofit
Sports and Games
Technology and Communications
Transportation and Logistics
Travel and Leisure
Uncategorized
Utilities

6.6.4 Tags

A score can also have a number of **tags** associated with it that can help to retrieve it via the BigML API or to provide scores with some extra information. Scores inherit the tags from the anomaly detector used to create it.

Each tag is limited to a maximum of 128 characters. Each score can have up to 32 different tags.

6.7 Anomaly Scores Privacy

The link displayed in the **privacy panel** is the private URL of your score, so only a user logged into your account is able to see it. Neither single nor batch scores can be shared from the BigML Dashboard by sharing a link as you can with other resources.

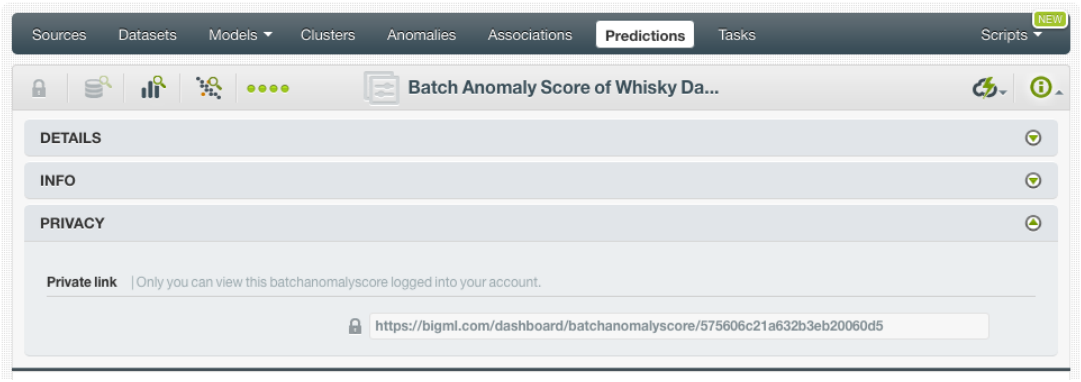

Figure 6.31: Private link of an anomaly score

6.8 Moving Scores

When you create an anomaly score it will be assigned to the same project where the original anomaly is located. You cannot move scores between projects as you do with other resources.

6.9 Stopping Scores Creation

Batch anomaly scores are asynchronous resources so you can stop their creation before the task is finished. You can use the DELETE BATCH SCORE option from the **1-click action menu**. (See Figure 6.32.)

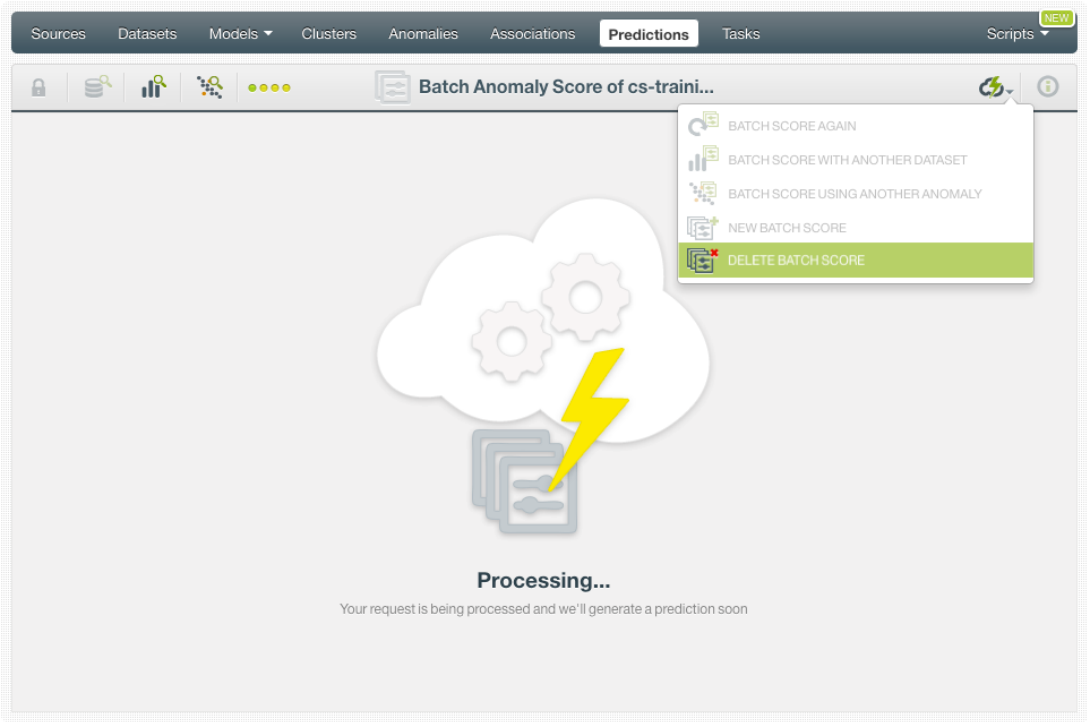

Figure 6.32: Stop batch anomaly score from the 1-click action menu

Alternatively, you can use the DELETE BATCH SCORE from the **pop up menu** on the anomaly score list view. (See Figure 6.33.)

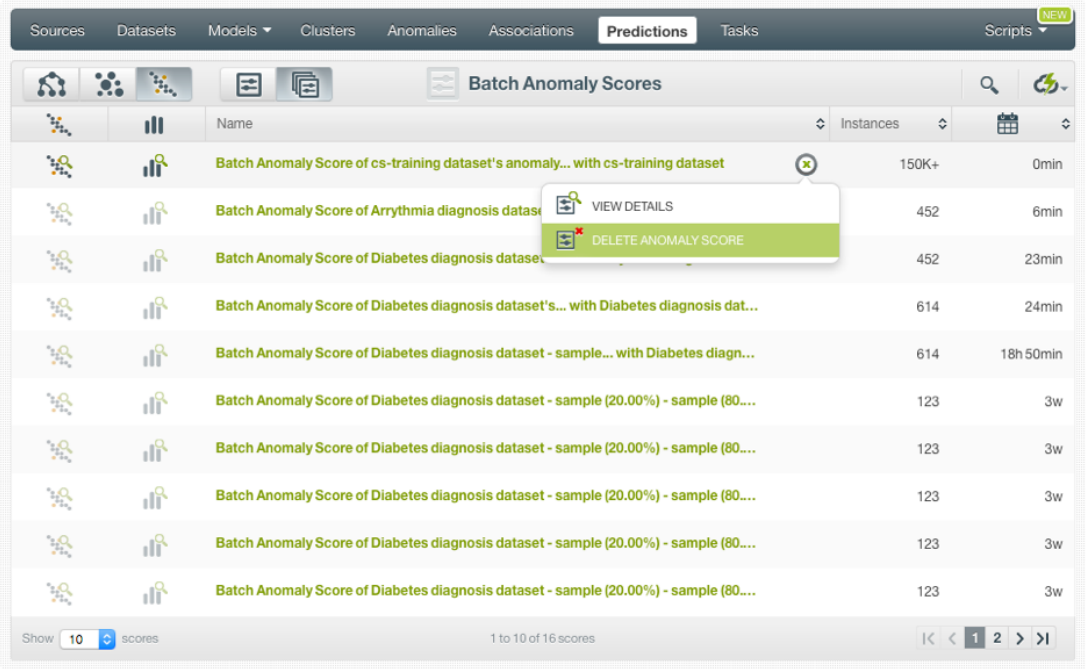

Figure 6.33: Stop batch anomaly score from the pop up menu

Note: if you stop the scoring during its creation you will not be able to resume the same task again. So if you want to create the same anomaly score you will have to start a new task.

6.10 Deleting Anomaly Scores

You can delete your single and batch anomaly scores by clicking on the DELETE ANOMALY SCORE or DELETE BATCH SCORE option in the **1-click action menu** (see Figure 6.34).

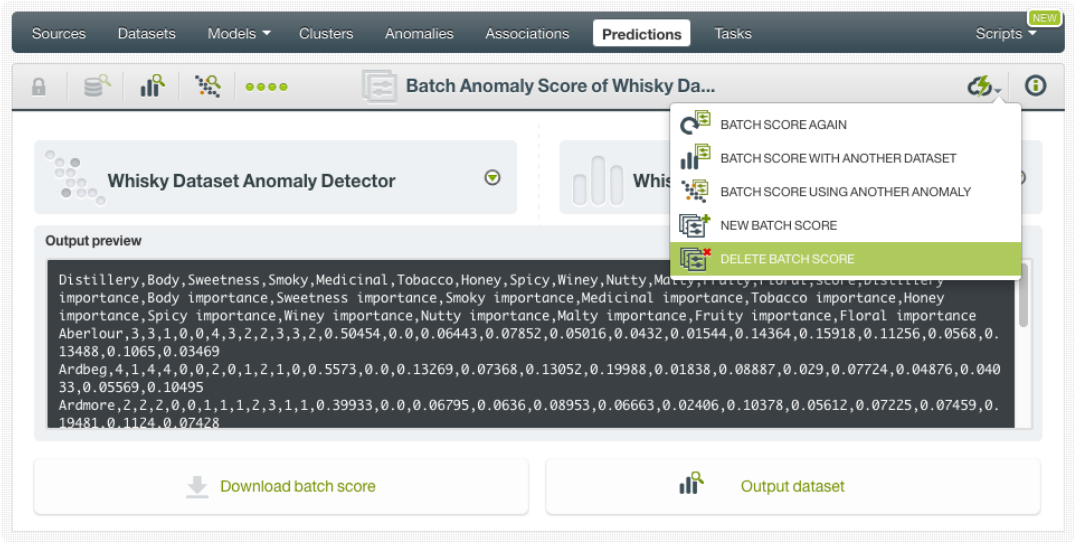

Figure 6.34: Delete batch anomaly score from the 1-click menu

Alternatively, you can click the DELETE ANOMALY SCORE in the **pop up menu** from the score list view (see Figure 6.35.)

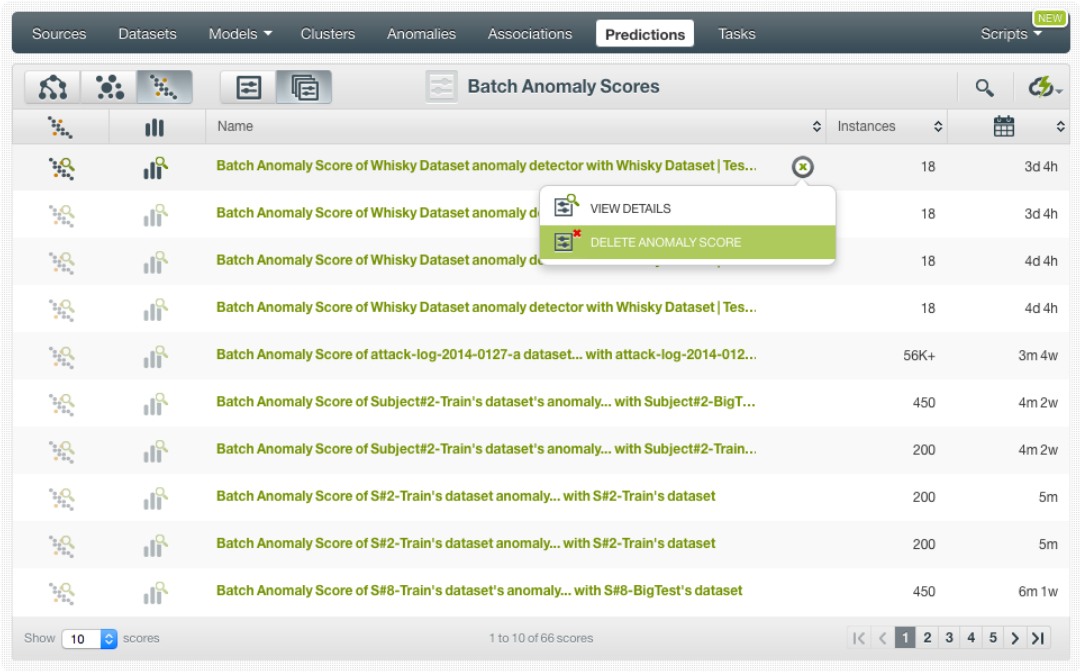

Figure 6.35: Delete batch anomaly score from pop up menu

A modal window will be displayed asking you for confirmation. Once a score is deleted, it is permanently deleted and there is no way you (or even the IT folks at BigML) can retrieve it. (Figure 6.36.)

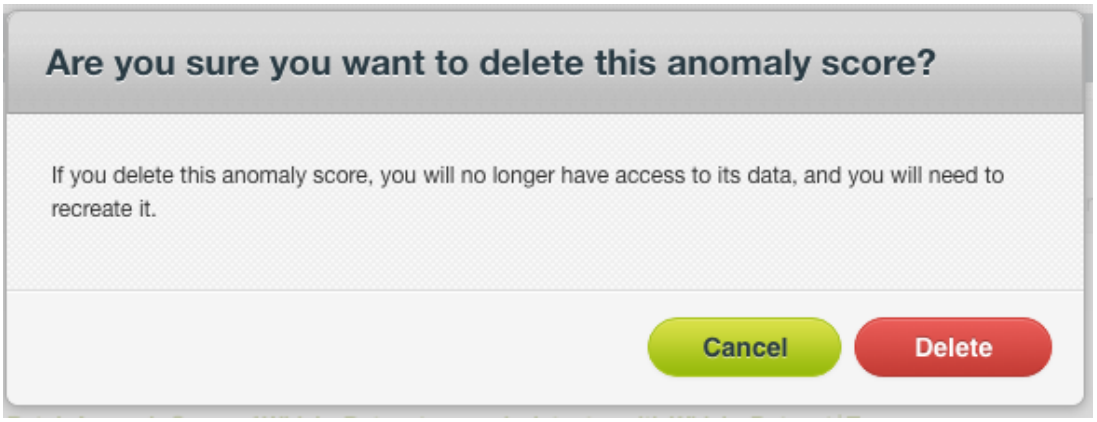

Figure 6.36: Delete score confirmation

CHAPTER 7

Consuming Anomalies

Similarly to other models in BigML, you can **download** anomalies and use them locally to make predictions. You can also create and consume your anomalies programmatically via the **BigML API and bindings**. The following subsections explain those three options.

7.1 Downloading Anomalies

You can download your anomaly in a number of languages including Python, JSON PML or Node.js. Just click on the DOWNLOAD ACTIONABLE ANOMALY menu option and select your preferred language. (See Figure 7.1.)

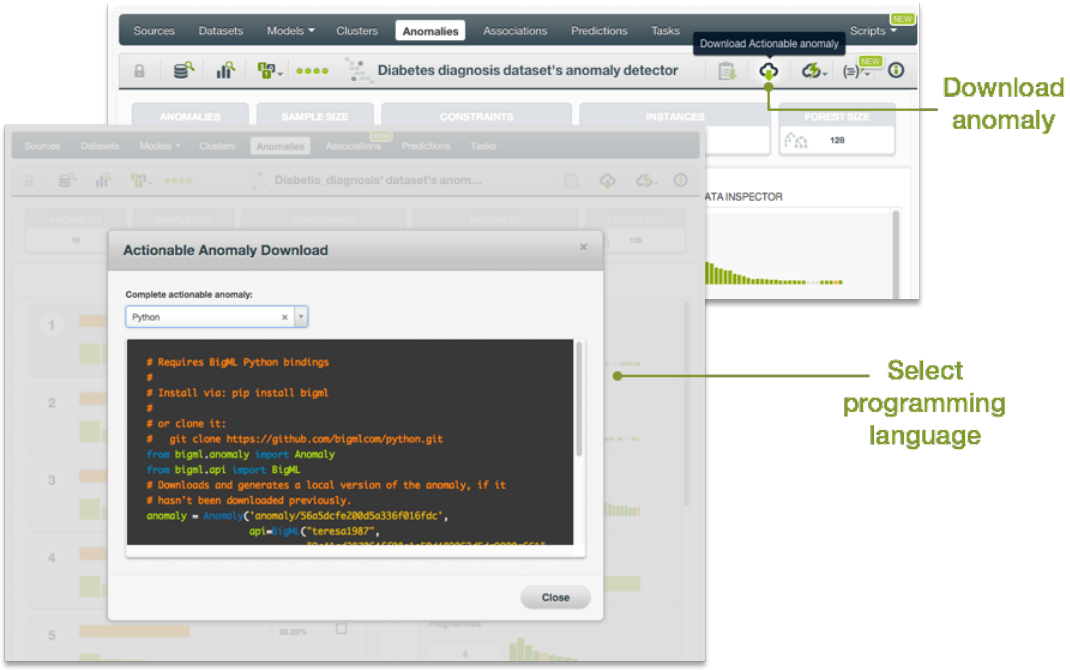

Figure 7.1: Downloading anomaly detector

You can predict the anomaly score for your new data locally, free of latency, and at no cost by downloading your anomalies. It works the same way as local predictions for models and ensembles.

7.2 Using Anomalies Via the BigML API

Anomalies have full citizenship in the BigML API which allows you to programmatically create, configure, retrieve, list, update, delete, and use them to score new data.

See in the example below how to create an anomaly using an existing dataset after you have properly set the `BIGML_AUTH` environment variable to contain your authentication credentials:

```
curl "https://bigml.io/anomaly?$BIGML_AUTH" \
     -X POST \
     -H 'content-type: application/json' \
     -d '{"dataset": "dataset/50650bdf3c19201b64000322"}'
```

For more information on using anomalies through the BigML API, please refer to anomalies REST API documentation[1].

7.3 Using Anomalies Via the BigML Bindings

You can also create and use anomalies via **BigML bindings** which are libraries aimed to make it easier to use the BigML API from your language of choice. BigML offers bindings in multiple languages including Python, Node.js, Java, Swift and Objective-C. You can find below an example to create an anomaly with the Python bindings.

```
from bigml.api import BigML
api = BigML()
anomaly = api.create_anomaly('dataset/57506c472275c1666b004b10')
```

For more information on BigML bindings, please refer to the bindings page[2].

[1] https://bigml.com/api/anomalies
[2] https://bigml.com/tools/bindings

CHAPTER 8

Anomalies Limits

BigML anomalies have a few limitations regarding the type of input data they can support. BigML also impose some limits on the configurable parameters to create an anomaly. You can find all limits listed below:

- **Categorical fields**: a maximum number of 1,000 distinct classes per field is allowed.
- **Text fields**: anomalies do not support text as input fields.
- **Item fields**: anomalies do not support items as input fields.
- **Top Anomalies**: a maximum of top 1,024 anomalies is allowed. You can score all your dataset instances by performing a batch anomaly score.
- **Forest size**: a maximum of 256 trees for the Isolation Forest is allowed.

CHAPTER 9

Anomalies Descriptive Information

Anomalies have an associated **name**, **description**, **category**, and **tags**. The following subsections briefly describe each concept. See in Figure 9.1 the options under MORE INFO menu to edit anomalies.

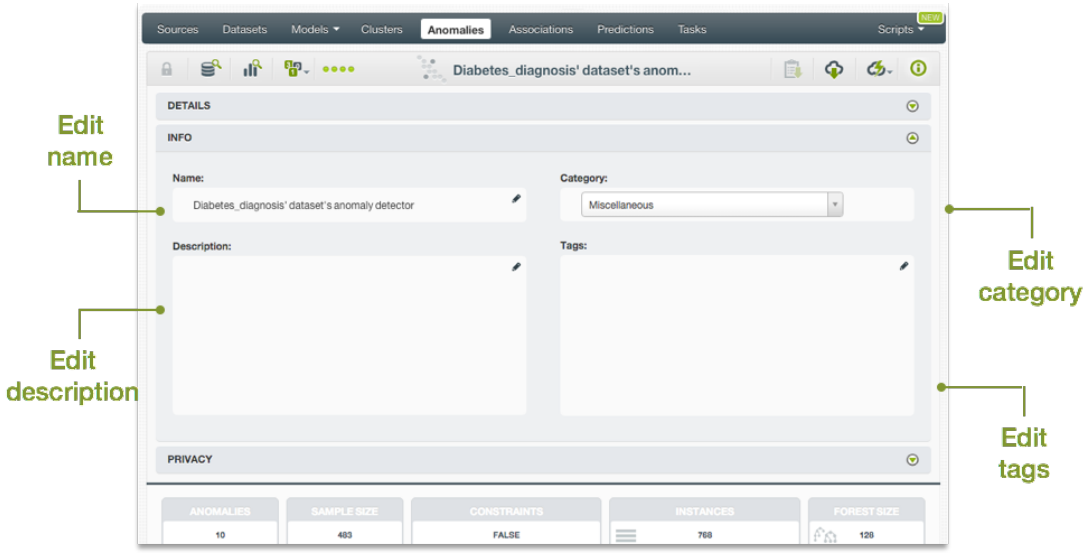

Figure 9.1: Editing anomalies

9.1 Anomalies Name

Each anomaly has a name that is displayed in the anomalies list view and also on the top bar of the anomalies view. Anomalies names are indexed to be used in searches. When you create an anomaly, it gets a default name. You can change it using the MORE INFO menu option on the right corner of the anomalies view. The name of an anomaly cannot be longer than **256** characters. More than one anomaly can have the same name even within the same project, but they will always have different identifiers.

9.2 Description

Each anomaly also has a **description** that is very useful for documenting your Machine Learning projects. Anomalies take the description of the datasets used to create them by default.

Descriptions can be written using plain text and also markdown[1]. BigML provides a simple markdown editor that accepts a subset of markdown syntax. (See Figure 9.2.)

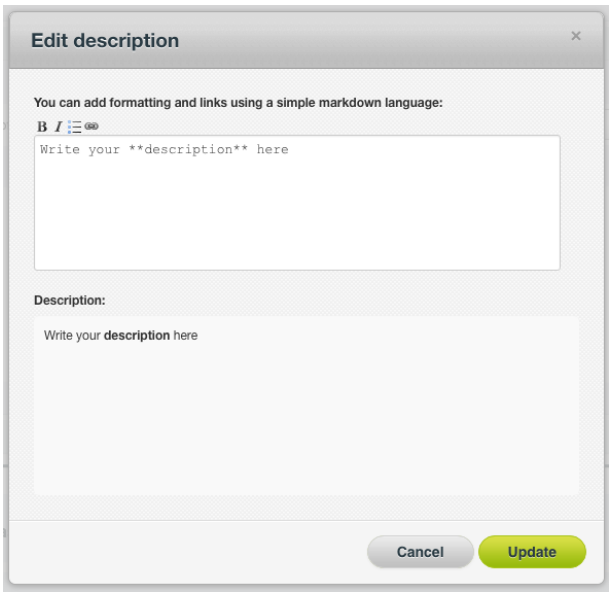

Figure 9.2: Markdown editor for anomalies descriptions

Descriptions cannot be longer than **8192** characters and can use almost any character.

9.3 Category

A **category**, taken from the dataset used to create it, is associated with each anomaly. Categories are useful to classify anomalies according to the domain which your data comes from. This is useful when you use BigML to solve problems across industries or multiple customers.

An anomaly category must be one of the **24** categories listed in Table 9.1.

[1] https://en.wikipedia.org/wiki/Markdown

Table 9.1: Categories used to classify anomalies by BigML

Category
Aerospace and Defense
Automotive, Engineering and Manufacturing
Banking and Finance
Chemical and Pharmaceutical
Consumer and Retail
Demographics and Surveys
Energy, Oil and Gas
Fraud and Crime
Healthcare
Higher Education and Scientific Research
Human Resources and Psychology
Insurance
Law and Order
Media, Marketing and Advertising
Miscellaneous
Physical, Earth and Life Sciences
Professional Services
Public Sector and Nonprofit
Sports and Games
Technology and Communications
Transportation and Logistics
Travel and Leisure
Uncategorized
Utilities

9.4 Tags

An anomaly can also have a number of **tags** associated with it that can help to retrieve it via the BigML API or to provide anomalies with some extra information. Anomalies inherit the tags from the dataset used to create them. Each tag is limited to a maximum of 128 characters. Each anomaly can have up to **32** different tags.

9.5 Counters

For each anomaly, BigML also stores a number of counters to track the number of other resources that have been created using the anomaly as a starting point. Display the counters by mousing over the menu option at the top of the anomaly view. Click on VIEW # ANOMALY SCORE FROM THIS ANOMALY menu option to quickly access the single anomaly scores and VIEW # BATCH ANOMALY SCORE FROM THIS ANOMALY to see all batch anomaly scores. (See Figure 9.3.)

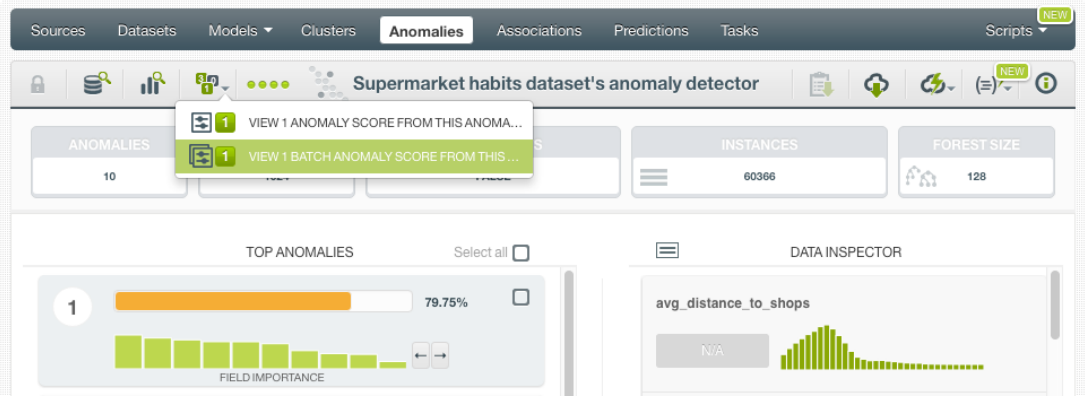

Figure 9.3: Counters for anomalies

CHAPTER 10

Anomalies Privacy

Privacy options for anomalies can be defined in the **More Info** menu option. (See Figure 10.1.) There are two levels of privacy for BigML anomalies:

- **Private**: only accessible by authorized users (the owner and those who have been granted access by him or her).
- **Shared**: by enabling the **secret link** you will get two different links to share your anomalies. The first one is a sharing link that you can copy and send to others so they can visualize and interact with your anomalies. The second one is a link to embed your anomalies directly on your web page. This is very useful if you want to make local scoring predictions at no cost.

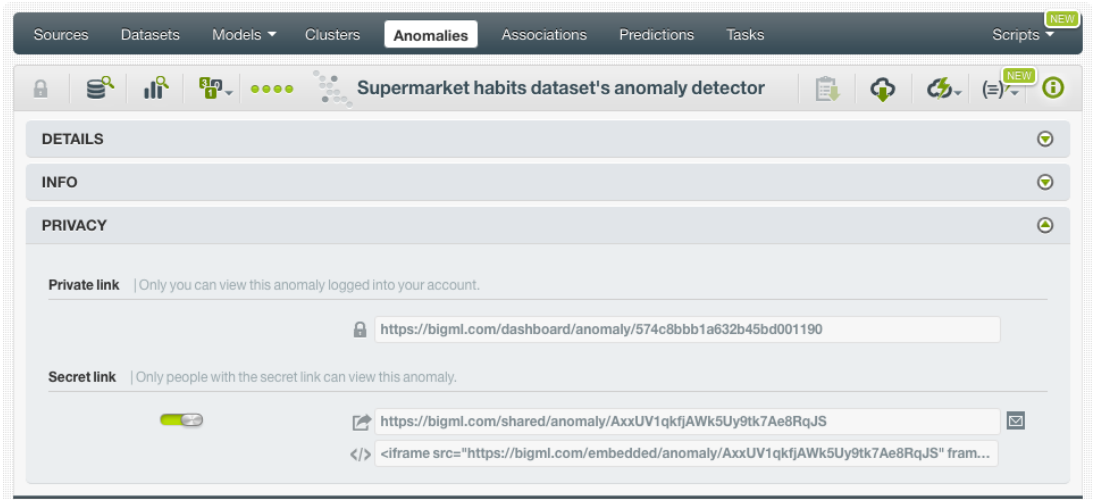

Figure 10.1: Anomalies privacy

CHAPTER 11

Moving Anomalies

When you create an anomaly, it will be assigned to the same project where the original dataset is located. Anomalies can only be assigned to a single project. However, you can move anomalies between projects. The menu option to do this can be found in two places:

1. Click MOVE TO... within the **1-click action menu** from the anomaly view. (See Figure 11.1.)

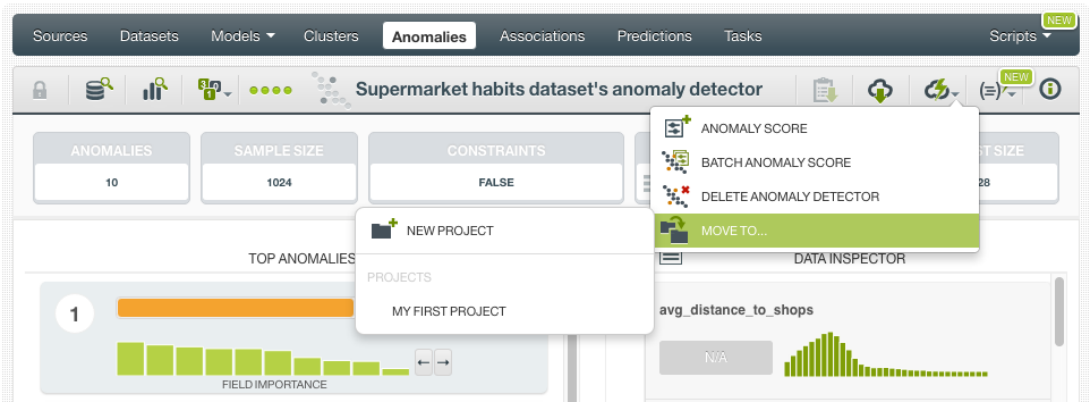

Figure 11.1: Change project from 1-click menu

2. Click MOVE TO... within the **pop up menu** from the anomaly list view. (See Figure 11.2)

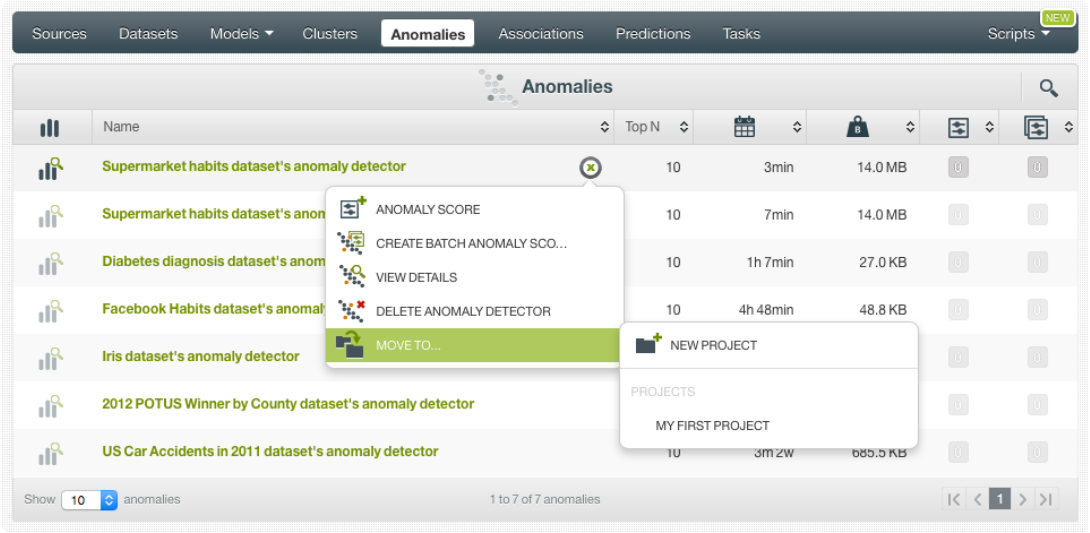

Figure 11.2: Change project from pop up menu

CHAPTER 12

Stopping Anomalies Creation

You can stop the creation of an anomaly before the task is finished by clicking the DELETE ANOMALY DETECTOR option in the **1-click action menu** from the anomaly view. (See Figure 12.1.)

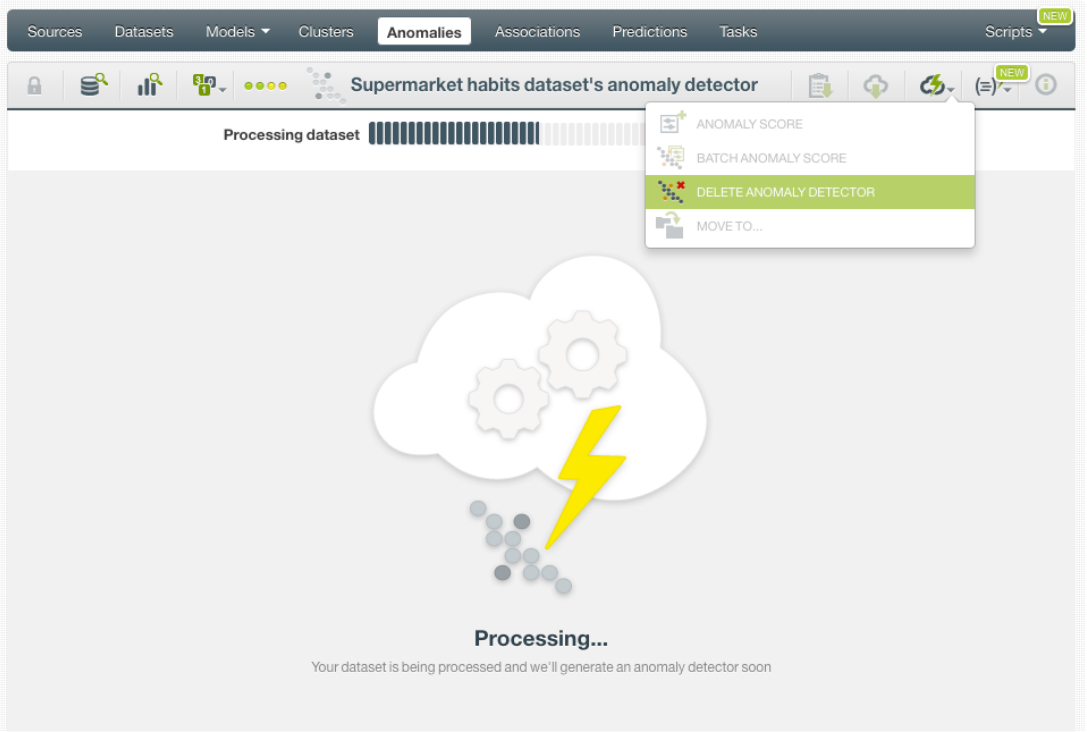

Figure 12.1: Stop anomalies creation from 1-click menu

Alternatively, click the DELETE ANOMALY DETECTOR in the **pop up menu** from the anomaly list view. (See Figure 12.2.)

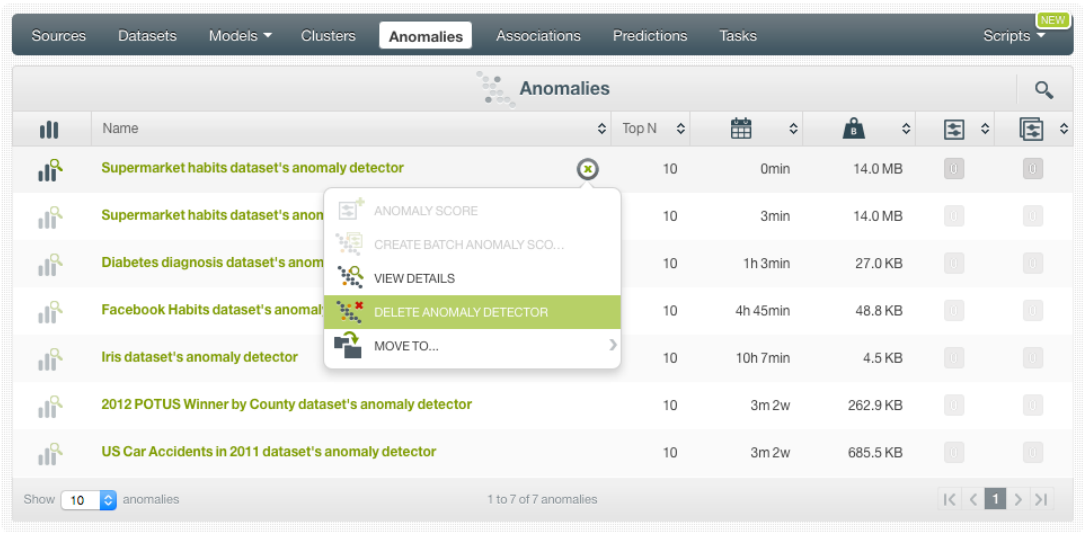

Figure 12.2: Stop anomalies creation from pop up menu

Note: if you stop the anomaly during its creation, you will not be able to resume the same task. If you want to create the same anomaly, you will have to start a new task.

CHAPTER 13

Deleting Anomalies

You can delete your anomalies by clicking the DELETE ANOMALY DETECTOR option in the **1-click action menu** from the anomaly view. (See Figure 13.1.)

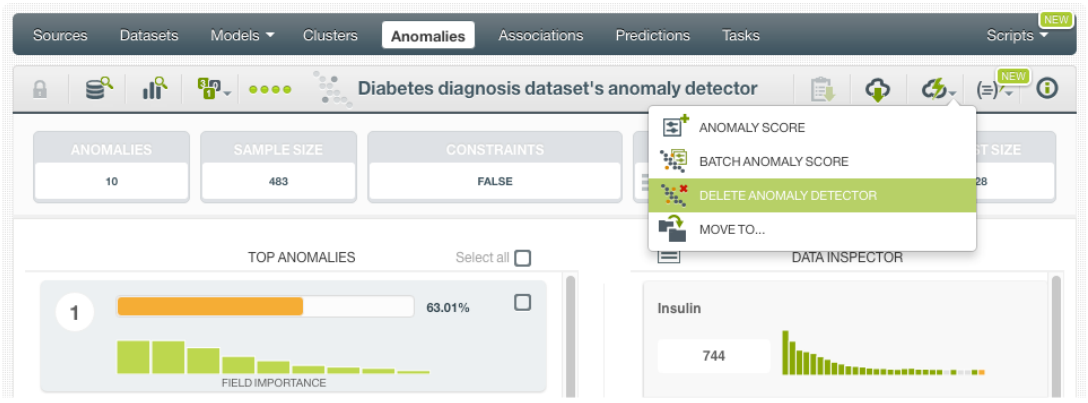

Figure 13.1: Delete anomalies from 1-click menu

Alternatively, click the DELETE ANOMALY DETECTOR in the **pop up menu** from the anomaly list view. (See Figure 13.2.)

Figure 13.2: Delete anomalies from pop up menu

A modal window will be displayed asking you for confirmation. After an anomaly is deleted, it is permanently deleted, and there is no way you (or even the IT folks at BigML) can retrieve it.

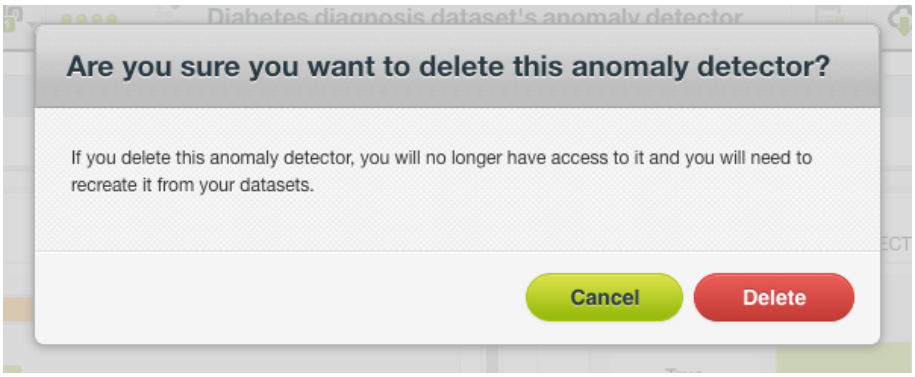

Figure 13.3: Confirmation message to delete an anomaly

CHAPTER 14

Takeaways

This document covered anomalies in detail. We conclude it with a list of key points:

- Anomaly dectection is an **unsupervised** learning method used to detect instances that do not follow a regular pattern.
- BigML anomaly use an optimized implementation of the **Isolation Forest** algorithm, a highly scalable and efficient method that usually yields the best results compared to other anomaly detection techniques.
- BigML computes an anomaly score for each instance and a measure to indicate the relative contribution of each input field to the anomaly score.
- BigML anomalies support categorical and numeric fields as inputs, text and items fields will not be taken into account to compute the anomaly score.
- BigML anomalies also supports missig data.
- To create anomalies you just need an existing **dataset**. Then anomalies can be used to make a single score prediction or a batch score prediction. Additionally, you can create a dataset from anomalies. (See Figure 14.1.)
- You can use the **1-click option** to create your anomaly or you can **configure** the several parameters provided by BigML before.
- When the anomaly has been created, you get a list of your TOP ANOMALIES ranked by score.
- You can inspect your anomalous instances values in the DATA INSPECTOR.
- You can create a new **dataset** removing your anomalous instances or including them.
- You can use your anomaly to **score** single or multiple instances in batch not seen before by the model.
- You can create, configure, update, and use your anomalies programmatically via the **BigML API and bindings**.
- You can download your anomalies to **locally** score your new instances.
- You can add **descriptive information** to your anomalies.
- You can **move** your anomalies between projects.
- You can **share** your anomalies with other people using the secret link or embedding them into your own applications.
- You can **stop** your anomalies creation by deleting them.
- You can permanently **delete** your existing anomalies.

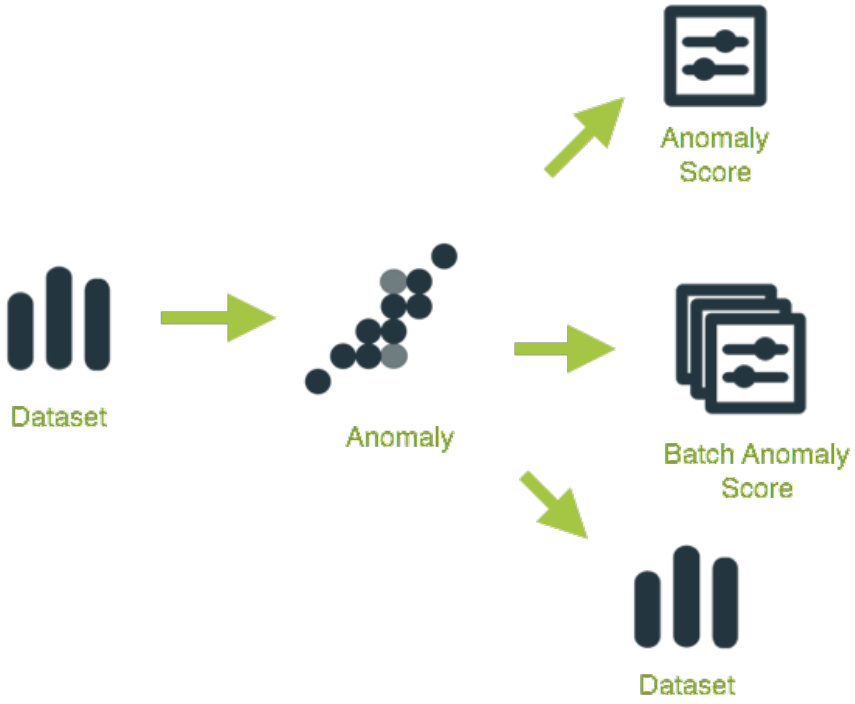

Figure 14.1: Anomalies workflow

List of Figures

List of Tables

Glossary

Anomaly Score an anomaly detector assigns an anomaly score to each instance of th input dataset. Additionally, you can use an anomaly detector to calculate the anomaly score of new data instances. An anomaly score is a percentage between 0% and 100%, with higher scores indicating higher anomaly. ii, 1, 12, 44

Anomaly Detection an unsupervised Machine Learning task which identifies instances in a dataset that do not conform to a regular pattern. ii

Dashboard The BigML web-based interface that helps you privately navigate, visualize, and interact with your modeling resources. ii, 1, 25

Decision Trees a class of Machine Learning algorithms used to solve regression and classification problems. Decision trees are composed of nodes and branches that create a model of decisions with a tree graph. Nodes represent the predictors or labels that have an influence in the predictive path, and the branches represent the rules followed by the algorithm to make a given prediction. 3

Ensembles a class of Machine Learning algorithms in which multiple independent classifiers or regressors are trained, and the combination of these classifiers is used to predict an objective field. An ensemble of models built on samples of the data can become a powerful predictor by averaging away the errors of each individual model. 3

Instances the data points that represent the entity you want to model, also known as observations or examples. They are usually the rows in your data with a value (potentially missing) for each field that describes the entity. 1

Local predictions the predictions made in your local environment, faster, at no cost, by downloading your model. 44

Project an abstract resource that helps you group related BigML resources together. 2, 22, 52

Supervised learning a type of Machine Learning problem in which each instance of the data has a label. The label for each instance is provided in the training data, and a supervised Machine Learning algorithm learns a function or model that will predict the label given all other features in the data. The function can then be applied to data unseen during training to predict the label for unlabeled instances. ii

Unsupervised learning a type of Machine Learning problem in which the objective is not to learn a predictor, and thus does not require each instance to be labeled. Typically, unsupervised learning algorithms infer some summarizing structure over the dataset, such as a clustering or a set of association rules. ii, 1

References

[1] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. *Isolation Forest Algorithm*. Jan. 2016. URL: https://feitonyliu.files.wordpress.com/2009/07/liu-iforest.pdf.

[2] The BigML Team. *Anomaly Detection with the BigML Dashboard*. Tech. rep. BigML, Inc., Jan. 2016.

[3] The BigML Team. *Association Discovery with the BigML Dashboard*. Tech. rep. BigML, Inc., Dec. 2015.

[4] The BigML Team. *Classification and Regression with the BigML Dashboard*. Tech. rep. BigML, Inc., May 2016.

[5] The BigML Team. *Cluster Analysis with the BigML Dashboard*. Tech. rep. BigML, Inc., May 2016.

[6] The BigML Team. *Datasets with the BigML Dashboard*. Tech. rep. BigML, Inc., Jan. 2016.

[7] The BigML Team. *Sources with the BigML Dashboard*. Tech. rep. BigML, Inc., Jan. 2016.

[8] The BigML Team. *Time Series with the BigML Dashboard*. Tech. rep. BigML, Inc., July 2017.

[9] The BigML Team. *Topic Models with the BigML Dashboard*. Tech. rep. BigML, Inc., Nov. 2016.


bigml