

# Лабораторная работа 6

## «AutoML с BigML»

Автоматизированное машинное обучение, которое также называется автоматизированным ML или AutoML, представляет собой процесс автоматизации трудоемких и многократно повторяющихся задач разработки моделей машинного обучения. С его помощью специалисты по обработке и анализу данных могут создавать модели машинного обучения с высокой масштабируемостью, эффективностью и производительностью, сохраняя при этом качество модели.

**BigML** один из сервисов AutoML.

***Внимание! Сервис доступен для бесплатного использования только 14 дней***

Он предоставляет набор надежных алгоритмов машинного обучения, которые доказали свою эффективность в решении реальных задач. BigML обеспечивает неограниченное количество приложений прогнозирования в различных отраслях, включая аэрокосмическую, автомобильную, энергетическую, развлекательную, финансовые услуги, продукты питания, здравоохранение, Интернет вещей, фармацевтику, транспорт, телекоммуникации и многое другое.

Доступные типы алгоритмов:

- обучение с учителем: классификация и регрессия (деревья, ансамбли, линейные регрессии, логистические регрессии, глубокие сети) и прогнозирование временных рядов;
- обучение без учителя: кластерный анализ, обнаружение аномалий, тематическое моделирование, обнаружение ассоциаций и анализ главных компонент (PCA).

Все прогнозные модели на BigML оснащены функциями интерактивной визуализации и пояснения, которые делают их интерпретируемыми. Их можно экспортировать и использовать для локального автономного прогнозирования на любом периферийном вычислительном устройстве или мгновенно развертывать как часть распределенных производственных приложений в реальном времени.

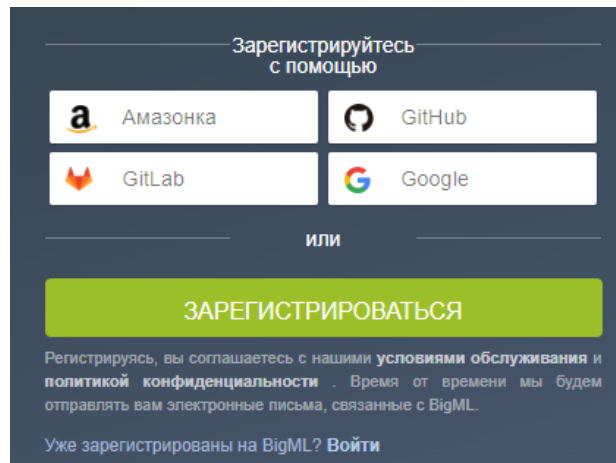
Модели BigML полностью экспортируются через JSON PML (и PMML) и могут использоваться со всеми популярными языками программирования. Это означает, что вы можете легко подключать свои модели к веб-приложениям, мобильным приложениям или сервисам IoT.

Будучи компанией, ориентированной на API, BigML сначала внедряет каждую новую функцию в REST API. Привязки и библиотеки доступны для всех популярных языков, включая Python, Node.js, Ruby, Java, Swift и другие.

Все ресурсы BigML являются неизменяемыми и хранятся с уникальным идентификатором и параметрами создания, что позволяет вам отслеживать любой рабочий процесс машинного обучения в любое время.

## Задание 1.

1. Перейдите на сайт сервиса: <https://bigml.com/>
2. Зарегистрироваться можно, например, с помощью google-аккаунта



Зарегистрируйтесь  
с помощью

Амазонка GitHub

GitLab Google

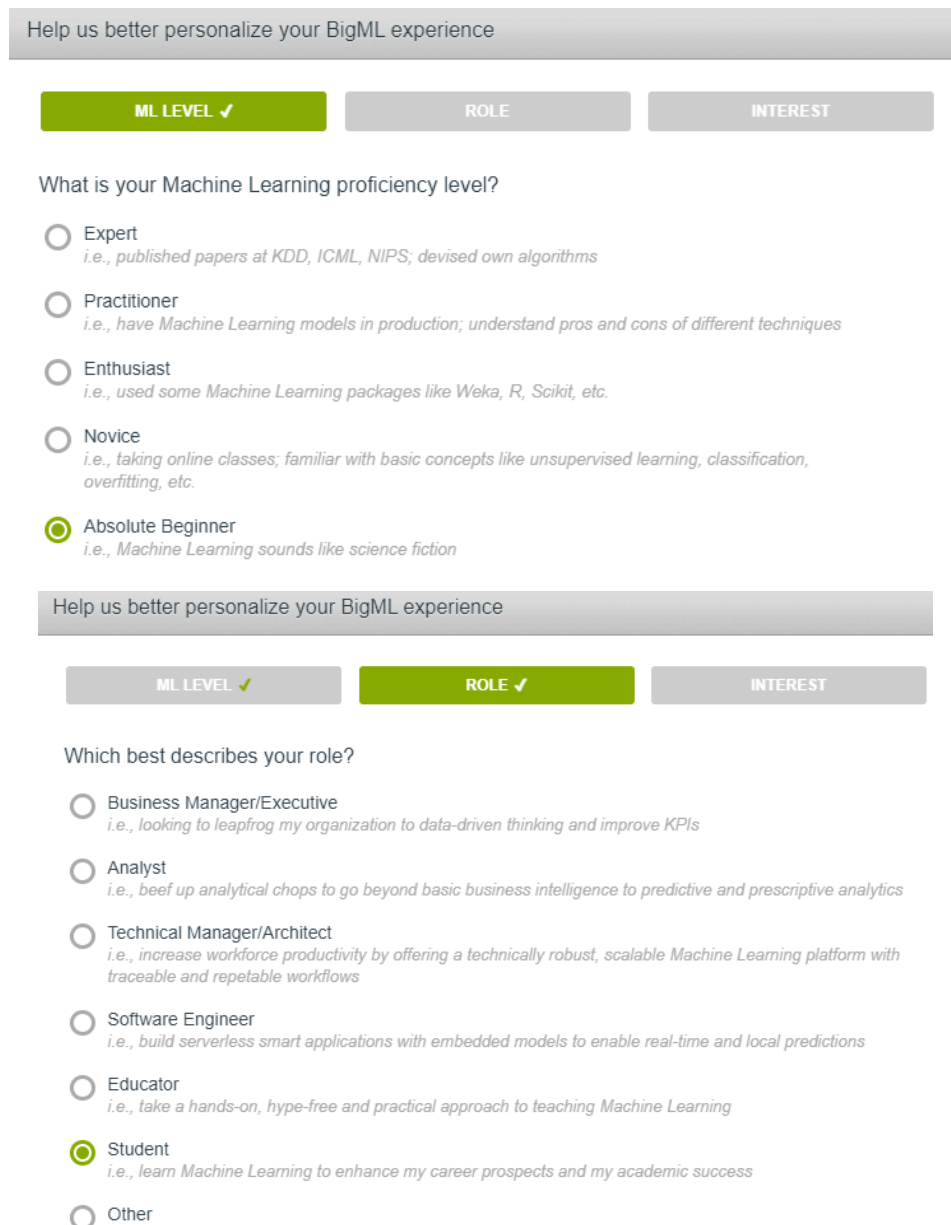
или

**ЗАРЕГИСТРИРОВАТЬСЯ**

Регистрируясь, вы соглашаетесь с нашими **условиями обслуживания** и **политикой конфиденциальности**. Время от времени мы будем отправлять вам электронные письма, связанные с BigML.

Уже зарегистрированы на BigML? **Войти**

3. Далее предлагается заполнить небольшую анкету:



Help us better personalize your BigML experience

**ML LEVEL ✓** ROLE INTEREST

What is your Machine Learning proficiency level?

☐ Expert  
*i.e., published papers at KDD, ICML, NIPS; devised own algorithms*

☐ Practitioner  
*i.e., have Machine Learning models in production; understand pros and cons of different techniques*

☐ Enthusiast  
*i.e., used some Machine Learning packages like Weka, R, Scikit, etc.*

☐ Novice  
*i.e., taking online classes; familiar with basic concepts like unsupervised learning, classification, overfitting, etc.*

☒ **Absolute Beginner**  
*i.e., Machine Learning sounds like science fiction*

Help us better personalize your BigML experience

ML LEVEL ✓ **ROLE ✓** INTEREST

Which best describes your role?

☐ Business Manager/Executive  
*i.e., looking to leapfrog my organization to data-driven thinking and improve KPIs*

☐ Analyst  
*i.e., beef up analytical chops to go beyond basic business intelligence to predictive and prescriptive analytics*

☐ Technical Manager/Architect  
*i.e., increase workforce productivity by offering a technically robust, scalable Machine Learning platform with traceable and repeatable workflows*

☐ Software Engineer  
*i.e., build serverless smart applications with embedded models to enable real-time and local predictions*

☐ Educator  
*i.e., take a hands-on, hype-free and practical approach to teaching Machine Learning*

☒ **Student**  
*i.e., learn Machine Learning to enhance my career prospects and my academic success*

☐ Other

Help us better personalize your BigML experience

ML LEVEL ✓

ROLE ✓

INTEREST ✓

Why are you interested in BigML?

- ☐ Adopt a Machine Learning platform at my company  
*i.e., establish a standard framework to accelerate actionable insights and streamline Machine Learning in production*
- ☐ Implement a specific analytical use case or application  
*i.e., solve a real-life business problem by applying Machine Learning instead of hard-coded business rules*
- ☒ Find the best tool to teach Machine Learning to my students  
*i.e., take a hands-on, hype-free and practical approach to teaching Machine Learning*
- ☐ Learn Machine Learning on my own  
*i.e., enhance my analytical skills to become ML-literate and stay ahead of the curve*
- ☐ Other

**Сервис доступен для бесплатного использования только 14 дней**

4. При необходимости можно изучить обучающие видео: <https://bigml.com/education/videos>

5. На вкладке Sources можно найти наборы данных, которые можно использовать для обучения моделей (после нажатия на конкретный набор можно увидеть его описание и содержание).

6. Создайте новый проект

The screenshot shows the BigML dashboard for user SEMENOVASSAU. The top navigation bar includes links for PRODUCT, GETTING STARTED, PRICING, and SUPPORT, along with a user profile and a 'Dashboard' button. The main content area is titled 'SEMENOVASSAU - My Dashboard' and features a 'Sources' tab. A dropdown menu is open, showing options: 'NEW PROJECT', 'VIEW ALL PROJECTS 1', 'NEW ORGANIZATION', and 'DASHBOARD SETTINGS'. Below the menu, a table lists three datasets:

Type	Name	Size	Fields	Path	Image
open, image	grape-strawberry classification	22min	13.3 MB	0	0
open, image	Firetruck Anomalies	22min	8.5 MB	0	0
open, image	Hot Dog Or Not Hotdog	22min	15.8 MB	0	0

Появится следующее окно:

The screenshot shows the 'Project Titanic' window. It includes a header with the project name and a description: 'Построение прогноза: выжил бы человек с конкретными характеристиками или нет, если бы он был пассажиром Титаника'. Below the description, there is a 'Tags' section and a 'Resources' section. The 'Resources' section displays a grid of icons representing various data sources and models, with a 'TOTAL: 0' indicator.



7. Добавьте источник данных в проект. Для этого нажмите на кнопку

Появится следующее окно:

The screenshot shows the WhizzML dashboard for 'SEMENOVASSAU - My Dashboard'. The 'Project Titanic' view is selected. The 'Sources' tab is active, and a dropdown menu is open, showing 'Project Titanic' and 'BigML Intro Project'. The 'Sources' table is empty, showing 'No sources'.


Пока список источников в проекте пуст. Перейдите в полный список источников, из которого выберите Titanic Survival – информация о пассажирах Титаника.

Type	Name			
Image	grape-strawberry classification	open, image, 70 sources, 237 fields (1 categorical, 234 numeric, 1 path, 1 image)	30min	13.3 MB
Image	Firetruck Anomalies	open, image, 118 sources, 514 fields (512 numeric, 1 path, 1 image)	30min	8.5 MB
Image	Hot Dog Or Not Hotdog	open, image, 160 sources, 3 fields (1 categorical, 1 path, 1 image)	30min	15.8 MB
CSV	Country Stats Mashup	open, table, 8 fields (8 numeric)	30min	12.0 KB
TSV	Fictional Wine Sales	open, table, 6 fields (3 categorical, 3 numeric)	30min	51.9 KB
CSV	<u>Titanic Survival</u>	open, table, 5 fields (3 categorical, 2 numeric)	30min	78.0 KB

А уже из самого источника укажите, что вы хотели бы перенести его в наш проект.

The screenshot shows the WhizzML dashboard for 'Titanic Survival'. The 'Sources' tab is active. A context menu is open, showing options like 'NEW PROJECT', 'PROJECTS', 'Project Titanic', and 'BigML Intro Project'. The 'Titanic Survival' source is highlighted.


## 8. Снова вернитесь на страницу проекта



[PRODUCT ▾](#)
[GETTING STARTED](#)
[PRICING ▾](#)
[SUPPORT](#)

SEMENOVASSAU

Dashboard

 SEMENOVASSAU - My Dashboard
 

Project Titanic

Sources





Datasets


Supervised ▾

Unsupervised ▾





Predictions ▾

Tasks

 Titanic Survival

Name	Type	Instance 1	Instance 2	Instance 3
Age	1 2 3	29	0.9	2
Class/Dept	A B C	1st Class	1st Class	1st Class
Fare today	1 2 3	£16,300.00	£11,700.00	£11,700.00
Joined	A B C	Southampton	Southampton	Southampton

 NEW PROJECT
  VIEW ALL PROJECTS 2
  NEW ORGANIZATION
  DASHBOARD SETTINGS

Search by name

Теперь в проекте есть источник данных:

[illegible]

9. На основании источника данных можно создать Dataset – набор данных, на которых будет обучаться наша модель.

Для этого нужно перейти на вкладку Sources проекта и выбрать пункт, позволяющий создать Dataset на основании этого источника

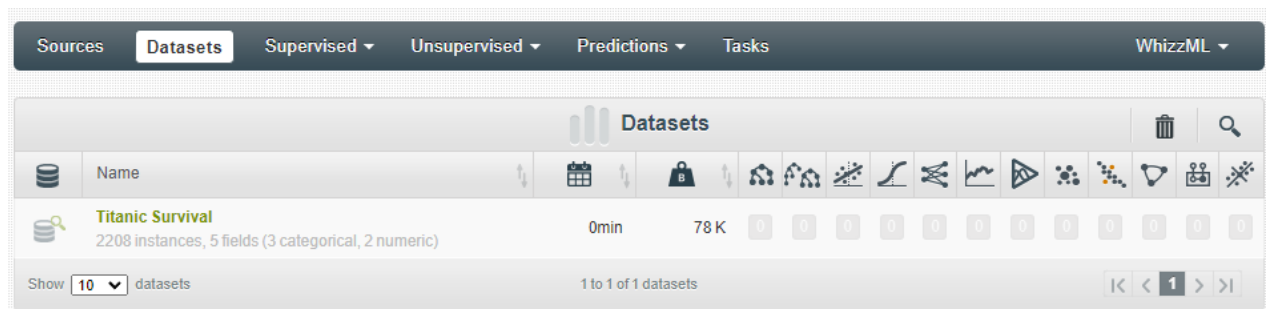
The screenshot shows the WhizzML interface with the 'Sources' tab selected. A context menu is open over the 'Titanic Survival' source. The menu options are:

- CLOSE THIS SOURCE
- 1-CLICK DATASET (highlighted)
- CREATE COMPOSITE WITH THIS SOUR...
- CLONE THIS SOURCE
- VIEW DETAILS
- DELETE SOURCE
- MOVE TO...

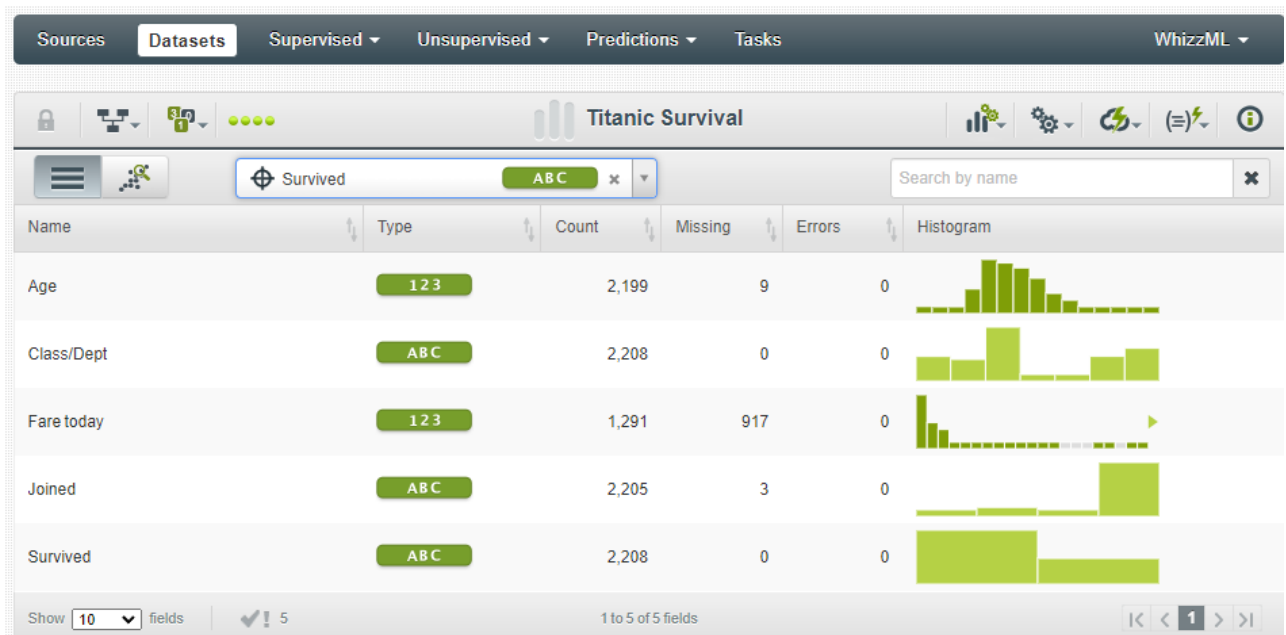
The source details for 'Titanic Survival' are:

- Type: CSV
- Name: Titanic Survival
- Description: open, table, 5 fields (3 categorical, 2 numeric)
- Size: 46min
- Weight: 78.0 KB
- Count: 0

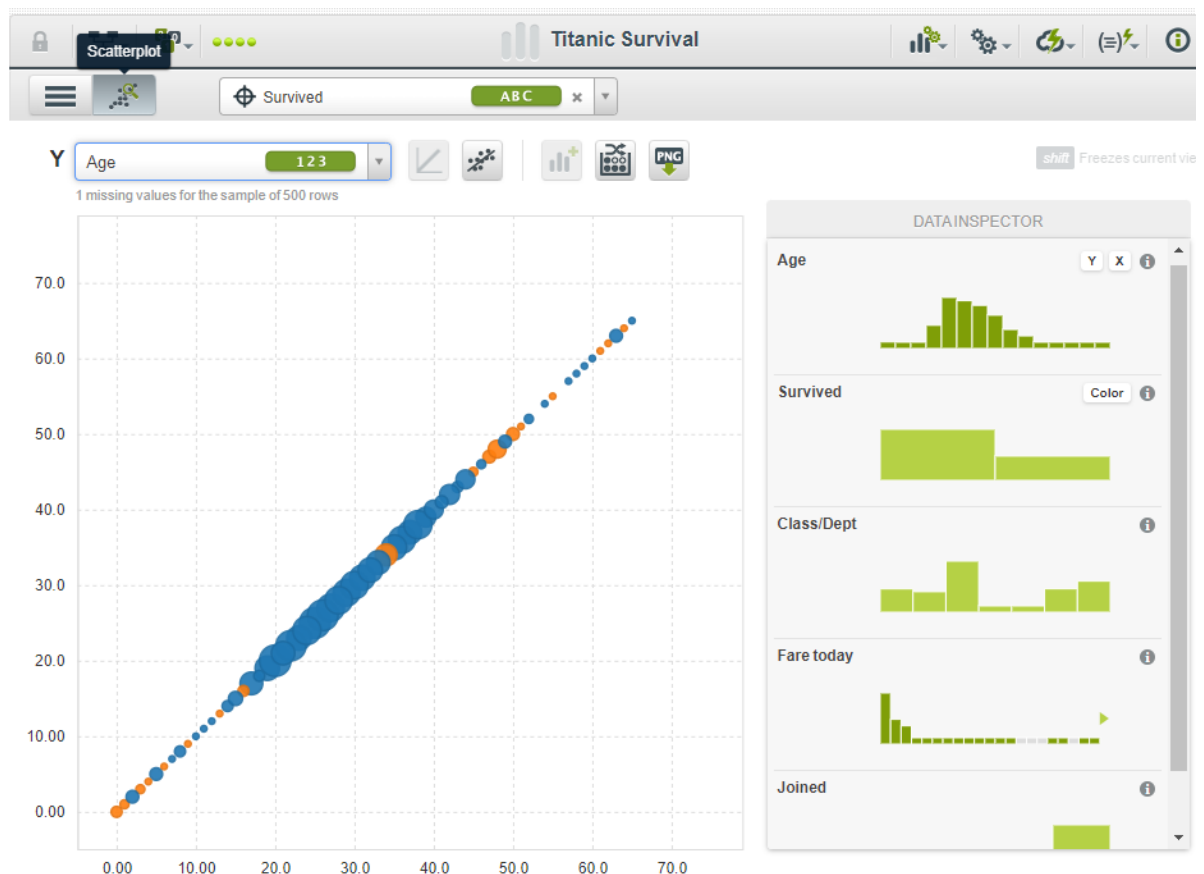
Теперь созданный набор данных можно увидеть и на вкладке Datasets:



Нажав на название можно увидеть детальную информацию о нем:

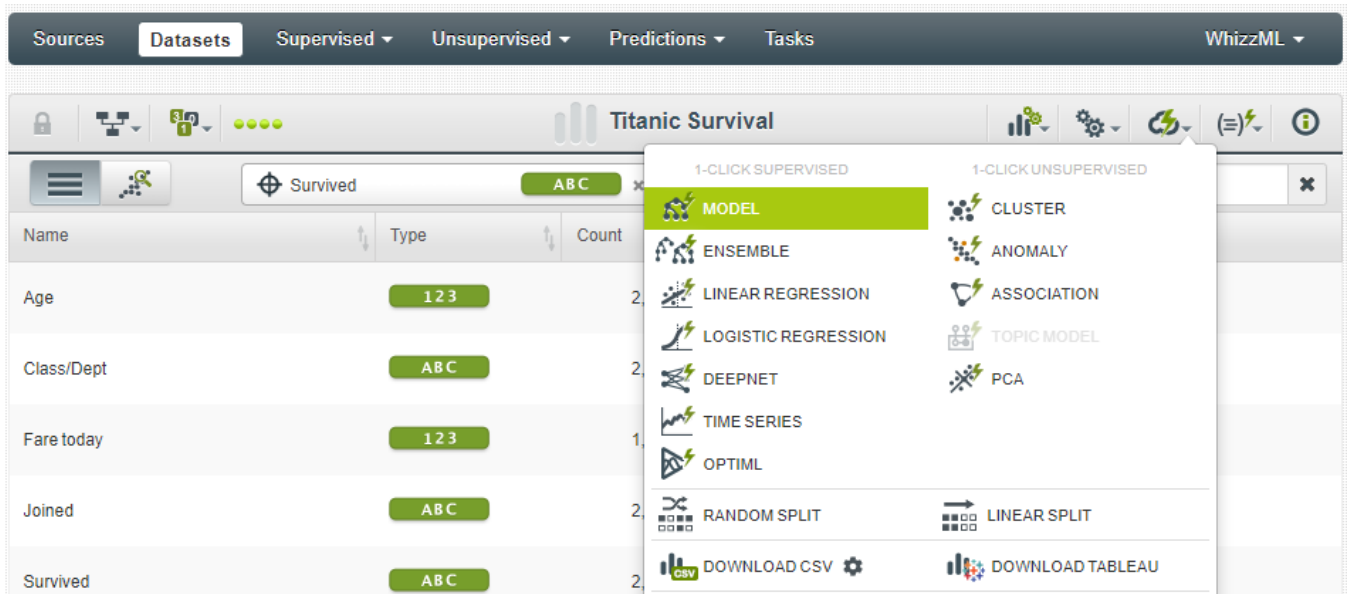


А нажав на соответствующий значок можно увидеть графическую визуализацию выбранного поля набора данных:



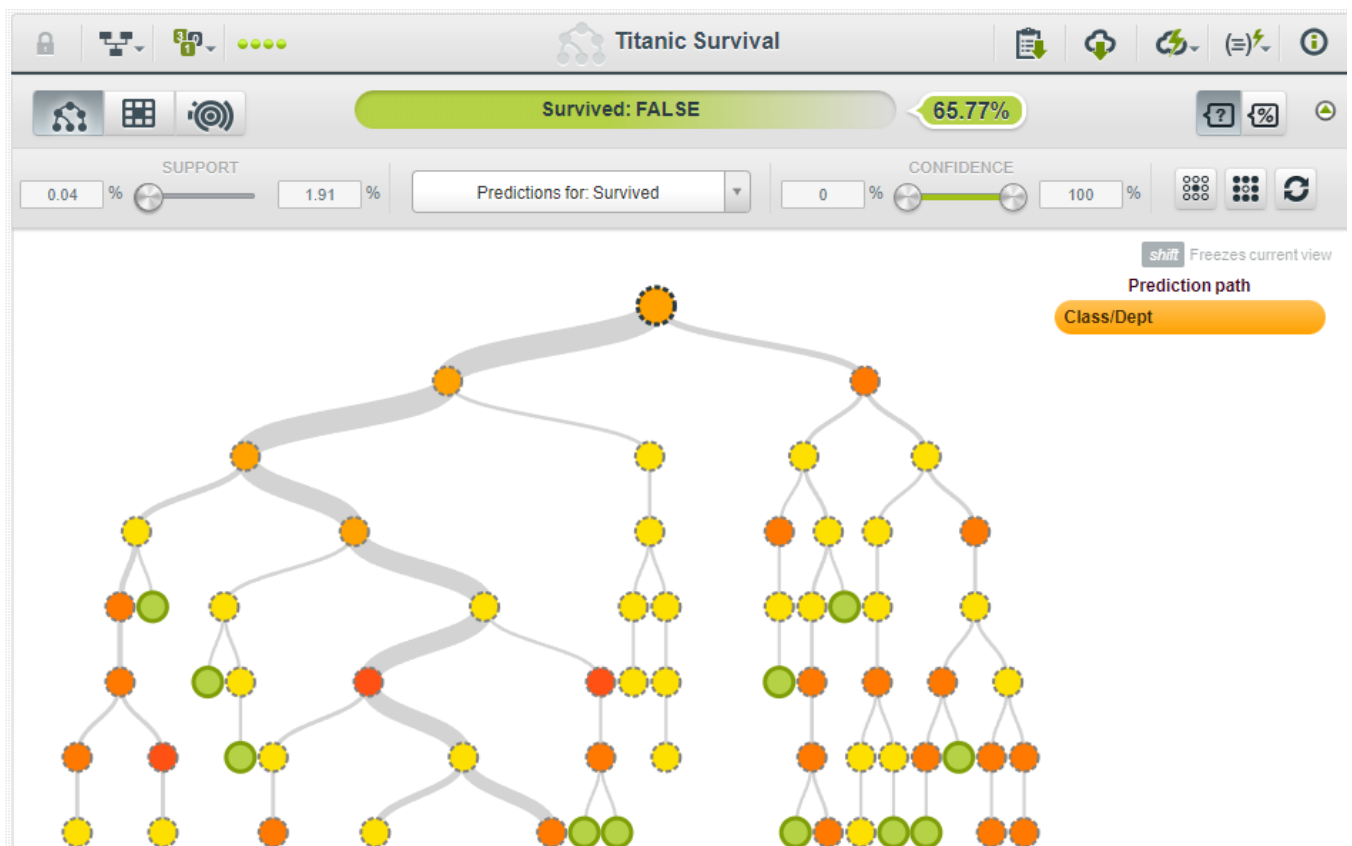
10. Получив набор данных, можно обучить на нем модель.

Для этого на странице Datasets необходимо выбрать соответствующий пункт меню

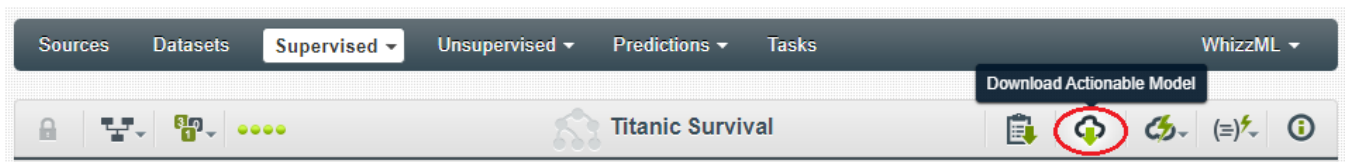


В качестве модели используется Дерево решений.

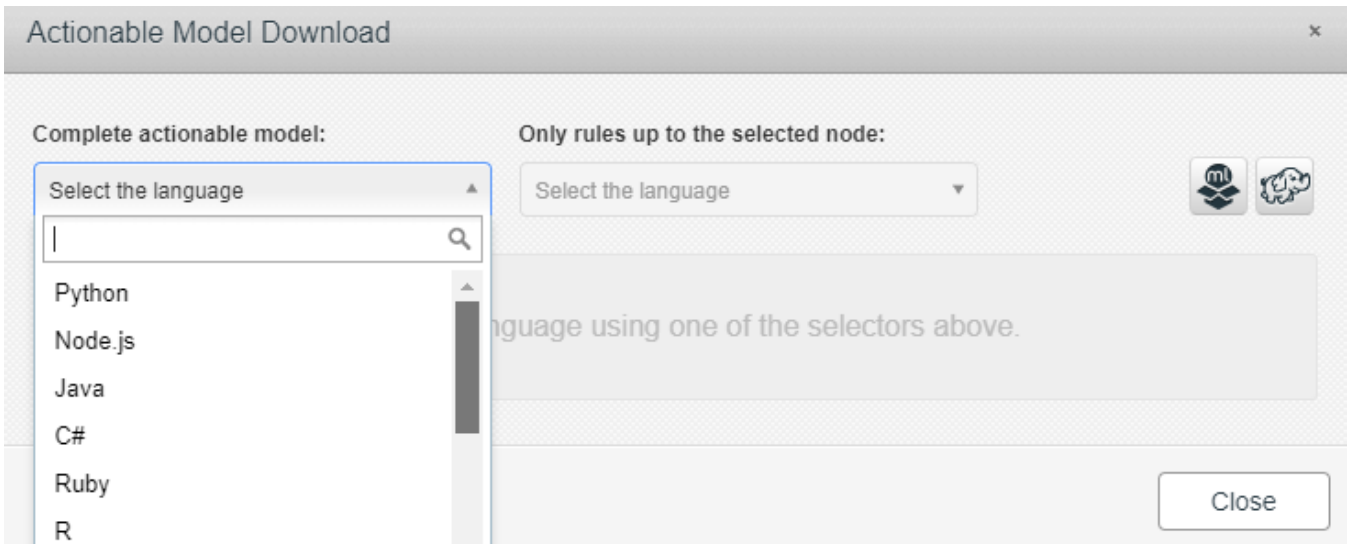
После завершения процесса обучения будет показано получившееся дерево. При наведении мышки на узлы дерева можно увидеть условия, которые в процессе обучения были привязаны к каждому узлу дерева.



11. Нажав на следующую кнопку, модель можно загрузить

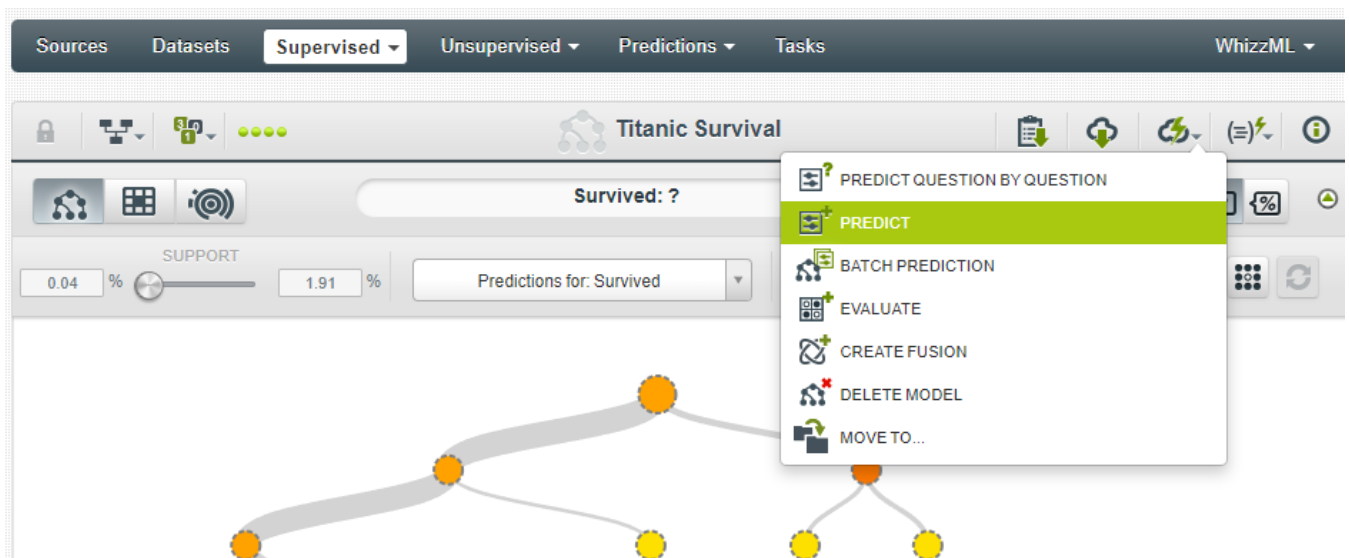


Далее нужно выбрать язык программирования



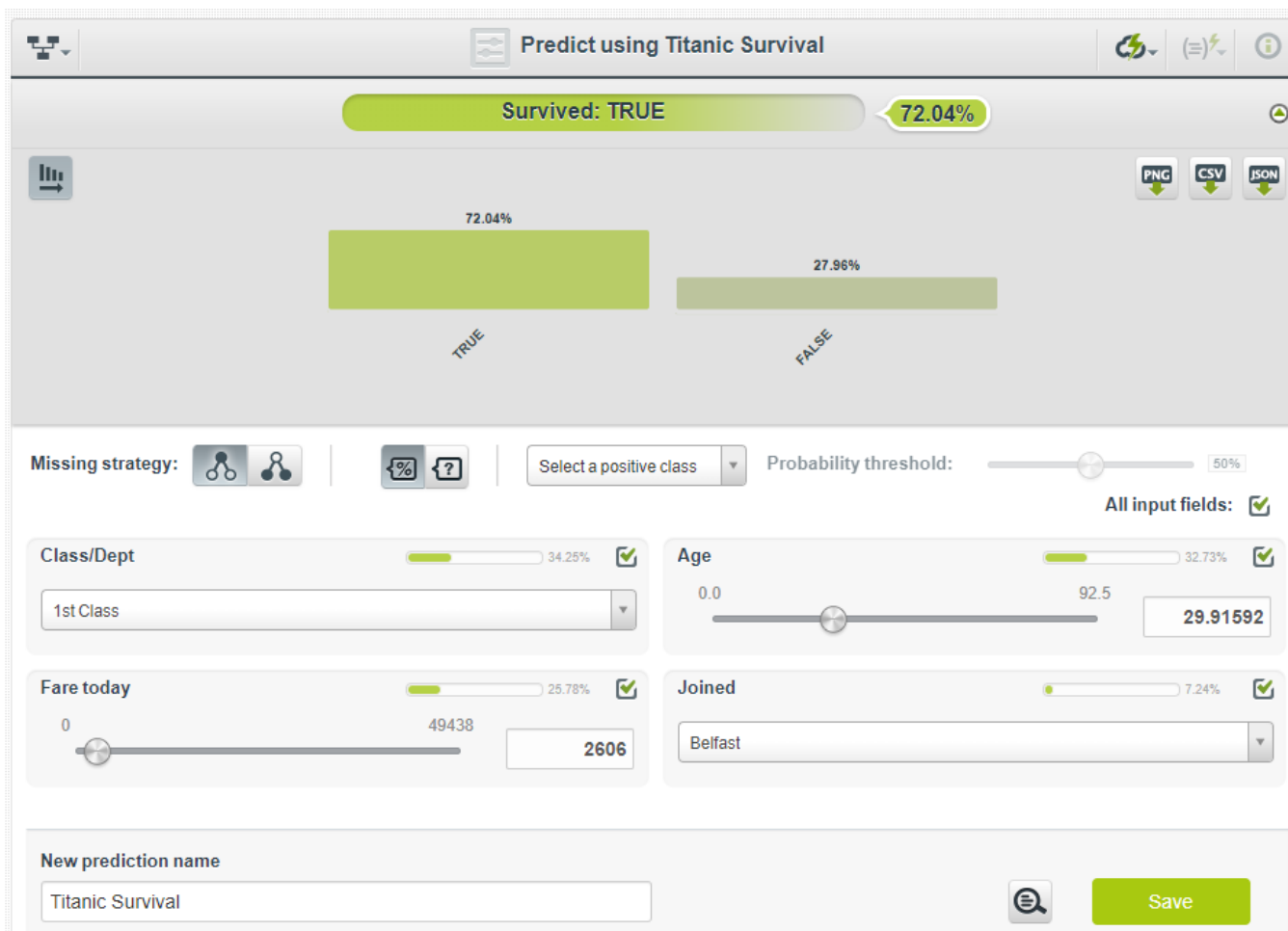
Правда выгрузка происходит именно дерева, а не полноценной модели машинного обучения.

12. Также можно использовать обученную модель непосредственно в сервисе (не выгружая). Для этого необходимо выбрать пункт для построения прогнозов:



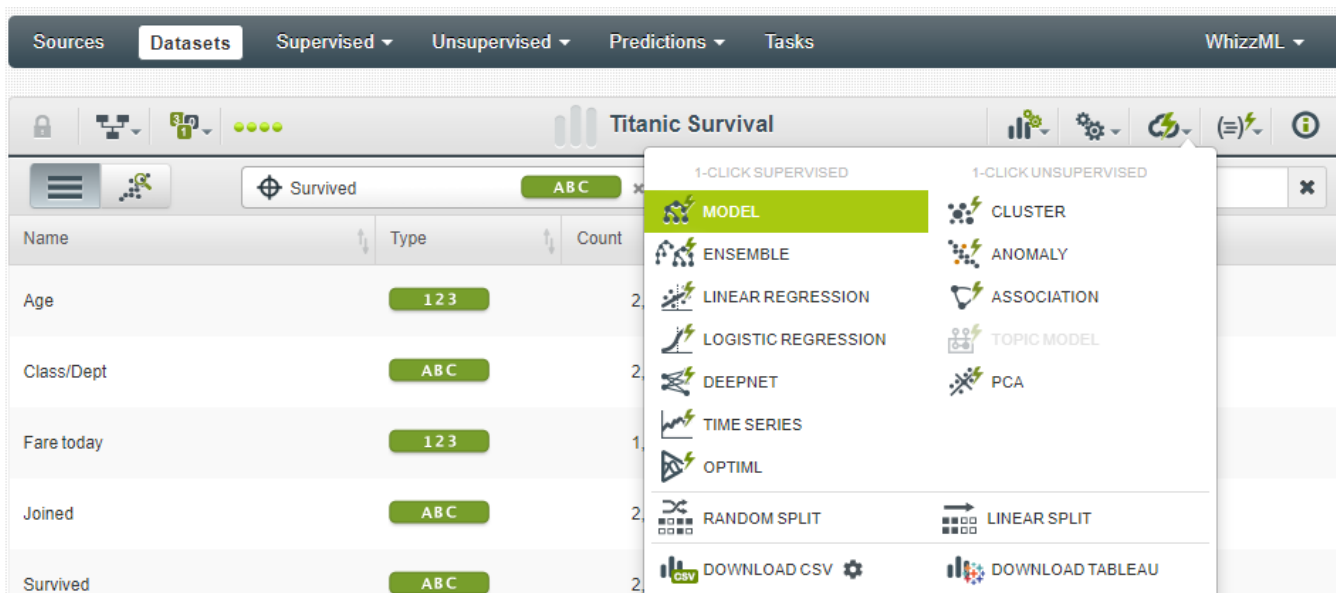
В появившемся окне можно задать параметры конкретного человека и увидеть: выжил бы он или нет, если бы был пассажиром Титаника.



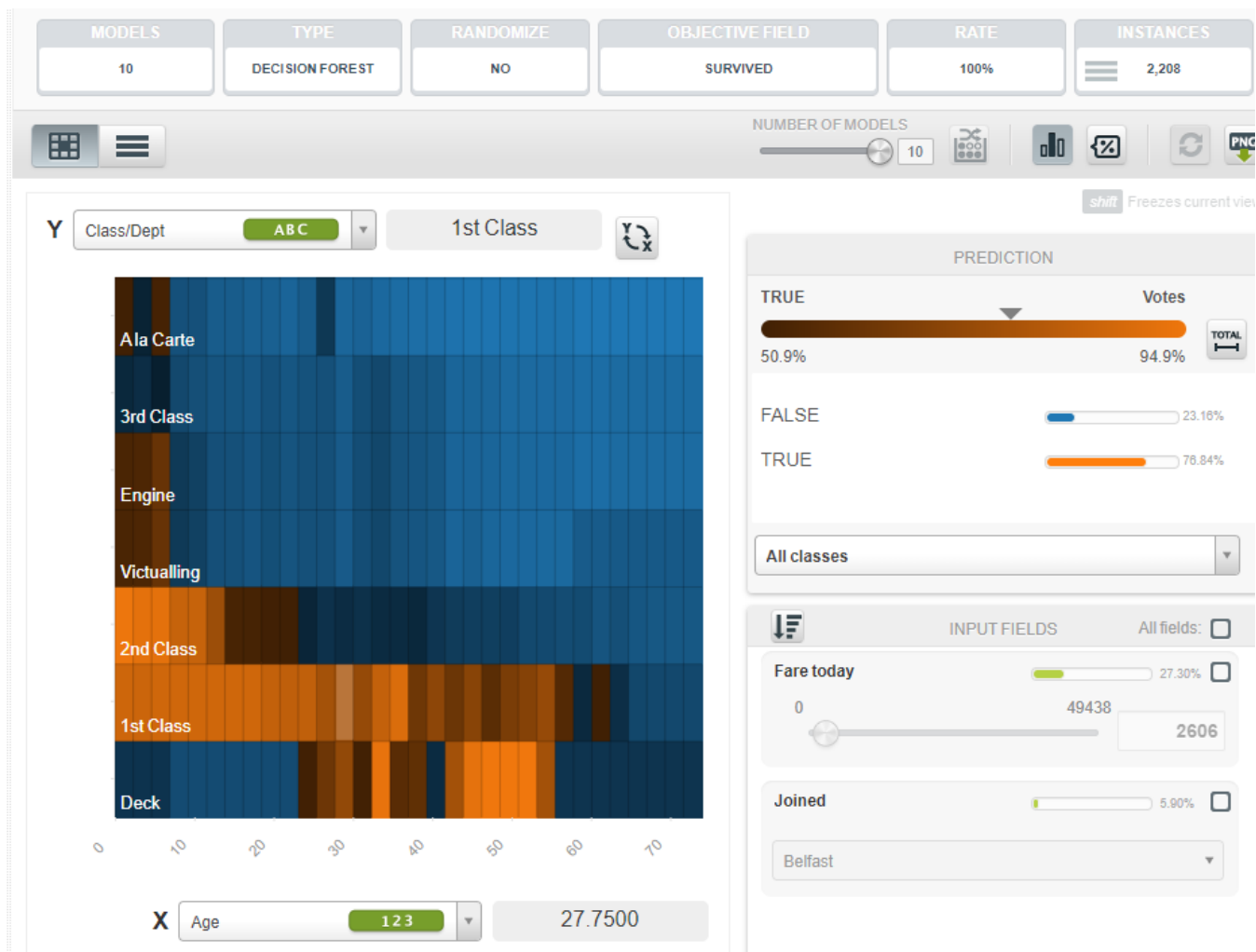


Нажав на кнопку Save, можно сохранить сам прогноз и используемые данные для него.

13. Перейдя на страницу Datasets, можно обучить модель, содержащую ансамбль деревьев. Для этого необходимо выбрать пункт Ensemble.



Так как при построении ансамблей используются несколько деревьев, то в результате будет показано окно, содержащее интерактивную диаграмму со значениями полей. Водя мышкой по этой диаграмме, вы будите попадать в какие-то ее точки, соответствующие конкретным параметрам человека, а справа можно увидеть результат предсказания: выжил этот человек или нет.



14. Теперь решите для этого же набора данных задачу регрессии, то есть задачу получения прогноза. Например, можно прогнозировать возраст человека на основании других характеристик, используемых в датасете Титаник (в том числе и по информации о том, выжил человек или нет). Для этого необходимо указать, что теперь для нас целевым будет являться поле Age. Для того, чтобы каждый раз для разных моделей не переключать целевые поля, можно создать копию имеющегося датасета и уже в ней указать новую целевую переменную.

В рамках этого же проекта повторите действия по созданию Dataset на основании все того же источника данных Titanic Survival.

Для того, чтобы не путаться в наборах данных поменяем у вновь созданного датасета имя на TitanicAge.

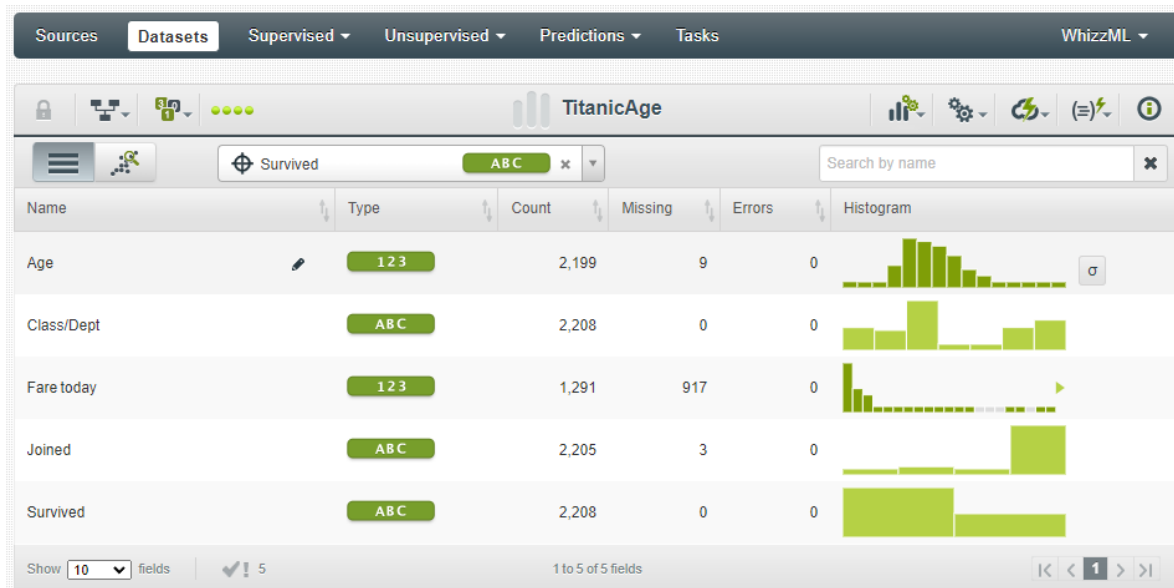
SEMENOVASSAU - My Dashboard | Project Titanic



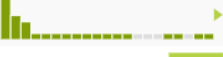


Sources	Datasets	Supervised	Unsupervised	Predictions	Tasks	WhizzML
Datasets						
Name						
Titanic Survival	2208 instances, 5 fields (3 categorical, 2 numeric)	1min	78 K	0	0	0
Titanic Survival	2208 instances, 5 fields (3 categorical, 2 numeric)	2h	78 K	1	1	0

Show 10 datasets | 1 to 2 of 2 datasets

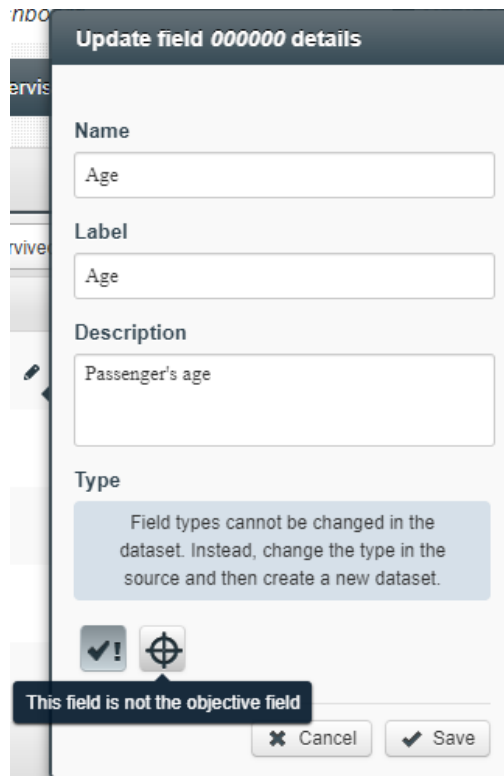
Перейдя в сам датасет, укажите, что теперь целевой будет переменная Age.

Для этого нажав на карандаш рядом с полем



Name	Type	Count	Missing	Errors	Histogram
Age	123	2,199	9	0	
Class/Dept	ABC	2,208	0	0	
Fare today	123	1,291	917	0	
Joined	ABC	2,205	3	0	
Survived	ABC	2,208	0	0	

нужно нажать следующую кнопку:



**Update field 000000 details**

Name: Age

Label: Age

Description: Passenger's age

Type: Field types cannot be changed in the dataset. Instead, change the type in the source and then create a new dataset.

☒ Active ☐ Objective

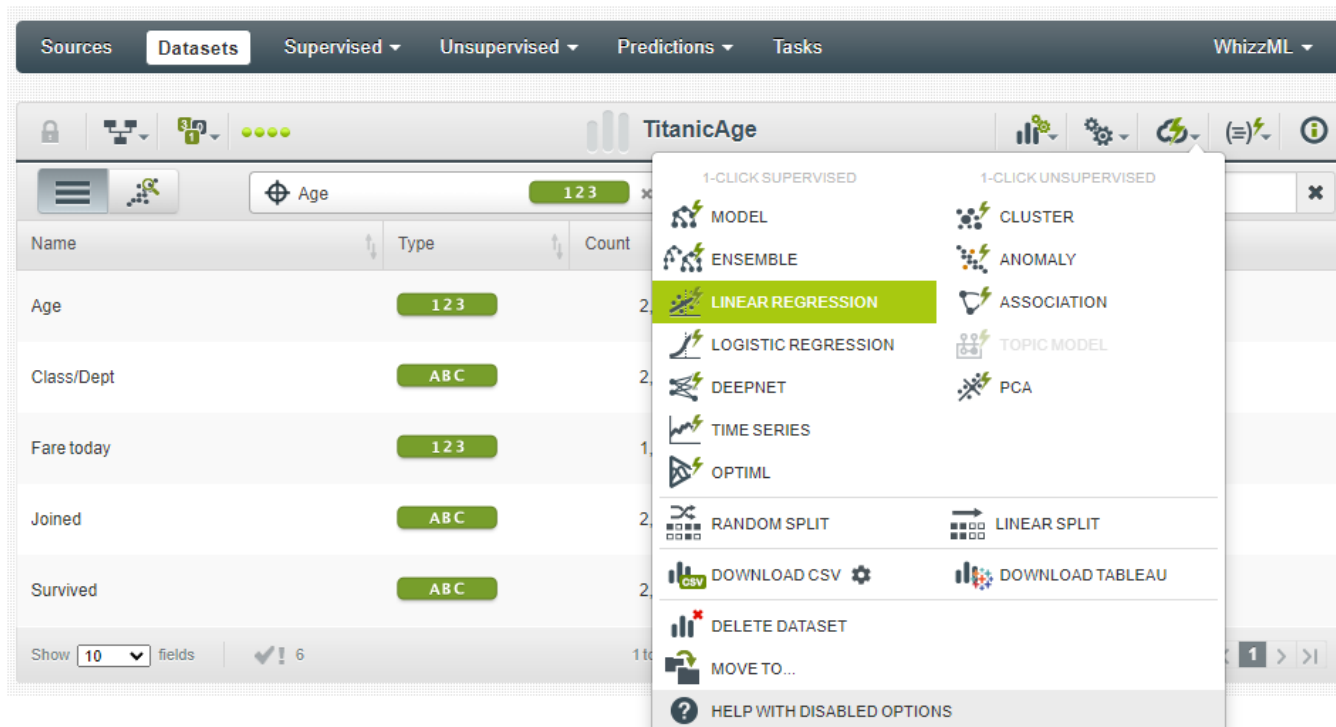
This field is not the objective field

Cancel Save

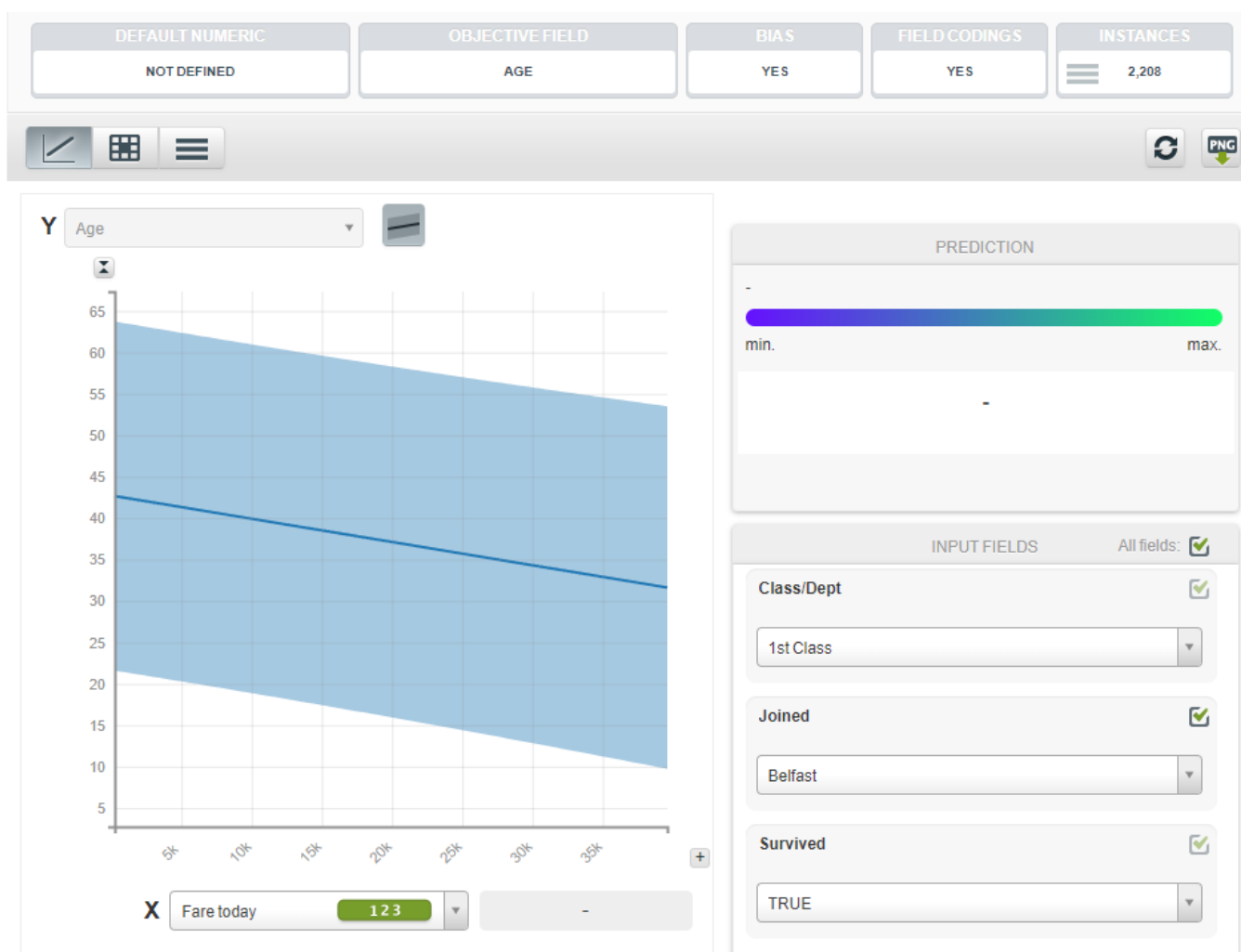
и сохранить сделанные изменения.

Проверьте, что в поле Survived такой значок стал неактивным.

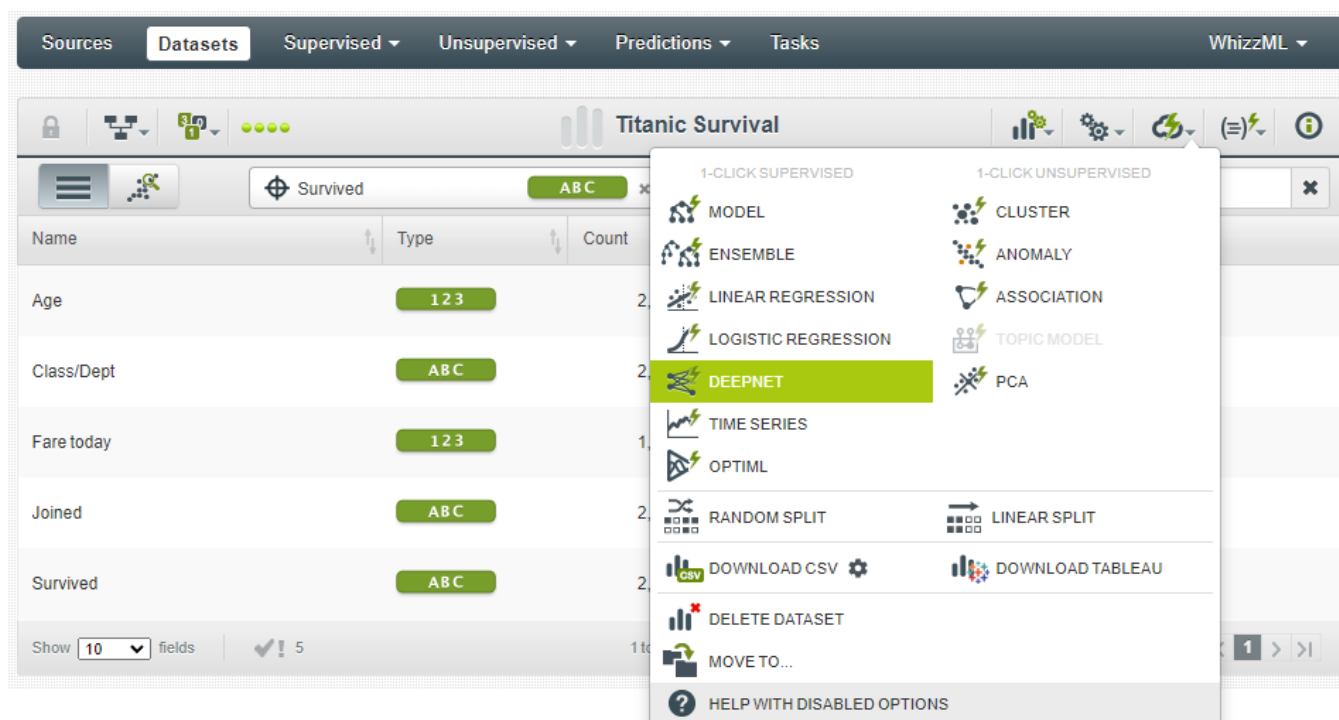
Теперь обучите модель линейной регрессии:



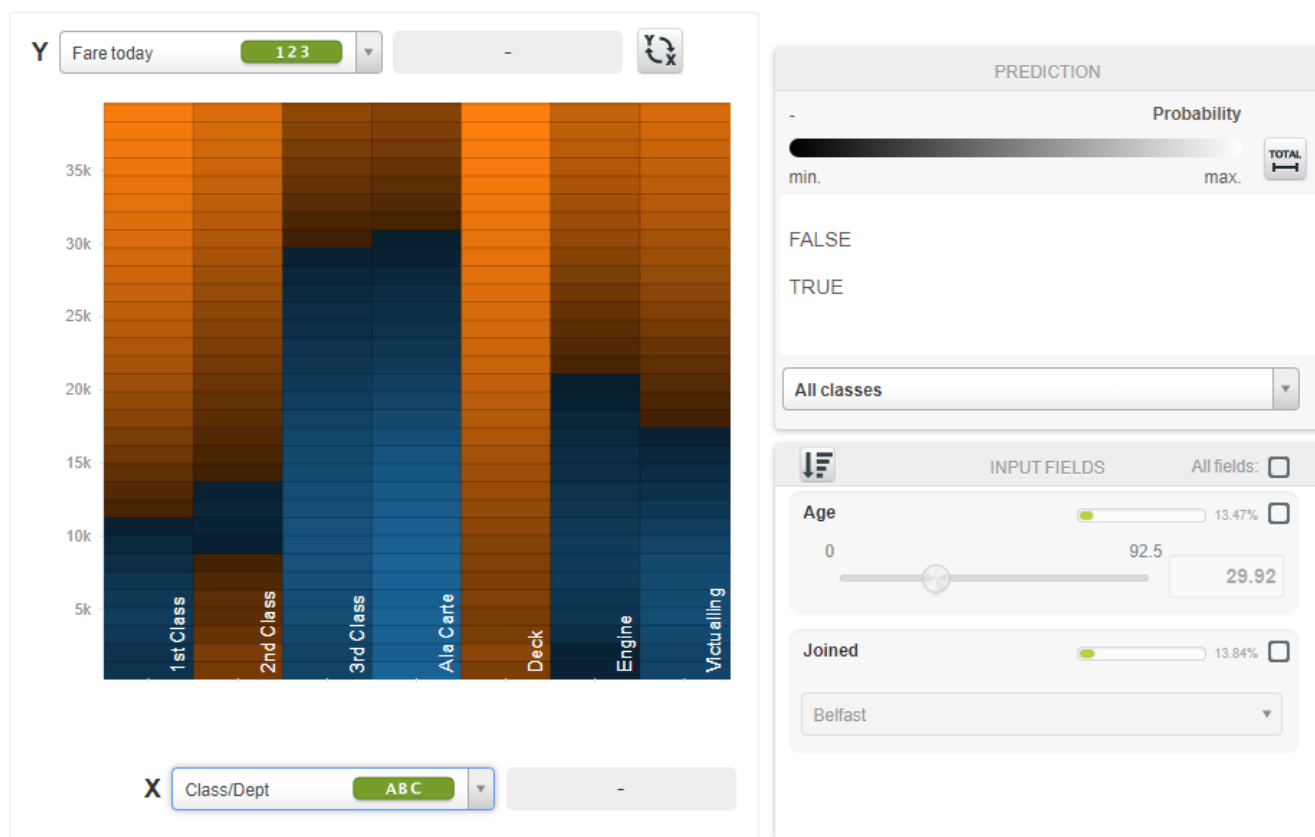
В появившемся окне можно увидеть, как изменяется возраст пассажиров в зависимости от значений других параметров:



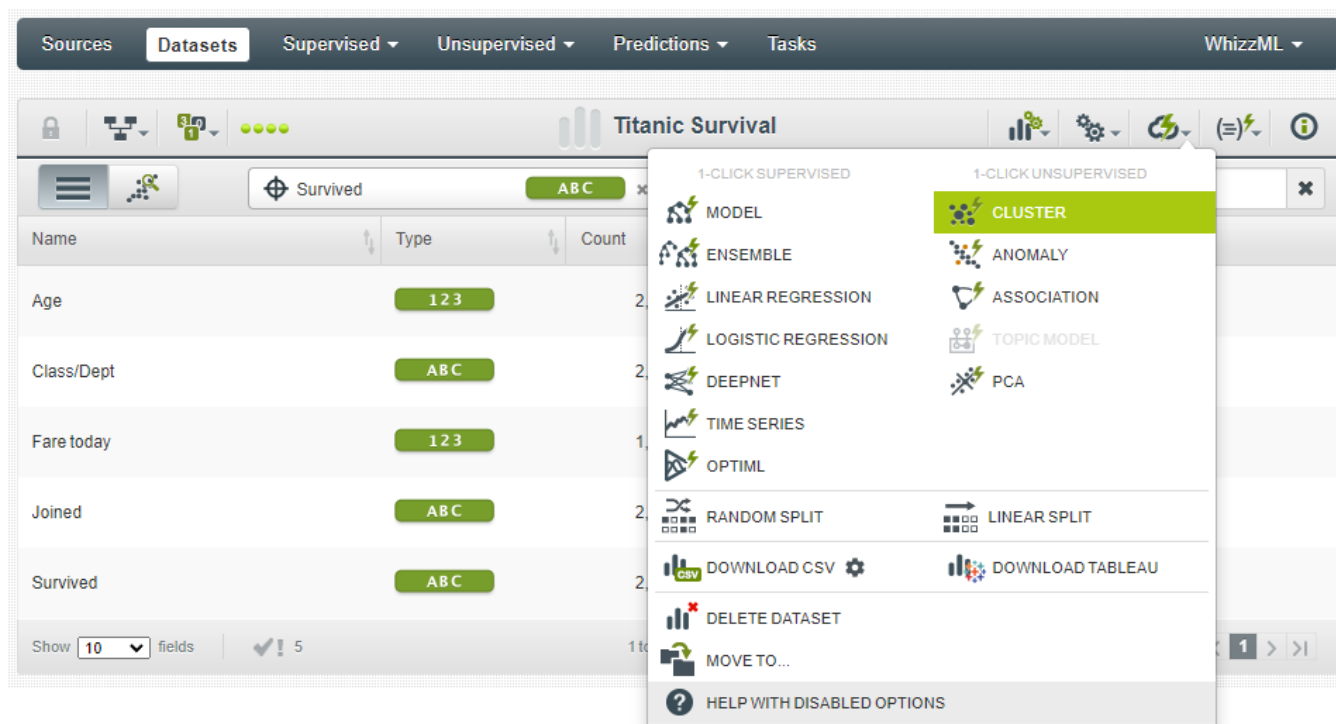
15. Аналогично для любого из созданных датасетов можно решить задачу с помощью нейронной сети:



Например, для первого датасета, в котором решалась задача классификации результат будет выглядеть следующим образом:



16. Кроме задач классификации и регрессии практически интересной задачей для рассматриваемого набора данных может быть задача кластеризации

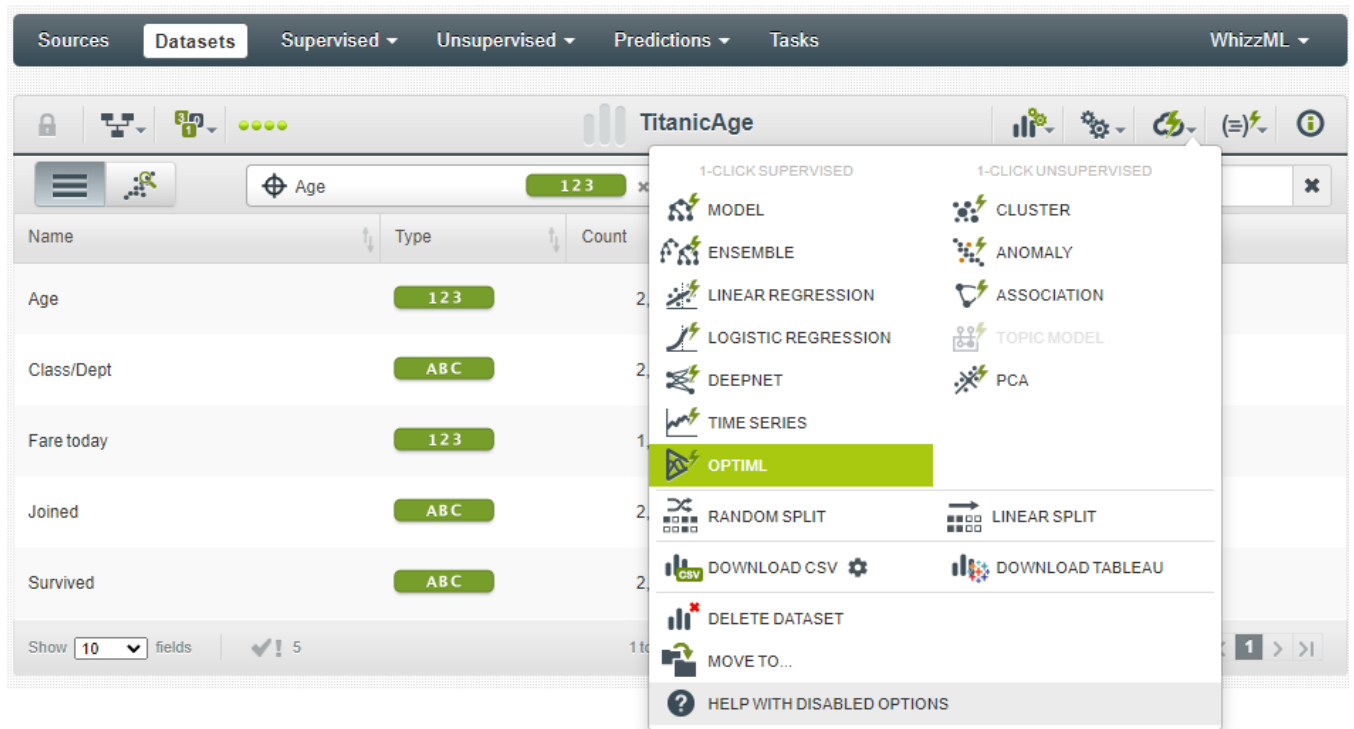


Ее результат может выглядеть так:



Вот на такие группы модель разделила всех пассажиров Титаника. При наведении мышки на каждый кружок можно увидеть значения параметров для этой группы.

17. В этом сервисе также есть возможность использования AutoML – подбора наилучшей модели и ее параметров для решения задачи.



Однако для этого потребует достаточно много времени (до нескольких часов).

## **Задание 2.**

1. Создайте новый проект
2. Выберите любой другой понравившийся источник данных, предлагаемый в сервисе
3. Создайте на его основе датасет
4. Обучите минимум 4 различных модели для этого источника данных
5. Продемонстрируйте получившийся результат.